relevance of the alternative variants to the protein functions, comprehensive information about the cDNA sequences is indispensable because sometimes protein motifs are embedded over a wide region of the protein sequences, and all of the combinations of the AS exons may not be allowed. Besides, for certain types of subcellular targeting signals, such as signal peptides, the position within the protein sequence is critical. Also, very recent reports, including ours (4), have demonstrated that many loci are subjected to complex patterns of AS in which two distinct genes seemed to be bridged (in which a variant uses exons from two adjacent loci), nested (in which a variant is located inside long intron of another locus) or degenerated (in which two variants use different reading frames in the shared exons. Its another name is multiple CDS): in all three cases, completely unrelated proteins are encoded by a single locus. These cases might not be regarded as alterative splicing in a strict sense. However, when those cases are also considered as extreme cases of functional diversification of a single locus and are subjected to be functional annotations, it is impossible to precisely characterize the combination of the exon usages.

Here, we introduce our new database of AS database, H-DBAS. We constructed this database exclusively using our unique dataset of completely sequenced and carefully annotated full-length cDNAs, which was produced by a human annotation meeting, H-Invitational (5,6). In H-Invitational, 56 419 cDNA sequences of human genes, which were fully sequenced with a sequence reliability higher than 99% [Phred values greater than 30; (7)] and whose potentially problematic sequences such as vectors and polyA tails were precisely trimmed, were subjected to manual annotation of AS variants. These cDNAs were clustered into 24 425 loci and of these, 6877 AS-containing loci, represented by 18 297 AS variants, were identified (4). As a specialized AS database, H-DBAS enables multifaceted analyses from various viewpoints, comprehensively aiming at elucidating functional consequences of widespread AS in human genes. [Note: We will use the word, 'locus', for the transcript cluster for the purpose of simplicity. However, the wording might be reconsidered, having observed highly diverse nature of the human transcriptome. Also see the reference (8)].

DATABASE CONTENTS

Data resources

In H-DBAS, the set of 167 992 so-called H-Invitational cDNAs was used (available from the URL). In addition, an option in which ASs represented by 23 210 RefSeq and 33 411 Ensembl transcripts were also considered is also implemented. In total, 167 564 transcripts were presented in the context of corresponding human genomic information as of UCSC hg17 (http://hgdownload.cse.ucsc.edu/downloads.html#human), cDNA information as of H-Invitational cDNA dataset (Table 1). The mapping and clustering procedures for the cDNAs were followed the annotation pipelines of the H-Invitational cDNAs. For details, see the help page of H-InvDB [http://jbirc.jbic.or.jp/hinv/; (9)].

Data processing

Patterning alternative splicings. Using the positional information for each of the transcripts on the human genome, representative AS patterns were defined for each locus as follows. First, in order to remove possible 5′/3′-end-truncated cDNAs, we excluded cDNAs whose 5′/3′-ends were located inside the second or later exons of any other cDNAs with compatible exon structure in the same locus. We accepted the cDNAs whose 5′/3′-ends were located inside of the first/last exons and considered as variations in the exact transcriptional starting/terminating sites. We also assumed that those cDNAs whose 5′-ends were located outside of the exonic regions of any other clones could not be truncated forms of any known types of transcripts, at least [for further detailed discussion of this subject, see reference (10)]. Second, using the resulting filtered set of putative full-length cDNAs, the genomic position of each exon–intron boundary

Table 1. Statistics of the data processing and of the AS variants and exons identified by genomic structure

	#Locus	#cDNA	#Total exon	#Alternative exon[a]	#Constitutive exon
H-Invitational cDNAs	35 005	167 992	1 164 482[b]	184 649	979 833
Successfully mapped	34 678	167 564	1 164 482	184 649	979 833
≥2 cDNAs per locus	15 445	89 687	795 175	184 649	610 526
Identified AS variants	11 744	74 378	687 841	184 649	503 192
Identified RASVs[c]	11 744	38 664	378 024	98 156	279 868
5′-end	7488	15 920	38 664	15 920	22 744
Internal	10 030	26 443	300 696	69 359	231 337
3′-end	5978	12 877	38 664	12 877	25 787
Retrotransposons[d]	7435	14 534	22 583	12 735	9848
LINEs	3548	5360	6620	3863	2757
SINEs	5849	10 188	14 114	8724	5390
Alu elements	4487	7323	10 240	6379	3861
Identified RASVs[c] including full-length ORF	11 382	30 389	311 409	78 078	233 331
5′-UTR	6660	14 230	26 310	10 238	16 072
CDS	11 382	30 389	272 780	64 270	208 510
3′-UTR	3519	5259	12 319	3570	8749

[a]The number of exons was simply counted in which indicated AS relation was not associated.
[b]Unmapped transcripts' exons could not be counted.
[c]Representative AS Variants.
[d]They were detected by RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org).

was compared with those of the other transcripts belonging to the same locus. For the comparison, a 10 bp allowance was made. If a cDNA had a part of the exonic sequence in the first/last exon inside confirmed intronic regions of the other isoforms, it was regarded as being a '5′/3′-end' AS variant. If a cDNA had a part of an internal exonic sequence inside a confirmed intronic region of other isoforms, it was recognized as being an 'internal' AS variant (4). At this point, we removed annotated genomic rearrangement genes such as Immunoglobulin (Ig) and T-cell receptor (TCR) and anomalistic high polymorphic genes such as Major histocompatibility complex (MHC).

Merging alternative splicing patterns with functional annotations of the encoded proteins. Obtained information of patterns of AS was merged with that of detailed ORF prediction and functional annotation of H-Invitational cDNAs regarding protein motifs, GO terms, predicted subcellular localization signals and transmembrane domains. The protein motif and GO term were identified by InterProScan (11), the subcellular localization was predicted by WoLF PSORT (12) and TargetP (13) and the transmembrane domain was predicted by TMHMM (14) and SOSUI (15). For further details in functional annotation pipeline, see H-InvDB help page (http://jbirc.jbic.or.jp/hinv/). The results of the computational identification and annotation of the AS were visually inspected by the members of the AS annotation team and whenever annotations were considered to be controversial, the caveats were inserted to flag possible annotation errors.

Complex patterns of alternative splicing. Several 'complex' patterns of AS were defined as follows and registered in the database: (i) 'bridged': a locus in which two AS variants were arrayed tandemly without sharing any exons and another transcript 'bridged' these two isoforms, sharing at least some of its exons with both of them; (ii) 'nested': a locus in which CDS region of one AS variant was not shared with another variant and (iii) 'multiple CDS': a locus in which different ORFs >200 bp in length were annotated independently for different AS isoforms sharing at least some of the exons but not sharing any reading frame.

Current statistics

Current statistics of the database are as summarized in Tables 1 and 2 (updated from those presented in the reference (4). In total, 38 664 AS patterns were identified from 11 744 loci. When focused on the consequence of the AS to the encoded amino acid sequences, 30 389 AS variants in 11 382 loci caused changes of 97 amino acids in length on average. Further detailed statistics about how the ASs changed amino acid sequences are presented in 'Statistics' page in the database. Especially, 14 550 AS variants changed the protein motifs. In 14 248 cases, different GO terms were assigned to different AS variants, thus, they could be considered as good targets for further analyzing functional diversification of the genes. Similarly, AS changing subcellular localization signals and transmembrane domains were identified in 17 718 and 3995 AS variants in 5323 and 1248 loci, respectively. As for 'complex' AS, 2336, 3629 and 258 AS variants in 472, 1223 and 101 loci were identified and registered in the database as bridged, nested and multiple CDS, respectively.

Table 2. Numbers of the loci in which AS variants should influence the possible protein functions

	#Locus	#cDNA
AS affecting function total	7630	24 092
Motif-changed	4624	14 550
GO-changed	4150	14 248
Subcellular localization-changed	5323	17 718
Transmembrane domain-changed	1248	3995
Complex AS pattern total	1512	5394
Bridged	472	2336
Nested	1223	3629
Multiple CDS	101	258

ACCESS TO DATABASE

Search system

A simple search form in the top page allows the user to retrieve from within H-DBAS by inputting word(s) of selected categories such as Keyword, HIX (H-Invitational cluster ID), HIT (H-Invitational transcript ID), corresponding Accession/Refseq/Emsembl ID, HUGO gene symbol and definition. In the advanced search form, the user can search the database by more detailed features of AS. The advanced search form consists of three categories: (i) 'Genomic Location' in which the user can specify in which chromosome and where in the chromosome the AS should be searched; (ii) 'AS Structure' in which the user can look for the number of representative AS variants in the locus, particular patterns of AS (such as cassette, internal acceptor, internal donor, mutually exclusive and retained intron) and their locations (5′/3′-end and internal); (iii) 'AS Functional Annotation' in which the user can specify the length difference of encoded protein, protein motifs, GO terms, predicted subcellular localization signals and transmembrane domains invoked by the AS: 'Complex' AS patterns can be also specified here. It is possible to use any combinations of the above search conditions which are within the same or different categories (Figure 1A). For example, the users can perform the search by querying the AS, which should be located on 'chromosome 21', having 'internal' 'cassette' exons, affecting '50–100 amino acids' and 'protein motif'. When multiple entries are hit, the user can see the Result summary and select which should be further examined (Figure 1B). Text-based summarized information can be also selected instead of showing a Java-based dynamic user interface.

