http://www.jstage.jst.go.jp/browse/biophysics

BIOPHYSICS

Vol. 10, pp. 55–62 (2014)
doi: 10.2142/biophysics.10.55

Familial clustering of mice consistent to known pedigrees enabled by the genome profiling (GP) method

Harshita Sharma[1], Fumihito Ohtani[1], Parmila Kumari[1], Deepti Diwan[1], Naoko Ohara[2], Tetsuya Kobayashi[2], Miho Suzuki[1], Naoto Nemoto[1], Yoshibumi Matsushima[3] and Koichi Nishigaki[1]

[1]*Department of Functional Materials Science, Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Saitama 338-8570, Japan*
[2]*Department of Regulatory Biology, Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Saitama 338-8570, Japan*
[3]*Research Institute for Clinical Oncology, Saitama Cancer Center, Japan*

Received May 16, 2014; accepted August 21, 2014

Familial clustering without any prerequisite knowledge becomes often necessary in Behavioral Science, and forensic studies in case of great disasters like Tsunami and earthquake requiring body-identification without any usable information. However, there has been no well-established method for this purpose although conventional ones such as short tandem repeats (STR) and single nucleotide polymorphism (SNP), which might be applied with toil and moil to some extent. In this situation, we could find that the universal genome distance-measuring method genome profiling (GP), which is made up of three elemental techniques; random PCR, micro-temperature gradient gel electrophoresis (μTGGE), and computer processing for normalization, can do this purpose with ease when applied to mouse families. We also confirmed that the sequencing approach based on the ccgf (commonly conserved genetic fragment appearing in the genome profile) was not completely discriminative in this case. This is the first demonstration that the familial clustering can be attained without a priori sequence information to the level of discriminating strains and sibling relationships. This method can complement the conventional approaches in preliminary familial clustering.

Key words: pedigree analysis,
universal genotyping method,
genome distance

Corresponding author: Koichi Nishigaki, Department of Functional Materials Science, Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Saitama 338-8570, Japan.
e-mail: koichi@fms.saitama-u.ac.jp

The genome sequence can be powerfully used for effectively confirming or refuting alleged familial relationships, which had previously been addressed by phenotype-based methods such as finger-prints, teeth shape/alignment, and blood types. Up to now, the genetic variation analysis has enabled us to do the parentage testing in civil and criminal investigations, disaster victim and missing person identifications, and other forensic and clinical purposes. However, with all genotyping methods such as AFLP (amplified fragment length polymorphism), RAPD (random amplified polymorphic DNA), STRs (short tandem repeats) and SNP (single nucleotide polymorphism)[1–5], familial clustering without any prerequisite knowledge (familial clustering is a concept that those individuals as belong to the same family line are binned in the same box) has never been established. The reliable and general genotyping techniques must have the properties of: *i*) being universally applicable, *ii*) being robust in testing low-quality samples and, *iii*) being sufficient in the information amount to provide.

The most acknowledged approach for familial clustering is currently the STR method[3,4], which is based on the genetic polymorphism in the repeat number of particular short sequences[6]. Since around 20 different loci in the human genome have been assigned for this purpose, this approach can be readily performed only by running PCR and gel electrophoresis and can identify individuality uniquely and calculate the relative closeness from the degree of correspondence (the number of matched loci out of the total loci examined) so as to build up the familial relationship. In this process, the detailed information of homo/heterozygosity can be utilized for confirmation of assignment. However, in the STR method, information of point mutation in the genome is not positively used though it often behaves as an obstacle occurred at the primer binding site with the signal band vanishing[7]. On the other hand, the GP method exploits the point mutation contained in the genome sequence and calculates the genome distance to obtain the relationship of different genomes (organisms). Besides, the GP method can be performed without any preceding information on the genome such as STR loci and their primer sequences. In other words, it can be applied directly to any organisms, of which the genome sequence is available or not. These are the biggest methodological differences between two approaches. Besides, the STR analysis is applicable to a limited number of species which are well-investigated such as human being and mouse. For the wild life research for a particular species, STR becomes possible after having been available of its genome sequence. Obviously, sequencing-based approaches such as 16S/18S rDNA sequencing and MLST (multi-locus sequence typing) require the purified DNAs and time-consuming processes as a standard protocol. From these reasons, a general and reliable familial clustering method has not been presented till now.

In this context, the GP (genome profiling) method, which comprises three main steps: (*i*) random PCR, (*ii*) µTGGE (Micro Temperature Gradient Gel Electrophoresis) and (*iii*) computer-aided normalization (which are explained in detail in Materials and Methods) and has been applied to species identification of various taxa and groups of organisms (bacteria[8], fungi[9], protozoa[10], insects[11], vertebrates[12], and plants[13]) together with measuring the genome distance for mutagen assay[14] and others[15], was novelly explored for this purpose. Through these steps, a parameter called genome distance (being equivalent to the difference in genome sequences) can be obtained and was exploited to examine the familial clustering. In this study, two lines of experiments were performed rearing experimental mice with our novel approach.