AS Viewer

A main part of H-DBAS is a user-friendly Java-based interface, which is subjected to dynamic operations of the user (Figure 2). The browser can be zoomed from the genomic level to the sequence level (genomic/cDNA and amino acid sequences can be viewed). RefSeq and Ensembl transcripts can be viewed together with H-Invitational cDNAs as references. By using the clone view controller, the users can select which items should be viewed. Functional annotation view controllers can be used for selecting which protein motifs identified in the locus should be highlighted/erased. This page is designed so that the user can empirically recognize the positions and patterns of AS in the context of the

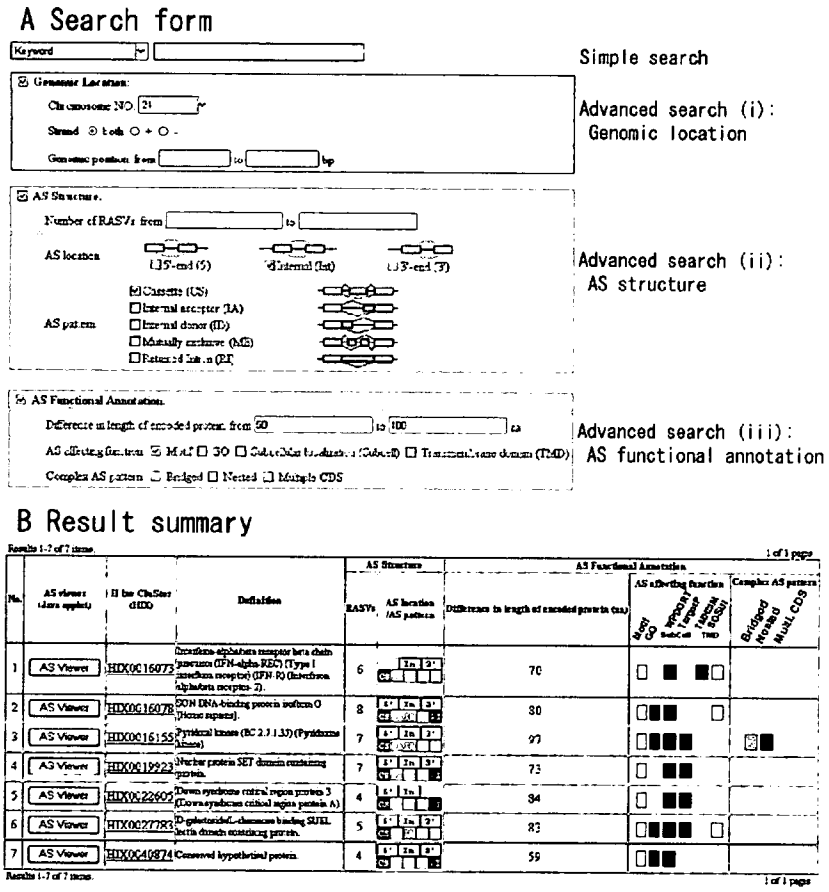


Figure 1. Search system of H-DBAS is shown. (**A**) Search form. H-DBAS has two sorts of search form named simple search and advanced search. The simple search allows the user to find AS locus by using keyword, H-Inv cluster ID (HIX), H-Inv transcript ID (HIT), accession number, RefSeq ID, Ensembl ID, HUGO gene symbol and definition. The advanced search allows the user to find AS locus by using various combinations of three categories such as Genomic location (i), AS structure (ii) and AS functional annotation (iii). In this search system, any combinations including simple search are available. (**B**) Result summary. The result of the query written in text and Figure 1A is shown.

full-length form of each transcript. It should be especially advantageous that the user can view possible influence of the AS on various kinds of protein motifs. When an AS exon in 'Exonic Segment' window is clicked, the positions which are regarded as mutually AS are highlighted in 'AS Event View' window. At the same time, if the protein motifs and transmembrane domains are identified, the corresponding exonic region of the cDNA(s) in 'Entry cDNA' window is colored aqua on the ORF region colored pink.

Example of the search

In Figure 3A, we show an example of AS affecting a motif by using AS Viewer. This is the IKK-related kinase epsilon gene. In this gene, while a cDNA (D63485) contains a prtein motif, 'protein kinase (InterPro ID; IPR000719)', another cDNA (AK093798) does not contain it. The lack of exons 2–7 in the latter cDNA because of cassette type AS is responsible for this putative functional difference. Figure 3B shows an example from complex AS pattern. AJ276409, which is Ssu72-like protein family protein looks as if 'bridging' AK127149 and AK023110 (Figure 3B), both of the latter two transcripts are of known genes and are reported to be protein-coding.

Glossary and download

Use of the database as well as the archives of the raw data is freely available to anonymous public users without any restrictions. A detailed user manual and technical terms used, definitions and parameters for the annotations are precisely described in the 'Glossary' page in H-DBAS. The users can follow the links to further detailed information from each items displayed here. In the 'Download' page, archives of raw data, containing all kinds of AS information and sequence data about all AS variants in our database, are made publicly and anonymously downloadable.

FUTURE DEVELOPMENTS

We are currently interconnecting H-DBAS with H-ANGEL [http://jbirc.jbic.or.jp/hinv/h-angel/; (16)], in which gene expression patterns of the H-Invitational cDNAs are registered. We are also adding precisely annotated mouse

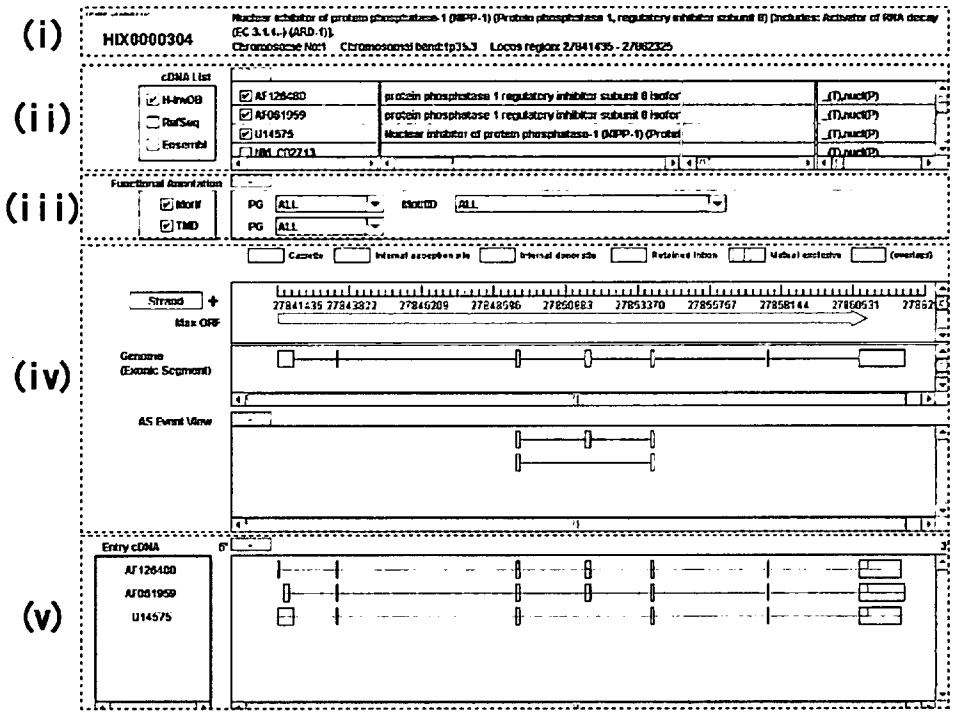


Figure 2. AS Viewer of H-DBAS is shown. Java applet for operating AS event and checking AS exon and protein functions such as protein motif and transmembrane domain. The user can also compare with representative AS variants (RASVs) by nucleic and amino acid sequence level. AS Viewer is separated following parts: (i) Definition and genomic information of the locus; (ii) Selection function and definition of RASVs including RefSeq and Ensembl transcripts as references; (iii) Selection function of protein motif and transmembrane domain; (iv) All exons of selected RASVs are located on genome and AS exons are colored red. By clicking an AS exon in Exonic Segment field, the AS events on the location are shown in AS Event Viewer. Max ORF means total ORF range on genome of selected RASVs; (v) Selection function of AS structure on genome and on cDNA. Selected RASVs' structures are shown and these ORFs are colored pink and protein motifs and transmembrane domains are colored aqua. They are also shown nucleic and amino acid sequences by using zoom function.