Materials and Methods

Genome Profiling (GP)

The GP method consists of three main steps: *1)* random DNA sampling from the genomic DNA by random PCR[16,17], *2)* extraction of the DNA sequence information without sequencing but with the µTGGE (micro temperature gradient gel electrophoresis) method[18], and *3)* computer-based data processing as to the genome profiling pattern for obtaining the normalized values, *spiddos* (species identification dots)[19]. Random PCR involves the PCR starting with mismatch or bulge-containing primer-template hybrid structures and performs random sampling from the whole genomic DNA (see Supplementary Fig. 1). Next, random PCR products are processed by µTGGE, resulting in sequence-specific DNA melting profile for each DNA. The mobility transition point of each of the DNA band in a genome profile can be denoted as 'featuring point' and corresponds to the initial DNA melting. These featuring points are then subjected to the computer aided normalization utilizing internal reference points to generate *spiddos*[19]. Each coordinate of a *spiddos* is unique, made of two factors temperature and mobility. A set of *spiddos* (~10 points) are proven to provide a sufficient amount of information for identifying species[19]. A set of *spiddos* of two genomes are compared and reduced to *PaSS* (pattern similarity score) as follows (for details see Supplementary protocol and Supplementary Fig. 2)

$$PaSS = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\left|\vec{P}_i^{(1)} - \vec{P}_i^{(2)}\right|}{\left|\vec{P}_i^{(1)}\right| + \left|\vec{P}_i^{(2)}\right|} \qquad [1]$$

Here $\vec{P}_i^{(1)}$ and $\vec{P}_i^{(2)}$ represent the normalized positional vectors (function of temperature and mobility) for *spiddos* $P_i^{(1)}$ and $P_i^{(2)}$ from genome (1) and genome (2) respectively, while *i* designates serial number of *spiddos*. *PaSS* will be 1 for a complete match in two sets of *spiddos*. In general,

$$0 \leq PaSS \leq 1. \qquad [2]$$

A measure of genome distance d_G can be derived from *PaSS*:

$$d_G = 1 - PaSS \quad (0 \leq d_G \leq 1) \qquad [3]$$

Higher the d_G value, higher the distance between two genomes. In other words, d_G value is 0 for a perfect match and near 0 for members of the same species.

Samples and their DNA

All blood samples were obtained from three *Mus musculus* families of strains SKC/Stm (n=8), A/JmsSlc (n=3), and C3H/HeSlc (n=5) reared at Research Institute for Clinical Oncology, Saitama Cancer Center (Japan) (Fig. 1A). In another independent experiment for single family-multigeneration genome analysis blood samples of four generations of a single growth-retarded (grt) mouse (originally derived from DW/J strain of mouse[20]) family (n=31) were obtained from the Department of Regulatory Biology, Saitama University (Fig. 3A, and Supplementary Table 1). DNAs from all of the samples were extracted using Dr. GenTLE DNA extraction Kit (Takara Bio, Japan) following the manufacturer's protocol. The three distinct mice families' genome analysis experiments were covered by the permission of Regulation on Animal Experimentation at Saitama Cancer Center. This study was carried out in strict accordance with the Guidelines for the Care and Use of Experimental Animals at Saitama University, Japan. All protocols for animal experiments were approved by the Saitama University Institutional Animal Care and Use Committee.

Random PCR

Isolated DNA samples were randomly amplified using PCR mix containing 200 µM dNTPs (N=G, A, T, or C), 0.5 µM primer, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, 0.03 U/µL Taq DNA polymerase (Takara,