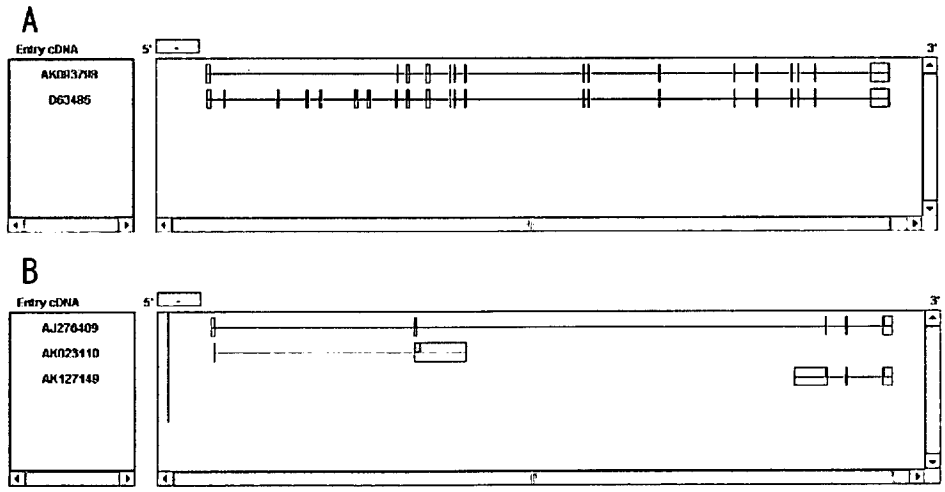


Figure 3. Examples of the alternative splicing affecting motif (**A**) and bridged complex AS pattern (**B**) from AS Viewer in H-DBAS. Exons and introns are represented by boxes and lines. ORF region is colored pink and protein motif region is colored aqua.

full-length cDNA information as well and developing comparative genomics interfaces. The upcoming two major categories of extensive data will allow us to start determining how the ASs were acquired during evolution and how they fulfill the functional diversification of a single locus in various cellular circumstances. Furthermore, in the phase of further detailed experimental validation of the AS, H-DBAS should serve as an important interface for looking for cDNA clone resources, as the H-DBAS represents physical full-length cDNAs, which should serve as

indispensable reagents for many kinds of experimental purposes.

Finally, we realize that we have a long way ahead for improving the web-page and database contents. We sincerely welcome any feedbacks from the users.

ACKNOWLEDGEMENTS

We thank Y. Fujii, Y. Sato, T. Habara, H. Nakaoka, F. Todokoro, Y. Imamizu, M. Ogawa and C. Yamasaki for genome mapping, ORF prediction and functional annotation of the H-Invitational cDNA dataset. We are grateful to C.Gough for critical reading of the manuscript. This research was financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC). Funding to pay the Open Access publication charges for this article was provided by JBIC.

Conflict of interest statement. None declared.

REFERENCES

1. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
2. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
3. Lee,C., Atanelov,L., Modrek,B. and Xing,Y. (2003) ASAP: the alternative splicing annotation project. *Nucleic Acids Res.*, **31**, 101–105.
4. Takeda,J., Suzuki,Y., Nakao,M., Barrero,R.A., Koyanagi,K.O., Jin,L., Motono,C., Hata,H., Isogai,T., Nagai,K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56,419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.*, **34**, 3917–3928.
5. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
6. Nakao,M., Barrero,R.A., Mukai,Y., Motono,C., Suwa,M. and Nakai,K. (2005) Large-scale analysis of human alternative protein isoforms: pattern classification and correlation with subcellular localization signals. *Nucleic Acids Res.*, **33**, 2355–2363.
7. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
8. Suzuki,M. and Hayashizaki,Y. (2004) Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *Bioessays*, **26**, 833–843.
9. Yamasaki,C., Koyanagi,K.O., Fujii,Y., Itoh,T., Barrero,R., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M., Takeda,J., Fukuchi,S. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
10. Kimura,K., Wakamatsu,A., Suzuki,Y., Ota,T., Nishikawa,T., Yamashita,R., Yamamoto,J., Sekine,M., Tsuritani,K., Wakaguri,H. *et al.* (2006) Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
11. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
12. Horton,P., Park,K.-J., Obayashi,T. and Nakai,K. (2006) Protein subcellular localization prediction with WoLF PSORT. *The 4th Annual Asia Pacific Bioinformatics Conference APBC06*, 39–48.
13. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
14. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
15. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
16. Tanino,M., Debily,M.A., Tamura,T., Hishiki,T., Ogasawara,O., Murakawa,K., Kawamoto,S., Itoh,K., Watanabe,S., de Souza,S.J. *et al.* (2005) The human anatomic gene expression library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**, D567–D572.

第2章 ヒトゲノム解析と疾患

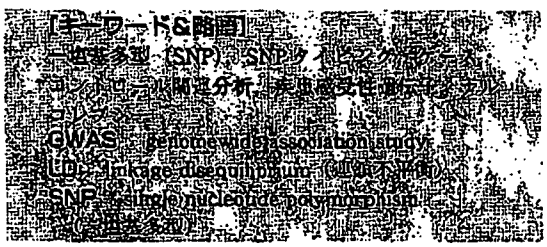
1. 大規模SNPタイピングによる多因子疾患遺伝子の探索

西田奈央，徳永勝士

ヒトゲノム計画をはじめとする遺伝情報解析の成果としてデータベースに蓄積された600万種を超える一塩基多型（SNP）のうち，約50万種のSNPを同時にタイピングすることのできる手法が近年になって実用化された．われわれは，500K Human Mapping Array Set（Affymetrix社）によるSNPタイピングを効率的に行うためのシステムを構築し，大規模なケース・コントロール関連分析を行っている．一例として，睡眠障害の1つであるナルコレプシーの疾患感受性遺伝子の探索研究を紹介する．

はじめに

現在，ヒトのさまざまな多因子疾患に関わる遺伝子を探索する戦略として最も注目を浴びているのが，本稿で紹介するゲノムワイド関連分析法である．例を挙げると，米国NIH（国立衛生研究所）はGWAS（genomewide association study）計画を提案し，いくつかのcommon diseasesについて大規模な研究チームを公募した．また，大規模な疫学研究として知られるFramingham Heart Study（フラミンガム研究）で収集された試料のうち9,000検体について，本稿で紹介する技術を用いて解析することにより，心，肺，血液，睡眠疾患に関与する遺伝子変異を探索する計画も発表された．



このような動向の背景には，多因子疾患感受性遺伝子を探索する統計遺伝学的手法の選択がある．大別して連鎖分析（linkage analysis）と関連分析（association analysis）があり，従来のゲノムワイド探索のほとんどは連鎖分析法が用いられてきた．多数の成功例があるものの，連鎖分析のみで疾患感受性遺伝子のほとんどを同定するのは困難であると考えられるようになり，また後述する大規模なSNP解析技術の進展も相まって，ゲノムワイド関連分析が大きく取り上げられることとなった．そもそも連鎖分析は患者家族を対象として文字通り疾患遺伝子と多型マーカーの連鎖を検出する手法であり，ゲノム全域に分布する1万から数万種のSNPを用いればよい．一方，関連分析の代表であるケース・コントロール関連分析法[注1]は，非血縁の患者集団と健常者集団を対象として疾患遺伝子と多型マーカーの連鎖不平衡（LD）を検出する手法であり，ゲノム全域に分布する数十万種以上のSNPを解析することが必要となる[1]（統計遺伝学手法の原理ならびに集団遺伝学的基礎については別冊に解説した[2]）．

Search for susceptibility genes to multifactorial diseases by large-scale SNP typing
Nao Nishida/Katsushi Tokunaga：Department of Human Genetics, Graduate School of Medicine, University of Tokyo（東京大学大学院医学系研究科人類遺伝学分野）

ヒトゲノム上には，マイナーアレル頻度が10％以上のSNPsが500万種，マイナーアレル頻度が1％以上のSNPsとなると1,100万種も存在すると試算されており[3]，これらのSNPsを大規模にかつ正確にタイピングのできる手法の確立が求められてきた．

SNPタイピング法として最もよく知られている方法は，個々の多型部位を含むゲノム断片を特異的にPCR増幅した後にアレルを識別する方法である[4)~7)]．これらの方法では，1,000種程度のSNPのタイピングであれば，PCRプライマーはじめ各種試薬にかかるコストを考えて実用可能であるといえるが，数千，数万種を超える数のSNPをタイピングすることは困難である．一方，近年になって多型部位特異的なPCRを行わずに大規模なSNPタイピングを行う方法が実用化された[8) 9)]．その一つであるAffymetrix社によって確立された方法では，まず制限酵素反応でゲノムDNAの断片化を行い，続いてそれら断片の両端にアダプター配列を付加し，まとめて増幅した後にマイクロアレイを用いたアレル特異的なハイブリダイゼーションを行う[8)]．現在では，この手法を用いて50万種を超えるSNPを同時にタイピングするキットが市販されている．われわれは500K Human Mapping Array Setを用いる大規模なSNP解析システムを構築し，いくつかの多因子疾患についてゲノムワイド関連分析を実施している．その一例としてナルコレプシーの疾患感受性遺伝子の探索研究を紹介する．

1 技術原理

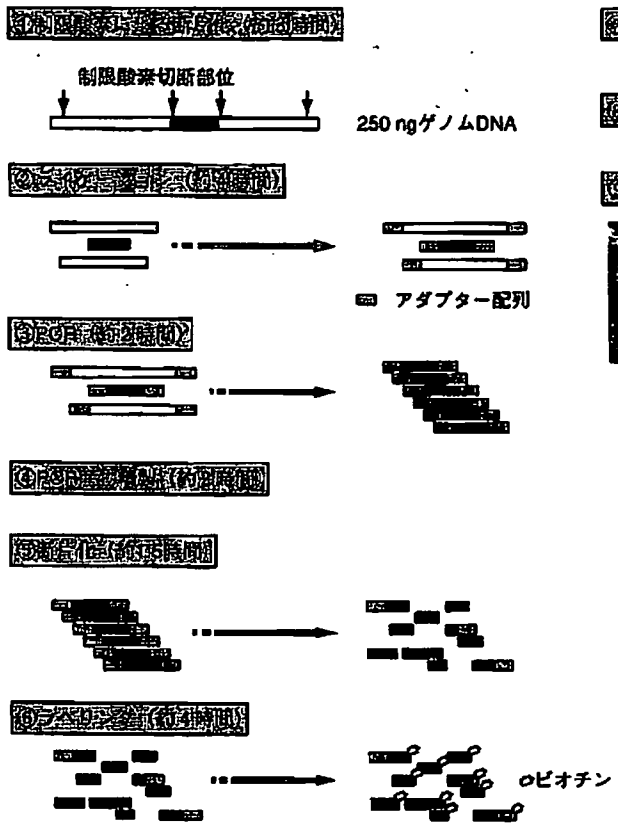
500K Human Mapping Array Set（以下，500K Array Set）は，制限酵素によるゲノムDNAの断片化とマイクロアレイによるタイピングの手法に改良を加えることにより，大規模なタイピングを行える手法として確立された[8)]．解析対象となるSNPは，公共のSNPデータベースおよびPerlegen社に登録されている約220万種のSNPから遺伝学的情報量が最大化されるように，また連鎖不平衡（LD）やHapMapプロジェクトからの情報も考慮して選択された約50万種のSNPである．

500K Array SetによるSNPタイピングは，ゲノムの複雑さを低減しマイクロアレイへのハイブリダイゼーション効率を上げるための酵素反応ステップと，洗浄・染色装置およびマイクロアレイ用スキャナーを用いた検出ステップに分けることができる（図1）．50万種のSNPタイピングは，2種類の制限酵素（*Sty*Ⅰ，*Nsp*Ⅰ）を用いてそれぞれ約25万種のSNPを独立にタイピングすることで実現される．制限酵素によるゲノムDNAの断片化を行った後，断片化されたゲノムDNAの両末端にアダプター配列をライゲーション反応により付加する．アダプター配列は，続くPCRで使用されるプライマーと相同な配列をもち，また制限酵素認識配列を突出端としてもつ二本鎖DNAである．2種類の制限酵素（*Sty*Ⅰ，*Nsp*Ⅰ）のそれぞれに対して用意されるアダプター配列は，制限酵素認識配列を除いて共通の配列をもっているので共通のプライマーを使用してPCRを行うことができる．PCRでは，目的の長さをもったゲノムDNA断片（200～1,100 bp）だけが選択的に増幅される．ここまでの酵素反応により，もともと30億塩基対のゲノムDNAが5億塩基対程度のPCR混合産物となる．マイクロアレイへの効率的なハイブリダイゼーションには，ゲノムの複雑さを低減することが大きな役割を果たすと考えられている[10) 11)]．続いて，PCR産物の精製を行った後，DNaseⅠ制限酵素によりPCR産物の断片化を行う．断片化されたPCR産物は平均長で180 bp以下となる．マイクロアレイへの効率的なハイブリダイゼーションには，ゲノムの複雑さを低減することに加えてPCR産物の断片化が重要になる．最後にterminal deoxynucleotidyl transferase酵素反応により，断片化されたPCR産物の末端にビオチンを導入する．

続いて，専用のマイクロアレイを用いてハイブリダイゼーションを行う．マイクロアレイには解析対象となる各SNPに対して合計24本のプローブが用意されている．プローブは25塩基長のオリゴDNAで，SNP部位を含む塩基配列をもっている．2種類のアレルに対して完全に相補的な塩基配列をもつプローブ（PMプローブ）と1塩基のミスマッチを含むプローブ（MMプローブ）を用意し，4種類のプローブを1組のプローブセットとしている．SNP部位を25塩基長のプローブの中心に置いたプローブセットを基本として，SNP

※1 ケース・コントロール関連分析

ある疾患に罹っている患者群（ケース）と健常者群（コントロール）とで遺伝子・ゲノム多型の頻度に差があるかどうかを検定することにより，疾患関連遺伝子を探索するための統計遺伝学的方法．



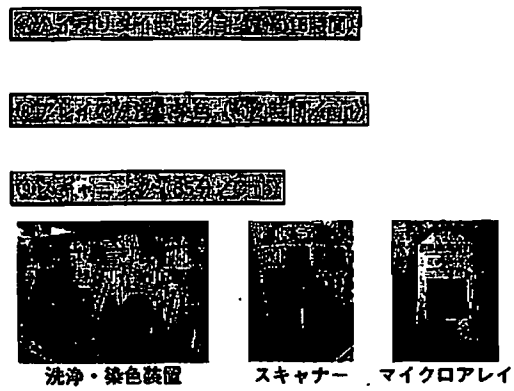


図1　500K Human Mapping Array Setによる SNPタイピングの原理
制限酵素による断片化からラベリングまでの酵素反応で得られた産物を洗浄・染色装置およびマイクロアレイ用スキャナーで検出することにより，約50万種のSNPのタイピングを行う

部位を中心から4塩基上流（＋4）にずらしたプローブセットから4塩基下流（－4）にずらしたプローブセットまで7組のプローブセット（－4，－2，－1，0，＋1，＋3，＋4）の中から3組のプローブセットを選択する．3組のプローブセットはゲノムDNAの両鎖に対して用意されているので，合計6組のプローブセットが各SNPに対して用意されている．また，遺伝学的に重要だとされる約52,000 SNPsに関してはプローブセットを合計10組とし，計40本のプローブを用意している．すべてのプローブセットからのPMプローブとMMプローブのシグナル強度差を検出することで遺伝子型の判定を行うことができる．

マイクロアレイへのハイブリダイゼーションが終了した後，洗浄・染色装置を用いてマイクロアレイの洗浄および蛍光染色を行う．蛍光染色は，蛍光分子で標識されたストレプトアビジンが，前述のビオチン導入されたPCR断片に結合することにより行われる．また，洗浄・染色装置内ではビオチン修飾された抗ストレプトアビジン抗体を用いてシグナルの増強が行われる．最後に蛍光染色されたマイクロアレイを専用のスキャナーで画像データとして読み取り，続いて専用の画像解析ソフトウェアを用いて各SNPの遺伝子型を決定する．

2 システム構築

1）ハードウェアの整備

500K Array SetによるSNPタイピングを効率的に行うために，環境，装置を整備し，作業マニュアルを作成した．まず，ゲノムDNAへのPCR産物のコンタミネーションを防ぐために，試料調製室とSNP解析室を設けた．試料調製室にはゲノムDNAを保管し，PCR以前の酵素反応を行うのに必要な装置（サーマルサイクラーなど）を用意した．制限酵素による断片化からPCRの反応溶液の調製までは試料調製室で行い，PCR以降の酵素反応はSNP解析室で行った．また，3台の洗浄・染色装置を用意し，1回のランで計

12枚のマイクロアレイを洗浄・染色することができるようにした．すべてのマイクロアレイはバーコードで管理され，洗浄・染色が終了したマイクロアレイはオートローダー付きのマイクロアレイ用スキャナーに装填され画像データが読み込まれる．オートローダー付きのマイクロアレイ用スキャナーは計64枚のマイクロアレイを装填することができ，バーコードを参照しながらすべてのマイクロアレイの画像データを自動的に読み込むことができる．