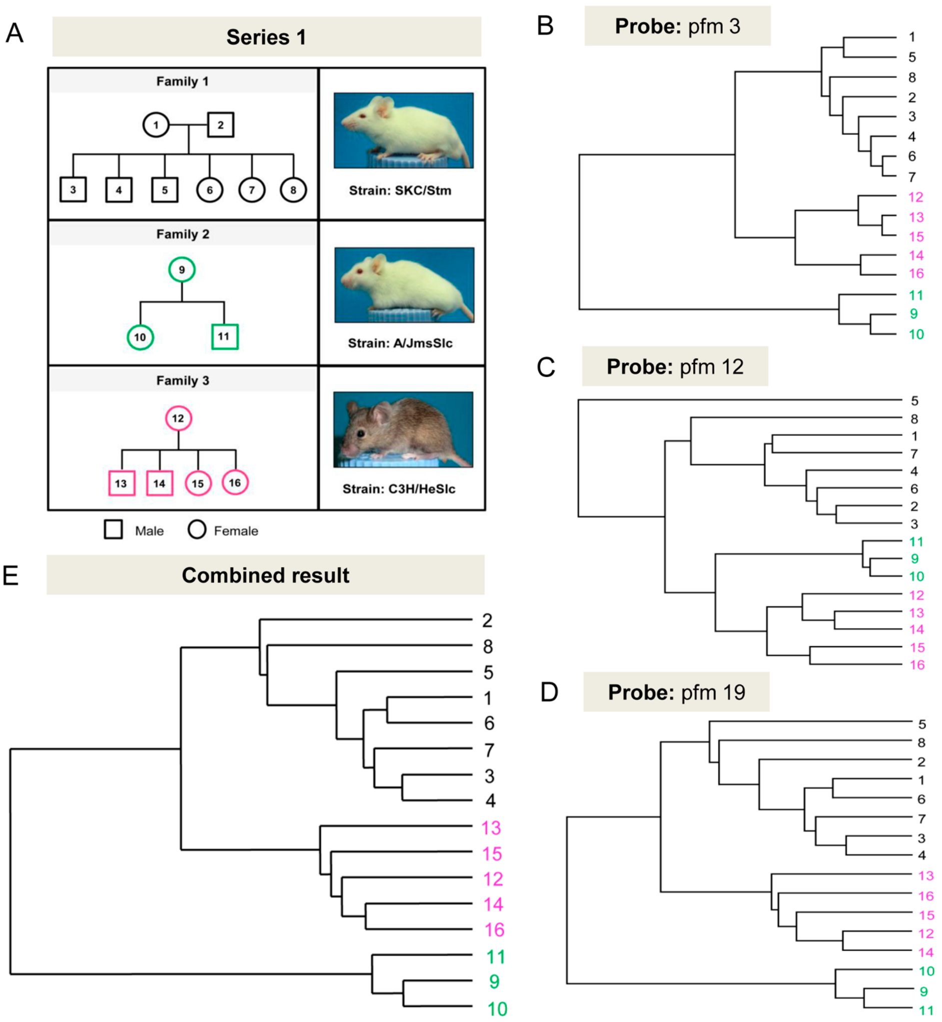

Figure 1 GP-based familial relationship analysis (Series 1) of three mouse families. (A) Pedigree charts of three distinctive mouse families used. Total 16 samples of three *Mus musculus* families of strains SKC/Stm (n=8), A/JmsSlc (n=3), and C3H/HeSlc (n=5) of a known pedigree were analyzed by GP. Familial clustering of samples amplified by primers (B) pfM 3, (C) pfM 12, and (D) pfM 19. (E) Combined genome distance-based clustering result of 3 different probe experiments. All the trees were constructed using d_G matrix data in DendroUPGMA web utility with Pearson coefficients and visualized using TreeView software.

Japan). All of the PCR mix contents except template DNA were treated with UV irradiation for 8 min in a laminar air flow, prior to the preparation of PCR master mix. In the first experiment, three different primers pfM 3, pfM 12, and pfM 19 were tested (sequence information in Table 1) for familial clustering of mice (16 individuals) from three distinctive mouse families. In another genome analysis of a single family-multigeneration, 31 samples of four generations of a single mouse family were dealt and their DNAs were amplified using pfM 12 primer. In case of pfM 12 and pfM 19, the standard random PCR conditions were applied (30 cycles of denaturation at 94°C for 15 s, annealing at 30°C for 30 s, and extension at 50°C for 30 s) using a Bio-Rad C1000 Touch thermocycler (Tokyo, Japan) whereas for the primer pfM 3-dependent random PCR, a modified thermal cycle program consisting 30 cycles of denaturation at 94°C for 30 s, annealing at 26°C for 60 s was used for the sake of obtaining an appropriate number of DNA bands.

Micro-TGGE and data processing

Random PCR products were analyzed by μTGGE, a miniaturized version of TGGE (Temperature Gradient Gel Electrophoresis)[18]. A slab gel [6% (w/v) denaturing polyacrylamide gel (acrylamide : bis = 19 : 1) comprising 500 mM Tris-HCl, 485 mM boric acid, 20 mM EDTA (pH 8.0) with 8 M Urea] of size 1 inch×1 inch was used[18]. The PCR product was loaded on the top of gel along with two internal reference DNAs: *i*) internal Ref1 of 200-bp, (a 191-bp fragment from the bacteriophage fd gene VIII, sites 1350~1540 attached to a 9-bp sequence, CTACGTCTC, at the 3'-end; T_m of 60°C under the standard conditions) and *ii*) internal Ref2 of 900-bp taken from pBR322 (T_m of 61.4°C under the

Table 1 Sequence information of primers used in this study

No.	Primer Name	Nucleotides	5′→3′ Sequence[a]	Purpose
1.	pfM 3	12mer	CY3- CTGGATAGCGTC	Genome Profiling
2.	pfM 12	12mer	CY3- AGAACGCGCCTG	Genome Profiling
3.	pfM 19	12mer	CY3- CAGGGCGCGTAC	Genome Profiling
4.	mcF	22mer	GTCAGTCCTCAGTGTCACATTA	CCGF amplification
5.	mcR	18mer	CCACAGACACAGAACTGG	CCGF amplification

[a] CY3 is a fluorescent dye labeled at 5′ end of each oligonucleotide.

standard conditions) and subjected to electrophoresis with a μTGGE apparatus Micro TG (Taitec, Japan) at 100 V for 10 min. The loaded sample DNA was migrated under the temperature gradient of 15°C–65°C set perpendicular to the direction of migration. The DNA bands were visualized with either intrinsic fluorophor CY3 or nucleic acid stain SYBR gold using a fluoroimager Molecular imager FX (BIO-RAD, USA). A set of featuring points, corresponding to the initial DNA melting point of each double-stranded DNA, were manually extracted from each genome profile. Initial transition points of internal reference bands of a known melting pattern were used to normalize each featuring point to obtain '*spiddos*' of normalized temperature and mobility. *PaSS* values were calculated in microTGGE software and d_G (genome distance) values were derived. Clustering was performed using d_G applying to DendroUPGMA web utility (http://genomes.urv.cat/UPGMA/) with Pearson coefficients[21] and visualized using TreeView software[22]. For the single family-multigeneration genome analysis, the UPGMA method in phylip-3.69 and MEGA 5.1 viewing software[23] were used as an identical alternative.