SNPタイピング作業のルーチン化にあたって，制限酵素による断片化からラベリングまでの5つの酵素反応ステップで使用するすべてのマイクロタイタープレートをバーコードで管理する（図2）．また，1枚のマイクロタイタープレートで32検体分の酵素反応を行うこととし，制限酵素による断片化で使用するマイクロタイタープレートのウェル位置をサンプルと対応させることでサンプルのID化を行った．制限酵素による断片化以降の酵素反応で使用するマイクロタイタープレートでもレイアウトを変えずに酵素反応を行うことで，ウェル位置をサンプルIDとして解析結果を得ることができる．また，酵素反応の各工程を管理するためにチェックシートを作成し，反応工程の進行を随時チェックシートで確認しながら進める．PCRおよび断片化の酵素反応の後にはアガロースゲル電気泳動を行い，PCR産物および断片化産物の平均長がそれぞれ200～1,100 bp，180 bp以下となっていることを確認する．加えて，同一のマイクロプレート上で酵素反応を行った32検体のうちから4検体だけを先行してハイブリダイゼーションを行い，SNPコール率に問題がないことを確認した後で残り28検体のハイブリダイゼーションを行った．28検体のハイブリダイゼーションを行う際に，次のマイクロタイタープレートから4検体を加えて合計32検体のハイブリダイゼーションを順次行っていくこととした．

2）ソフトウェアの開発

500K Array SetによるSNPタイピングではGeneChip® Operating Software (GCOS) とGeneChip® Genotyping Analysis Software (GTYPE) という2種類のソフトウェア（ともにAffymetrix社）を使用する．GCOSソフトウェアは洗浄・染色装置およびマイクロアレイ用スキャナーを操作する際に使用し，またGTYPEソフトウェアはマイクロアレイの画像データから遺伝子型を判定する際に使用する．GTYPEソフトウェアで決定された約25万種のSNPの遺伝子型は，*Sty*I，*Nsp*Iごとにテキストファイルとして転送することができる．

われわれは500K Array SetによるSNPタイピングから得られる約50万SNPsの遺伝子型情報を用いてケース・コントロール関連分析を行うためのソフトウェアを開発した．ケース・コントロール関連分析をするにあたって，*Sty*IおよびNspIごとにまとめられた約25万SNPsの遺伝子型データを検体ごとに統合し，さらに検体をケース群およびコントロール群に分けて新たなテキストファイルとして作成する機能をソフトウェアに加えた．続いて，作成したケース群およびコントロール群の解析結果のテキストファイルを使ってケース・コントロール関連解分析を行った．この際，各コントロール群における各遺伝子型の観察数からも，ハーディー・ワインベルク平衡[※2]の検定を行うこととした．ケース・コントロール関連分析の結果はレポートファイルとしてまとめられ，専用のビューアーを用いて表示することができる．

3 ナルコレプシー感受性領域のゲノムワイド探索

われわれは上に述べた大規模SNPタイピングシステムを用いて，文部科学省科学研究費特定領域研究「基盤ゲノム」におけるSNPタイピングセンターとして，数種の疾患のゲノムワイド関連分析を実施しており，すでに2種の疾患については約200名ずつの患者試料と約200名の健常者試料の解析を終了している．

またわれわれは睡眠障害の1つナルコレプシーの感受性・抵抗性遺伝子をゲノムワイド関連分析法によって探索している．すでに2万3千種のマイクロサテライト多型を用いたゲノムワイド関連解析を行い，11カ所の候補領域を検出するとともに，その1つから新たな疾患抵抗性遺伝子を同定した[13]．現在，他の候補領域についても詳細な解析を行っているが，これと平行して，新たに50万種のSNPsを用いたゲノムワイド関

※2 ハーディー・ワインベルク平衡

自然淘汰が働かず，突然変異によって新たな対立遺伝子が生じず，また移住や混血などが起こらない十分に大きな規模の集団においては，対立遺伝子の頻度は世代を経ても変化しないという集団遺伝学の基本的法則．

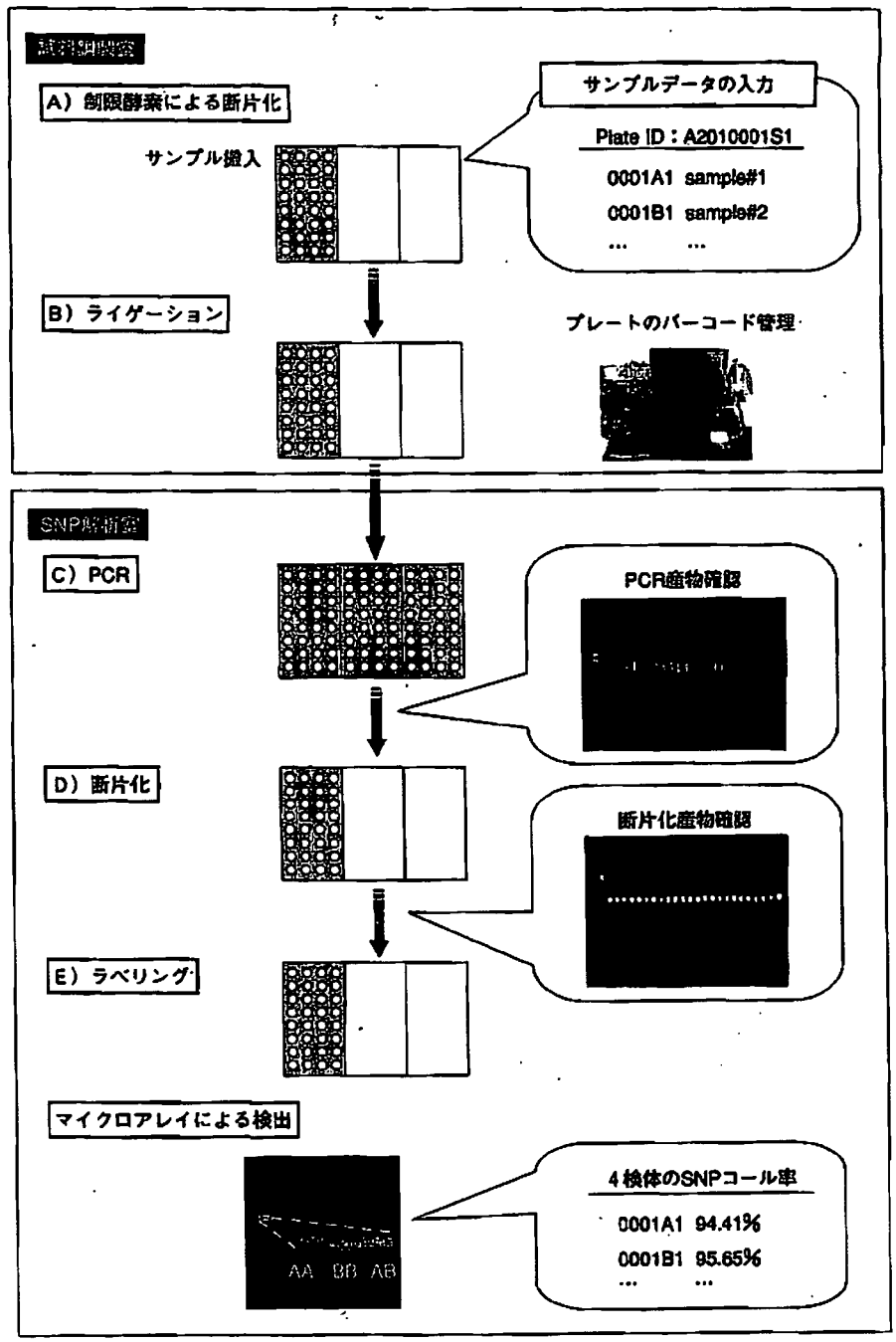


図2　SNPタイピングシステムの構築
500K Human Mapping Array Setを用いたSNPタイピングを効率的に行うためのシステム環境を構築し，作業マニュアルを作成した．96ウェルマイクロタイタープレートを使用して32検体ずつ解析を行う

表1　50万SNPsによるヒトナルコレプシーのゲノムワイド関連解析：関連SNP数の分布

有意水準	SNP数
$P<0.0001$	214
$p<0.001$	631
$P<0.01$	3,665
$P<0.05$	15,443
解析SNP総数	335,811

連分析も開始している．すでにタイピングおよび統計解析を終了した110名の患者試料と200名の健常者試料のデータから得られた関連SNPsの数の分布を表1に示す．call rate（＞90％）およびハーディ・ワインベルク平衡からのずれ（$p>0.001$）に基づいて選別された約33万6千種のSNPsのうち，$p<0.0001$の関連を示したSNPsが約200種検出された．なお，この解析には従来の遺伝子型判定ソフト（GTYPE4.0）を用いたデータを使用している．現在，新しい判定ソフト（GTYPE4.1）によってcall rateが向上していることから，より多くのSNPsについて統計解析することが可能になっている．ヒトのナルコレプシーについては，すでに確立した感受性遺伝子として6番染色体上の*HLA-DQB1*遺伝子が知られている．図3は今回の解析から*HLA*遺伝子領域について得られた結果であるが，予想通り*HLA-DQB1*遺伝子近傍をピークとする強い関連が認められた．現在われわれは，解析規模を2倍に拡大して新たなナルコレプシー感受性・抵抗性候補領域を検出している．

おわりに

数十万種以上のSNPを一挙にタイピングできる技術の実用化によって，従来は存在しなかった広範かつ詳細なヒトゲノム多型情報が得られる時代となった．このような情報は，疾患遺伝子探索研究のみならず，人類の進化や人類集団の歴史を解明する糸口を提供し，ヒトゲノム多様性に関連するさまざまな研究分野に画期的なインパクトを及ぼすことは疑いない．

しかしながら同時に，われわれはまだ得られる膨大な多型情報を十分に活用できるノウハウをもっていないことも指摘しておきたい．500K Array Setによる SNPタイピングで得られる1検体当たりのファイルデータのサイズは約2 Gbであるため，何百，何千検体のデータを保管し，必要な時に取り出して解析するためのコンピュータ環境を整備することは容易ではない．また，われわれの統計解析ソフトウェアはおのおのの SNPについて関連分析できるものの，まだSNPハプロタイプについて関連分析することはできない．市販

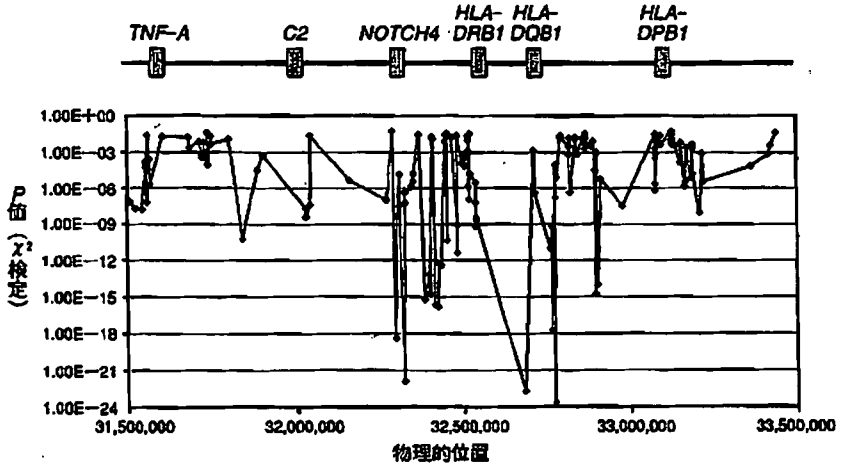


図3　ヒトナルコレプシーのゲノムワイド関連解析から
HLA-DQB1，*HLA-DRB1*遺伝子近傍領域におけるSNP関連マッピングによって強い疾患関連が検出された

のソフトウェアにも，限定した領域でハプロタイプ関連分析できるものはあるものの，ゲノム全域にわたって一挙に分析できるものはない．さらに，多数の検体について得られた50万SNPsデータから，これまで全く知られていなかった新しい遺伝子－遺伝子相互作用が見出される可能性がある．しかし残念ながら，従来の統計学的手法や計算アルゴリズムでは，このように膨大なデータを実用的に処理できない．このように，ゲノムワイド多型解析情報はバイオインフォマティクスに関わるさまざまな研究者にとって挑戦に値する多くの課題を提供してくれるとともに，その達成によって従来にない実り豊かな成果をもたらしてくれるに違いない．

文献

1）Ohashi, J. & Tokunaga, K. : J. Hum. Genet., 46 : 478-482, 2001
2）徳永勝士：「人類遺伝学ノート」，南山堂（印刷中）
3）Kruglyak, L. & Nickerson, D. A. : Nat. Genet., 27 : 234-236, 2001
4）Syvänen, A.-C. : Nat. Rev. Genet., 2 : 930-942, 2001
5）Kwok, P.-Y. & Chen, X. : Curr. Issues Mol. Biol., 5 : 43-60, 2003
6）Nishida, N. et al. : Anal. Biochem., 346 : 281-288, 2005
7）Rachagani, S. et al. : BMC Genetics, 7 : 31, 2006
8）Matsuzaki, H. et al. : Genome Research, 14 : 414-425, 2004
9）Oliphant, A. et al. : BioTechniques, 32 : S56-S61, 2002
10）Grant, S. F. et al. : Nucleic Acids Res., 30 : e125, 2002
11）Jordan, B. et al. : Proc. Natl. Acad. Sci. USA, 99 : 2942-2947, 2002
12）Kawashima, M. et al. : Am. J. Hum. Genet., 79 : 252-263, 2006

<著者プロフィール>

西田奈央：東京大学大学院総合文化研究科で博士号を取得後，東京大学大学院医学系研究科人類遺伝学分野（徳永勝士教授）にて研究員として従事．研究課題は遺伝子多型解析手法の開発．

徳永勝士：東京大学理学部，同医学部附属病院，日本赤十字中央血液センターを経て，1995年より東京大学大学院医学系研究科教授．研究課題はヒトゲノム多様性および複合疾患の遺伝要因とその機能．

Available online at www.sciencedirect.com
ScienceDirect
Analytical Biochemistry 364 (2007) 78–85
ANALYTICAL BIOCHEMISTRY
www.elsevier.com/locate/yabio

Further development of multiplex single nucleotide polymorphism typing method, the DigiTag2 assay

Nao Nishida [a,*], Tetsuya Tanabe [b], Miwa Takasu [a], Akira Suyama [c], Katsushi Tokunaga [a]

[a] *Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan*
[b] *Bio Business Division, Olympus Corporation, Hachioji, Tokyo 192-8512, Japan*
[c] *Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Meguro-ku, Tokyo 153-8902, Japan*

Received 21 December 2006
Available online 13 February 2007

Abstract

A number of single nucleotide polymorphisms (SNPs) are considered to be candidate susceptibility or resistance genetic factors for multifactorial disease. Genome-wide searches for disease susceptibility regions followed by high-resolution mapping of primary genes require cost-effective and highly reliable technology. To accomplish successful and low-cost typing for candidate SNPs, new technologies must be developed. We previously reported a multiplex SNP typing method, designated the DigiTag assay, that has the potential to analyze nearly any SNP with high accuracy and reproducibility. However, the DigiTag assay requires multiple washing steps in manipulation and uses genotyping probes modified with biotin for each target SNP. Here we describe the next version of the assay, DigiTag2, which works with simple protocols and uses unmodified genotyping probes. We investigated the feasibility of the DigiTag2 assay by genotyping 96 target SNPs spanning a 610-kb region of human chromosome 5. The DigiTag2 assay is suitable for genotyping an intermediate number of SNPs (tens to hundreds of sites) with a high conversion rate (> 90%), high accuracy, and low cost.

Keywords: Multiplex genotyping; SNPs; Mutation; Oligonucleotide ligation assay

As a consequence of the Human Genome Project and single nucleotide polymorphism (SNP)[1] discovery projects, several million SNPs have been uploaded onto public SNP databases. It is estimated that there are 5 million SNPs with a greater than 10% minor allele frequency and 11 million SNPs with a greater than 1% minor allele frequency in the human genome [1]. Among these SNPs, many are candidate susceptibility or resistance genetic factors for multifactorial diseases and have been identified based on linkage analysis in families or association analysis with unrelated patients (cases) and healthy controls [2–6]. Large-scale case–control analyses using a dense set of SNP markers across the human genome have revealed associations between various diseases and SNPs with the highest detection power [7–9].

During recent years, genome-wide association studies using SNP markers have attempted to search for susceptibility and/or resistance genes by using emerging genome-wide SNP typing technologies such as Affymetrix GeneChip arrays and Illumina BeadArray genotyping technology [10–13]. These genome-wide SNP typing technologies would detect candidate regions, including susceptibility or resistance genes. However, to identify primary SNPs or genes, it is necessary to perform association analysis using an intermediate number of SNPs (tens to hundreds of sites) located within the candidate regions. Currently, there are a variety of SNP genotyping methods that are suitable for genotyping large numbers of samples for a modest number of SNPs such as 5′ exonuclease

* Corresponding author. Fax: +81 3 5802 8619.
E-mail address: nishida-75@umin.ac.jp (N. Nishida).

[1] *Abbreviations used:* SNP, single nucleotide polymorphism; MALDI–TOF MS, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry; ED, end digit; D1, first digit; PCR, polymerase chain reaction; dNTP, deoxynucleoside triphosphate; ATP, adenosine triphosphate; DTT, dithiothreitol; NAD, nicotinomide adenosine dinucleotide; EDTA, ethylenediaminetetraacetic acid; Cy3–ED-1, Cy3-labeled ED-1; Cy5–ED-2, Cy5-labeled ED-2; SDS, sodium dodecyl sulfate; DCN, DNA coded number.