CCGF sequencing

CCGFs are defined as the corresponding DNA bands observed in different species genome profiles as close band patterns (similar melting temperature and mobility)[8]. These are assumed to originate from the corresponding genetic locus (ortholog, paralog, and like). Here, an around 300 bp band commonly appearing in all of the pfM 19-amplified samples was selected as a possible ccgf (Supplementary Fig. 3). The corresponding DNAs were collected basically following the procedure previously described[8]. The resulting PCR products were ligated to pGEM-T Easy Vector (Promega, Japan) at 4°C for overnight and then transformed in *E. coli* DH5α competent cells (Toyobo, Japan) and cloned. The plasmid DNAs purified using Wizard™ Plus SV Minipreps DNA Purification System (Promega, Japan) were commercially sequenced (Operon Bio-Technology Co., Ltd. Japan) and manually analyzed using BLASTn against NCBI mouse genome database.

Based on the sequence thus obtained, specific primers mcF (forward) and mcR (reverse) (Table 1) were designed. All of the 16 samples were specifically PCR amplified using PCR mix containing 200 μM dNTPs (N = G, A, T, or C), 0.5 μM primer mcF, 0.5 μM primer mcR, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, and 0.03 U/μL Taq DNA polymerase with 30 cycles of denaturation at 94°C for 30 s, annealing at 61°C for 60 s, and extension at 72°C for 60 s. PCR products were purified using PCR product purification kit (Quiagen) and commercially sequenced (Operon Bio-Technology Co., Ltd.).

Analysis of sequencing results

Sequences obtained for ccgf analysis were analyzed by MUSCLE alignment[24] and phylogenetic trees were drawn by neighbor-joining method using MEGA5.1 software[23].

Results and Discussion

The familial clustering has been a hard work to perform, which is often desperately required in such situations as disasters (Tsunami, earthquake, airplane crash, etc.) generating a large number of victims unidentified[25,26]. From the view of science, elucidation of the familial relationship among the observing field-animals such as monkeys is important especially for Behavioral and Resource Sciences[27]. Among the various techniques of AFLP[1], RAPD[2], STR[3,4], SSCP[28], and others, one of the most reliable and popular approaches is to depend on the STR method for forensic field and the 18S rDNA sequencing method for pure and applied sciences. However, these approaches require a huge amount of cost-and-labor to perform and, yet, often end in leaving works unfinished[29,30]. In this stream, the genome profiling (GP) method has already shown its potential to identify species in general[9–11].

However, the applicability of GP to the most detailed familial relationship (that is, pedigree) analysis has not been clarified, which is the point for the above mentioned purpose (especially, application to disaster cases). Therefore, this paper directly examined the possibility of familial clustering using three distinct mouse families (mutually inbred) and then a mouse family consisting of four successive generations (mutually congenic).

As shown in Figure 1, the members of the three mouse families of SKC/Stm, A/JmsSlc, and C3H/HeSlc strains were all clustered conforming to their pedigrees by all of

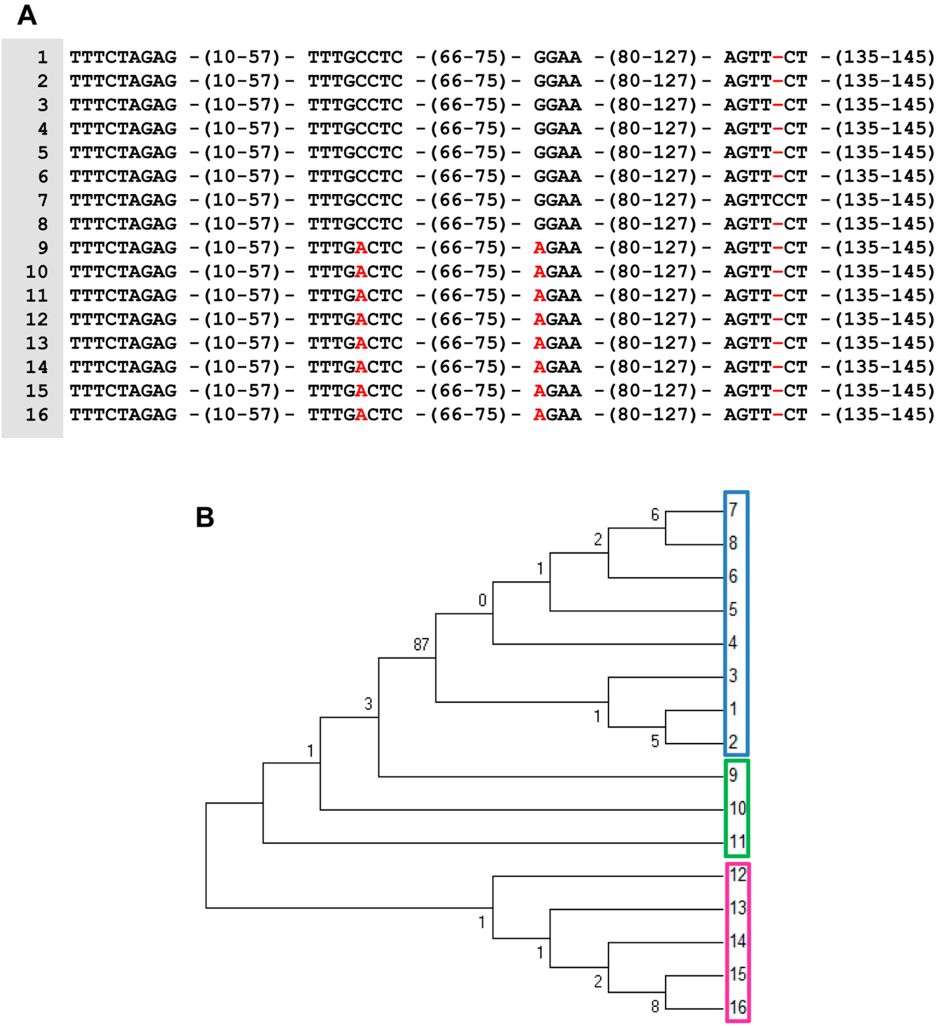
A
1 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
2 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
3 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
4 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
5 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
6 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
7 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTTCCT -(135-145)
8 TTTCTAGAG -(10-57)- TTTGCCTC -(66-75)- GGAA -(80-127)- AGTT-CT -(135-145)
9 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
10 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
11 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
12 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
13 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
14 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
15 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
16 TTTCTAGAG -(10-57)- TTTGACTC -(66-75)- AGAA -(80-127)- AGTT-CT -(135-145)
B
6
2
1
0
87
3
1
1
5
7
8
6
5
4
3
1
2
9
10
11
12
13
14
15
16
1
1
2
8

Figure 2 CCGF sequencing analysis of three mouse families. (A) Partial ccgf sequences aligned by MUSCLE. (B) CCGF sequence-based clustering of samples of three mouse families. Here, only partial sequences with difference among samples are shown for clarity. Letters in red indicate point mutations. Clustering tree is drawn by neighbor-joining method with 1,000 bootstraps in MEGA 5.1 software. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches.

three kinds of probe; pfM 3, pfM 12, and pfM 19 (Fig. 1B–D, see corresponding d_G-matrix data in Supplementary Table 2, Supplementary Table 3, and Supplementary Table 4 respectively). Naturally, the clustering based on the average genome distance obtained from these three probe results, could generate a tree with three families discretely separated as expected (Fig. 1E, see corresponding d_G-matrix data in Supplementary Table 5). Since each probe must have collected different portions of genomic DNA[16], the detailed relationship within a family is not the same among the three trees. In the sense of statistical reliability, the result of the three probe average is, naturally, most reliable. Even so, the fact that the single, simple run of the GP experiment could succeed in generating the three family discriminative clustering is very promising and useful. For comparison, regarding these mouse families, we have investigated the possibility of sequencing-based clustering with GP-derived ccgf (commonly conserved genetic fragments)[8] sequencing analysis. We also tested conventional 18S rDNA sequencing, but it failed to discriminate between families (Supplementary Fig. 4) supporting its less polymorphic nature.

Following the ccgf protocol[8], ~300-bp DNA band commonly appearing in all of the pfM 19-amplified genome profiles was extracted and sequenced. Thus obtained sequence was analyzed by BLASTn against NCBI mouse genome database, hitting the sequence of ID: ref|NC_000067.6| and sequence range: from position 125778705 to 125779000 from *M. musculus* strain C57BL/6J chromosome 1 (GRCm38.p1 C57BL/6J featuring G-protein coupled receptor 39). Based on this ccgf sequence, specific primers were designed so as to cover a 145 bp part of this. The sequences obtained for 16 mice samples were aligned by MUSCLE and a phylogenetic tree was drawn by neighbor joining method (Fig. 2A and 2B). In this case, the clustering was successful