0003-2697/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ab.2007.02.005

fluorescence-based assay (TaqMan) [14], pyrosequencing [15], single-base extension [16], matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI–TOF MS) [17,18], and SNPlex assay [19]. However, many applications need to select relevant SNPs for their assay by in silico assay design, and some candidate SNPs are then excluded from investigation. Moreover, it is difficult or impossible for some assays to perform multiplex SNP genotyping.

To accomplish successful SNP typing for all candidate SNPs at low cost, new technologies must be developed. We previously reported a multiplex SNP typing method, designated the DigiTag assay, that has a high conversion rate (>90%) and reliable accuracy [20]. However, the DigiTag assay requires improvement with regard to simplifying assay protocols and reducing assay cost. In this study, we developed the DigiTag2 assay, which has simplified assay protocols, and performed typing for 96 SNP sites located in a 610-kb region on human chromosome 5 using 48 individual genomic DNA samples.

Materials and methods

DNA samples

Genomic DNA samples from 48 unrelated healthy donors were obtained from the Japan Health Science Foundation (Osaka, Japan). All donors provided written informed consent, and samples were anonymized. For each sample, 1 μg of purified genomic DNA was dissolved in 20 μl of TE buffer (pH 8.0, Wako, Osaka, Japan) for use and was stored at −20 °C.

End digits and first digits

We designed the end digits (EDs) and first digits (D1s) to be 23-mer oligonucleotides and attached the EDs and D1s to 5′ query probes and 3′ query probes, respectively. We prepared two EDs (ED-1 and ED-2) for two alleles at each SNP. All EDs and D1s are used for the priming site in the labeling step, and D1s are also used as probes that are attached to DNA microarray in the detection step. The EDs and D1s have the following properties: (i) uniform melting temperature (58.8 ± 1.0 °C) and length, (ii) specific hybridization only to complementary EDs and D1s, (iii) minimal interaction with other EDs and D1s, and (iv) no formation of secondary structures [21]. These properties ensure uniform polymerase chain reaction (PCR) efficiency, even if all of the EDs and D1s are used in multiplex PCR. Furthermore, precise hybridization on DNA microarray is possible using a set of D1s with high reproducibility. Sequence information for EDs and D1s is listed in Supplementary Table 1.

Multiplex PCR from sample DNA

We designed multiplex PCR primers for each of the 96 SNP sites to have relatively long length (average length 40-mer) and to give PCR products of between 181 and 798 bp (average length 527 bp). Sequence information for the multiplex PCR primers is listed in Supplementary Table 2.

We performed multiplex PCR using a two-step protocol (denature and extension steps) with a 6-min extension step using specifically designed primer pairs. Multiplex PCR was performed with 2.5 μl genomic DNA and 250 fmol of each primer for 96 SNP sites in 10 μl of 2× Qiagen Multiplex PCR Master Mix containing HotStarTaq DNA polymerase, multiplex PCR buffer and deoxynucleoside triphosphate (dNTP) mix (Qiagen Multiplex PCR Kit, Qiagen, Valencia, CA, USA). Cycling was performed using a Bio-Rad PTC-200 Peltier thermal cycler (Bio-Rad, Hercules, CA, USA) as follows: 95 °C for 15 min, followed by 40 cycles of 95 °C for 30 s and 68 °C for 6 min. When necessary, fragment length of the 96 PCR products was confirmed by capillary electrophoresis (Agilent 2100 Bioanalyzer, Agilent, Palo Alto, CA, USA) to evaluate PCR efficiency.

Encoding reaction

We performed multiplex oligonucleotide ligation assay using the multiplex PCR products as targets. For 96-plex oligonucleotide ligation assay, we prepared mismatch-induced 5′ query probes for 91 target SNPs and perfect match 5′ query probes for 5 target SNPs (SNP 7, SNP 9, SNP 18, SNP 49, and SNP 93). The assignment of D1s to the SNPs analyzed in this study and sequence information for the probes are listed in Supplementary Table 3.

Prior to the encoding reaction, 96 unmodified 3′ query probes were simultaneously phosphorylated at the 5′ end in 40 μl of 1× protruding end kinase buffer containing 30 mM adenosine triphosphate (ATP), 40 U polynucleotide kinase, and 4 pmol of 3′ query probes for 96 SNP sites (Kination Kit, Toyobo, Osaka, Japan). The reaction mixture was incubated for 30 min at 37 °C and for 3 min at 95 °C using a Bio-Rad PTC-200 Peltier thermal cycler. The encoding reaction was prepared by mixing 1 μl of multiplex PCR products in 15 μl of *Taq* DNA ligase buffer containing 20 mM Tris–HCl (pH 7.6), 25 mM potassium acetate, 10 mM magnesium acetate, 10 mM dithiothreitol (DTT), 1 mM nicotinomide adenosine dinucleotide (NAD), and 0.1% Triton X-100 (New England Biolabs, Beverly, MA, USA) with 10 fmol of probes (192 5′ query probes and 96 phosphorylated 3′ query probes) and 10 U *Taq* DNA ligase. All components of the encoding reaction were mixed on ice. The encoding reaction initially was held at 95 °C for 5 min, followed by 58 °C for 15 min using a Bio-Rad PTC-200 Peltier thermal cycler. The reaction was stopped by holding the temperature at 10 °C.

Labeling reaction

For the labeling reaction, 6 μl of ligation products was directly mixed in 12 μl of *Ex Taq* buffer containing 20 mM Tris–HCl (pH 8.0), 100 mM KCl, 0.1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM DTT, 0.5% Tween 20, 0.5%

Nonidet P-40, 50% glycerol, and 2 mM each dNTP (TaKaRa, Shiga, Japan) with 6.0 pmol of Cy3-labeled ED-1 (Cy3–ED-1), 6.0 pmol of Cy5-labeled ED-2 (Cy5–ED-2), 30 fmol each of the 96 D1s, and 1.5 U *Ex Taq* polymerase. The reaction initially was incubated at 95 °C for 1 min, followed by 30 cycles of 95 °C for 30 s, 55 °C for 6 min, and 72 °C for 30 s, using a Bio-Rad PTC-200 Peltier thermal cycler. The reaction was stopped by holding the temperature at 10 °C.

Hybridization and detection on DNA microarray

We purchased a DNA microarray (NovusGene, Tokyo, Japan) that had 24 separated areas on the same slide glass. Each of the separated areas contained 100 types of oligonucleotide probe (96 probes for 96 SNPs and 4 probes for validation controls of the assay) identical to D1 sequences. Of the 4 validation control probes, 3 were not used in the DigiTag2 assay because these probes were prepared to validate the washing step with magnetic beads in the previous version of the DigiTag assay. The ready-to-use DNA microarrays were stored in a desiccator at room temperature until use.

A hybridization mixture was prepared by mixing 5 μl of labeled products in 12 μl of hybridization buffer containing 0.5× SSC, 0.1% sodium dodecyl sulfate (SDS), 15% formamide, and 1 mM EDTA with 1 μl of hybridization control. The hybridization control was prepared with 2.5 fmol of Cy3-labeled D1_100 and Cy5-labeled D1_100. Then 8 μl of the hybridization mixture was applied to each area on the DNA microarray. Hybridization was carried out for 30 min at 37 °C in a hybridization oven (ThermoStat plus, Eppendorf, Hamburg, Germany). After hybridization, DNA microarrays were washed in washing buffer (0.1× SSC and 0.1% SDS) with shaking at 60 rpm for 5 min. DNA microarrays were consecutively washed in distilled water with shaking at 60 rpm for 1 min and were then dried by centrifugation at 2000 rpm for 1 min. Hybridization images were scanned at photomultiplier voltages of 400 V for Cy3 and 480 V for Cy5 using a commercially available DNA chip scanner, and fluorescence image analysis was performed using commercially available software (GenePix 4000B unit and GenePix Pro 4.1 software package, Axon Instruments, Foster City, CA, USA). The genotype calls were determined using the SNPStar software (version 0.0.0.8, Olympus, Tokyo, Japan).

Results and discussion

DigiTag2 assay scheme

We previously reported a multiplex SNP typing method, the DigiTag assay, in which all of the SNP genotypes are encoded to the well-designed oligonucleotides, named DNA coded numbers (DCNs) [20]. The assignment of the DCNs to the SNPs is unconstrained; therefore, the DNA chips prepared to read out the types of DCN are universally available for any types of SNP. We revealed that the DigiTag assay has the potential to analyze nearly all kinds of SNP with high accuracy and reproducibility. However, the DigiTag assay needs the washing step with magnetic beads, which is a laborious step in manipulation. Also, the biotinylated probes, which are necessary for the washing step, are expensive. For the next version of the assay, we improved the protocol to exclude the washing step and named it the DigiTag2 assay.