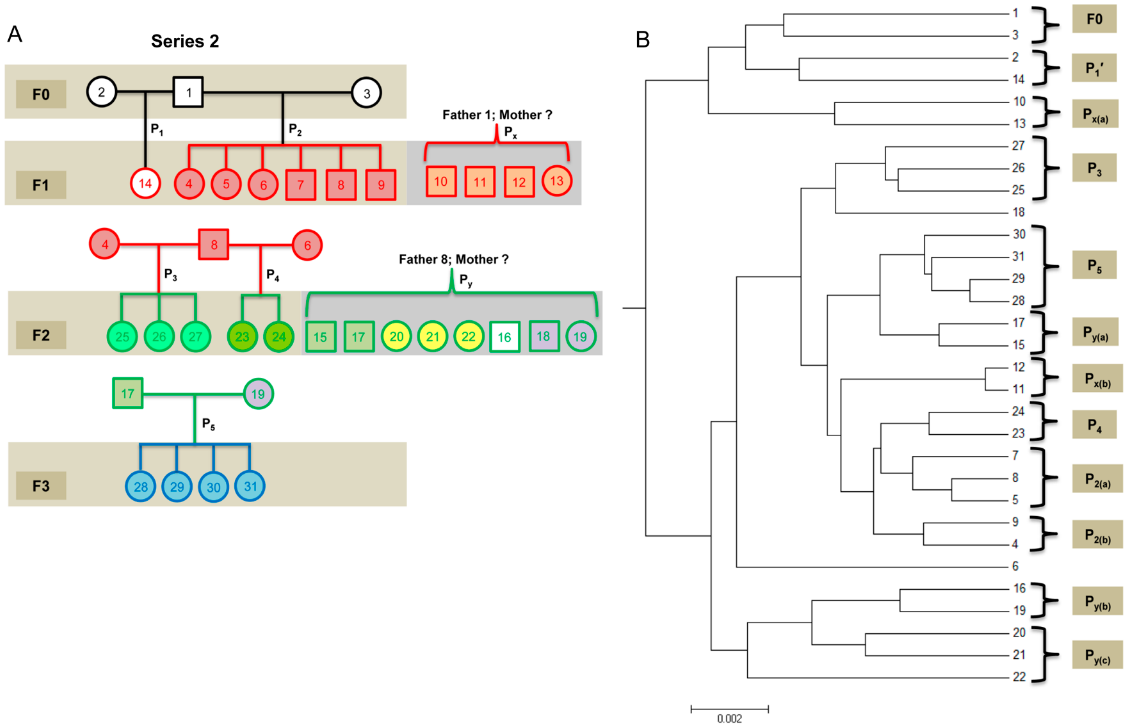

Figure 3 Single family-multigeneration genome analysis (Series 2) by GP. (A) Pedigree chart of four generations (F0, F1, F2, and F3) of a single mouse family. Here, male and female samples are indicated by square and circle, respectively. Each parentage is shown by one of P_1~P_5, P_x, and P_y. That is, samples filled with the same color share the same birthdate and parents. Samples 10 to 13 (P_x parentage) and 15 to 22 (P_y parentage) share the same father but the identity of mother is unknown. (B) d_G-based clustering of 31 samples constructed using UPGMA method in phylip3.69 and MEGA5.1 software. P_1' shows the mother and daughter relationship liked by the parentage line P_1. The sub-clustering within a same parentage is shown with the additional alphabet like $P_{x(a)}$ and $P_{x(b)}$.

in principle though the separation of families 2 and 3 is not so discrete. This is rather surprising since the size of DNA in this ccgf analysis is very short (145 bp: actually a portion of GPCR39 of mouse) which is quite arbitrarily selected and has no reliability established.

This fact is, in a sense, very interesting since there are some unidentified sequences that are more appropriate in use of classification-purpose though it is hard to know *a priori*.

It is also evident that the amount of information provided by a single sequencing of these sizes cannot offer sufficiently reliable results, requiring more amount of sequencing like whole genome and exome sequencing[31]. This means that the already labor-demanding process becomes more complicated, costly, and laborious. On the other hand, the GP approach is so simple and ready for strengthening reliability as it requires only additional, similar experiments using the other arbitrary primers (~10 nts) if the higher reliability is needed (theoretically, the more probes, the more reliability).

We further examined the effectiveness of the GP-based familial clustering by applying it to another mouse family consisting of four generations as shown in Figure 3A. In this case, a single run of the GP experiment using the pfM 12 probe could clarify most of the familial relationships, especially the same sibling relations (P_1~P_5, P_x, and P_y in Fig. 3B). Namely, those siblings born from the same parent (mother and father) were shown closely clustered independently without preceding knowledge. More close observation of this clustering result indicates:

i) Siblings sharing the same birthdate (Supplementary Table 1) are clustering together with exception of sample 18. However, upon considering the fact that same birthdate does not mean the same parentage in this case, the exceptional case may be rationalized somehow and thus, this result of genome profiling-based clustering is quite promising and intriguing for a preliminary clustering of a large number of members.

ii) The first generation (F0) clustered separately from the other three generations. This grouping may seem to be inadequate since the obvious separation from those individuals closest in the familial tree (such as siblings under the P_2 path). However, if we consider that the GP method provides the genome distance-based closeness measured by a particular probe, we notice that it cannot be the same with the actual familial relationship since each probe provides a different aspect of the genome distance and the true one will be the limit average of the genome distances obtained with all types of probes. Eventually, we may have to be satisfied with the result which is topologically correct (i.e., actual distance may not be correct).

iii) Although the parent-progeny relationships are hard to interpret from this clustering, this may be due to complex and interweaving relationships in the current pedigree of four generations. Despite this complexity, GP has successfully extracted the relationship information for some of the parent-progeny pairs: sample 14 is clustering with its mother (sample 2) and showing a distant relationship with the rest (differently mothered-siblings). Samples 10 and 13 are reasonably showing relative closeness to F0 generation. On the other hand, sample 17 (making a cluster with its sibling sample 15) is clustering in proximity of its progeny cluster (F3 generation). Similarly, sample 23 and 24 of F2 generation are also showing short genome distance with their father (sample 8).

iv) The distance between generations F0 and F3 depicting great grandparent-great grandchild relationship is evident from this cluster.

It is quite fascinating that such interpretations can be drawn without using popular molecular markers or any sequence information of samples, which can be utilized complimentarily for confirmation of results.

The current GP approach did not pay any attention to the nature of homo/heterozygosity of the genomes. The heterozygous DNA bands appear in the genome profile typically as two transitions from the same mobility DNA band (see Supplementary Fig. 5), of which the rightmost transition (i.e., the transition occurring at the higher temperature) is conventionally adopted as a representative one out of the DNA sequence of that size. This is reasonable from the viewpoint to collect as many as differently sequenced DNA bands for the representative of a particular genome. In addition, this principle can make us possible to avoid uneven counting of a particular locus in the genome resulting from the difference of homo- and heterozygosity (i.e., for both cases of homo- and heterozygosity, only a single *spiddos* is allotted one of the two for hetero and the overlapped one for homo). Empirically, there is almost no chance to meet the case where two DNAs of quite different sequence appear in the same size[32].

This important fact can be briefly explained from the probability viewpoint: in random PCR, the input primers are consumed up for generation of the dominantly amplified DNA fragments (say, top 10)[16]. Therefore, there is a competition among possible DNA fragments which can be amplified in random PCR. Those fragments which can be initiated by the relatively stable binding of primers to the template DNA (for both directions; thus can be evaluated by the sum of the ΔG for the forward and reverse primer binding structures) can be more abundantly amplified[16,32]. In this competition, there are usually more than hundreds of different sized DNA fragments (which depends on the resolution of gel electrophoresis; typically it can discriminate the difference of less than 3% in the length so that more than 100 discriminable DNA bands in the range of 50 to 1000 bp can be found). Hence, the probability of finding two DNAs of the same length is very rare in this competition. However, interestingly, heterozygous alleles can be easily observed as indicated above (Supplementary Fig. 5) due to the close stability of those sequences (only point mutation in general). As a result, in the GP method, the homo-/heterozygosity problem can be evaded though it still holds the potential to utilize such information by an additional secondary treatment.

The technology for the mass disaster case cannot require any extra information and can extract information only from victims themselves due to unavailability of relevant information on the relationship of each victim unless phenotypic traits and belongings can be employed. The STR method may pose difficulty in quantitative analysis of such cases due to requirement of prior information and increase in complexity of allelic differences among large number of query samples. The GP method holds promise in this field owing to its simplicity, universality and non-dependence on prior sample information (note that the pedigree was only used for the confirmation of the final conclusion in our current study). However, additional experiments are desired to determine its familial clustering potential in real mass disaster incidents.

Conclusion

In the field such as familial relationship analysis, which is often required of rapid and broad identification of samples, the GP method was shown to be very potent. This is based on the demonstration that only a single run of the GP experiment (i.e., in a rapid and low cost manner) could cluster three different mouse families discriminatively and also cluster siblings of the same parent without *a priori* knowledge, which has, to our knowledge, no precedent. Its universal nature combined with ease of handling and simple data analysis makes it a suitable technique for preliminary familial relationship analysis when phenotypic and genotypic information is not available.

Acknowledgement

We are grateful for the financial support of the special grant from Saitama University Research Organization Project [2011].

References

1. Otsen, M., den Bieman, M., Kuiper, M. T., Pravenec, M., Kren, V., Kurtz, T. W., Jacob, H. J., Lankhorst, A. & van Zutphen, B. F. Use of AFLP markers for gene mapping and QTL detection in the rat. *Genomics* **37**, 289–294 (1996).
2. Bardakci, F. & Skibinski, D. O. F. Application of the RAPD technique in tilapia fish: species and subspecies identification. *Heredity* **73**, 117–123 (1994).
3. Dixon, L. A., Dobbins, A. E., Pulker, H. K., Butler, J. M.,

Vallone, P. M., Coble, M. D., Parson, W., Berger, B., Grubwieser, P., Mogensen, H. S., Morling, N., Nielsen, K., Sanchez, J. J., Petkovski, E., Carracedo, A., Sanchez-Diz, P., Ramos-Luis, E., Brion, M., Irwing, J. A., Just, R. S., Loreille, O., Parsons, T. J., Syndercombe-Court, D., Schmitteri, H., Stradmann-Bellinghausenj, B., Benderj, K. & Gill, P. Analysis of artificially degraded DNA using STRs and SNPs—results of a collaborative European (EDNAP) exercise. *Forensic Sci. Int.* **164**, 33–44 (2006).

4. Coomber, N., David, V. A., O'Brien, S. J. & Menotti-Raymond, M. Validation of a short tandem repeat multiplex typing system for genetic individualization of domestic cat samples. *Croat Med. J.* **48**, 547–555 (2007).
5. Fan, J. B., Chen, X., Halushka, M. K., Berno, A., Huang, X., Ryder, T., Lipshutz, R. J., Lockhart, D. J. & Chakravarti, A. Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res.* **10**, 853–860 (2000).
6. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* **5**, 435–445 (2004).
7. Phillips, C., Fondevila, M., Garcia-Magarinos, M., Rodriguez, A., Salas, A., Carracedo, A. & Lareu, M. V. Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. *Forensic Sci. Int. Genet.* **2**, 198–204 (2008).
8. Naimuddin, M., Kurazono, T. & Nishigaki, K. Commonly conserved genetic fragments revealed by genome profiling can serve as tracers of evolution. *Nucleic Acids Res.* **30**, e42 (2002).
9. Hamano, K., Tsuji-Ueno, S., Tanaka, R., Suzuki, M., Nishimura, K. & Nishigaki, K. Genome profiling (GP) as an effective tool for monitoring culture collections: a case study with *Trichosporon*. *J. Microbiol. Methods* **89**, 119–128 (2012).
10. Kouduka, M., Matsuoka, A. & Nishigaki, K. Acquisition of genome information from single-celled unculturable organisms (radiolaria) by exploiting genome profiling (GP). *BMC Genomics* **7**, 135 (2006).
11. Ahmed, S., Komori, M., Tsuji-Ueno, S., Suzuki, M., Kosaku, A., Miyamoto, K. & Nishigaki, K. Genome profiling (GP) method based classification of insects: congruence with that of classical phenotype-based one. *PLoS ONE* **6**, e23963 (2011).
12. Suwa, N., Ikegaya, H., Takasaka, T., Nishigaki, K. & Sakurada, K. Human blood identification using the genome profiling method. *Leg. Med.* **14**, 121–125 (2012).
13. Kouduka, M., Sato, D., Komori, M., Kikuchi, M., Miyamoto, K., Kosaku, A., Naimuddin, M., Matsuoka, A. & Nishigaki, K. A solution for universal classification of species based on genomic DNA. *Int. J. Plant Genomics* **2007**, Article ID 27894 (2007).
14. Futakami, M., Salimullah, Md., Miura, T., Tokita, S. & Nishigaki, K. Novel mutation assay with high sensitivity based on direct measurement of genomic DNA alterations: comparable results to the Ames test. *J. Biochem.* **141**, 675–686 (2007).
15. Diwan, D., Komazaki, S., Suzuki, M., Nemoto, N., Aita, T., Satake, A. & Nishigaki, K. Systematic genome sequence differences among leaf cells within individual trees. *BMC Genomics* **15**, 142 (2014).
16. Sakuma, Y. & Nishigaki, K. Computer prediction of general PCR products based on dynamical solution structures of DNA. *J. Biochem.* **116**, 736–741 (1994).
17. Nishigaki, K., Naimuddin, M. & Hamano, K. Genome profiling: a realistic solution for genotype-based identification of species. *J. Biochem.* **128**, 107–112 (2000).
18. Biyani, M. & Nishigaki, K. Hundred-fold productivity of genome analysis by introduction of micro temperature-gradient gel electrophoresis. *Electrophoresis* **22**, 23–28 (2001).
19. Naimuddin, M., Kurazono, T., Zhang, Y., Watanabe, T., Yamaguchi, M. & Nishigaki, K. Species-identification dots: a potent tool for developing genome microbiology. *Gene* **261**, 243–250 (2000).
20. Yoshida, T., Yamanaka, K., Atsumi, S., Tsumura, H., Sasaki, R., Tomita, K., Ishikawa, E., Ozawa, H., Watanabe, K. & Totsuka, T. A novel hypothyroid 'growth-retarded' mouse derived from Snell's dwarf mouse. *J. Endocrinol.* **142**, 435–446 (1994).
21. Garcia-Vallve, S., Palau, J. & Romeu, A. Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in Escherichia coli and *Bacillus subtilis*. *Mol. Biol. Evol.* **9**, 1125–1134 (1999).
22. Page, R. D. M. TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357–358 (1996).
23. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
24. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
25. Clayton, T. M., Whitaker, J. P. & Maguire, C. N. Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Sci. Int.* **76**, 7–15 (1995).
26. Leclair, B., Frégeau, C. J., Bowen, K. L. & Fourney, R. M. Enhanced kinship analysis and STR-based DNA typing for human identification in mass fatality incidents: the Swissair flight 111 disaster. *J. Forensic Sci.* **49**, JFS2003311 (2004).
27. VandeBerg, J. L., Aivaliotis, M. J., Williams, L. E. & Abee, C. R. Biochemical genetic markers of squirrel monkeys and their use for pedigree validation. *Biochem. Genet.* **28**, 41–56 (1990).
28. Tanyi, M., Olasz, J., Lukács, G., Csuka, O., Tóth, L., Szentirmay, Z., Ress, Z., Barta, Z., Tanyi, J. L. & Damjanovich, L. Pedigree and genetic analysis of a novel mutation carrier patient suffering from hereditary nonpolyposis colorectal cancer. *World J. Gastroenterol.* **12**, 1192–1197 (2006).
29. Phillips, C., Fondevila, M., García-Magariños, M., Rodriguez, A., Salas, A., Carracedo, A. & Lareua, M. V. Resolving relationship tests that show ambiguous STR results using autosomal SNPs as supplementary markers. *Forensic Sci. Int. Genet.* **2**, 198–204 (2008).
30. Donato, M. D., Peters, S. O., Mitchell, S. E., Hussain, T. & Imumorin, I. G. Genotyping-by-Sequencing (GBS): a Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. *PLoS ONE* **8**, e62137 (2013).
31. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., Wei, Z., Wang, K. & Lyon, G. J. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
32. Nishigaki, K., Saito, A., Hasegawa, T. & Naimuddin, M. Whole genome sequence-enabled prediction of sequences performed for random PCR products of *Escherichia coli*. *Nucleic Acids Res.* **28**, 1879–1884 (2000).