The DigiTag2 assay involves four steps to accomplish the genotyping: target preparation, encoding, labeling, and detection (Fig. 1). During target preparation, target fragments (including target SNP sites) are prepared by multiplex PCR from genomic DNA. For multiplex PCR, we designed 40-mer primers (average length) and performed multiplex PCR using a two-step protocol (denature and extension steps) with a 6-min extension step. For encoding, we prepared two 5′ query probes and one 3′ query probe for each SNP site. The 5′ query probes have a sequence

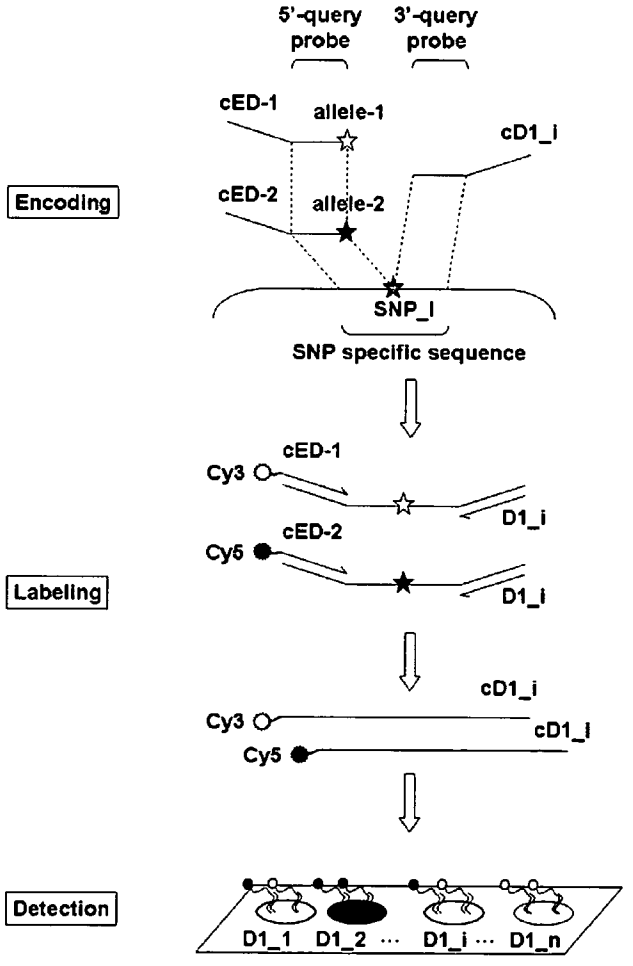


Fig. 1. Schematic representation of DigiTag2 assay. This assay involves four steps to accomplish SNP typing: target preparation, encoding, labeling, and detection. The 5′ query probes have EDs (cED-1 and cED-2) corresponding to each allele, and the 3′ query probes have a variable sequence (cD1_i) for each SNP. Each reverse complement sequence is depicted by a lowercase "c" before the sequence name.

complementary to the 5′-flanking region of the target SNP, and each of the probes has an allele-specific sequence. Two types of ED (ED-1 and ED-2) were attached to each of the 5′ query probes (see Materials and Methods), and we incorporated a mismatch base into the 5′ query probes at the fourth position from the SNP site to improve the precision of allele discrimination [20]. The 3′ query probe has a sequence complementary to the 3′-flanking region of the target SNP, and each of the probes has a D1 on its 3′ end. In the encoding step, the 5′ query and 3′ query probes are successfully concatenated by *Taq* DNA ligase, and the probes are fully complementary to adjacent regions on the target fragment [22]. The genotype is then converted into a type of ED and a type of D1. The types of ED and D1 designate the type of allele and SNP, respectively. After the encoding step, fluorescence is incorporated into the ligated products by asymmetric PCR using fluorescent-labeled primers (Cy3-ED-1 and Cy5-ED-2) and D1 primers. The D1 primers are a mixture of all D1s used in the assay. The Cy3- and Cy5-labeled PCR products are directly hybridized with the D1 probes on the DNA microarray to reveal SNP genotypes by reading signals from the various D1s. If the genomic DNA sample is homozygous for a certain SNP, a single color signal from Cy3 or Cy5 is detected from the corresponding spot on the DNA microarray. In contrast, both signals are present when the genomic DNA sample is heterozygous.

SNP selection and probe design

In a previous report, we investigated the ligation conditions in the encoding step using an SNP located in the *PLOD* gene on human chromosome 1p36 as a model SNP (JSNP ID IMS-JST068774) and determined the parameters for 5′ query and 3′ query probes [20]. We then randomly selected 96 SNPs from a 610-kb region, including the *IL-4* and *IL-13* genes on human chromosome 5q31-33, which contains various candidate genes related to immune and allergic disorders. We subsequently designed probes for the 96 SNP sites to have a uniform melting temperature as that of *PLOD* SNP so as to give similar ligation efficiency among the 96 SNP sites to be analyzed in a single tube. We also incorporated a mismatch base into the 5′ query probe at the fourth position from the SNP site for all target SNPs. The 20-mer mismatch-induced 5′ query probes and 3′ query probes (average length) had melting temperatures of 50.7 ± 2.1 °C and 52.4 ± 1.5 °C, respectively. Here we found that the length of the 3′ query probe influences the ligation efficiency in the encoding step; when a longer 3′ query probe was used in the encoding step, stronger signal intensities were acquired on microarray detection (data not shown). Therefore, we used the lengthened 3′ query probes to 30-mer, and the average melting temperature of the lengthened 3′ query probes was 66.1 ± 3.5 °C. The sequence information for 5′ query probes and lengthened 3′ query probes is listed in Supplementary Table 3.

Optimization of reaction conditions

When we used the mismatch-induced 5′ query probes, indistinct clusters were observed from 5 SNPs (SNP 7, SNP 9, SNP 18, SNP 49, and SNP 93) (Fig. 2A). However, these 5 SNPs can be discriminatively genotyped with perfect

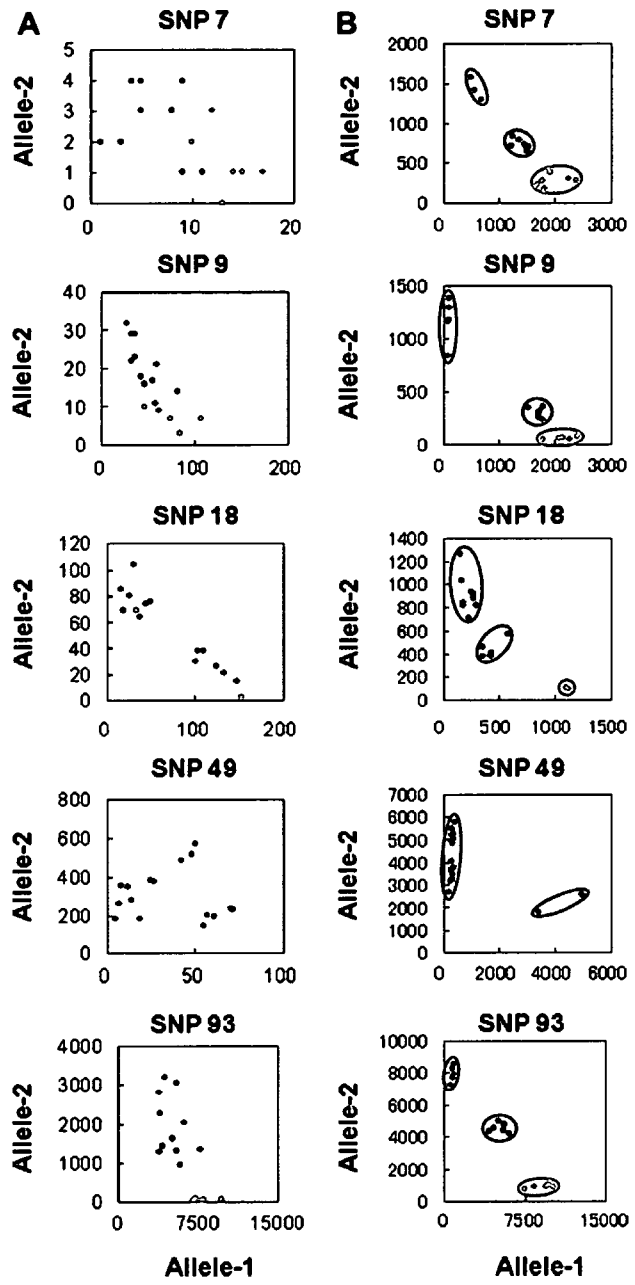


Fig. 2. Treatment of 5 failed SNPs with mismatch-induced 5′ query probes. Green dots and circle show allele-1 homozygous samples, red dots and circle show allele-2 homozygous samples, and blue dots and circle show heterozygous samples. (A) Mismatch-induced 5′ query probes, which have a mismatched base incorporated into the fourth position from the SNP base, were used. (B) Here 5′ query probes, which have a perfect match sequence for the target SNP, were used instead of mismatch-induced 5′ query probes.

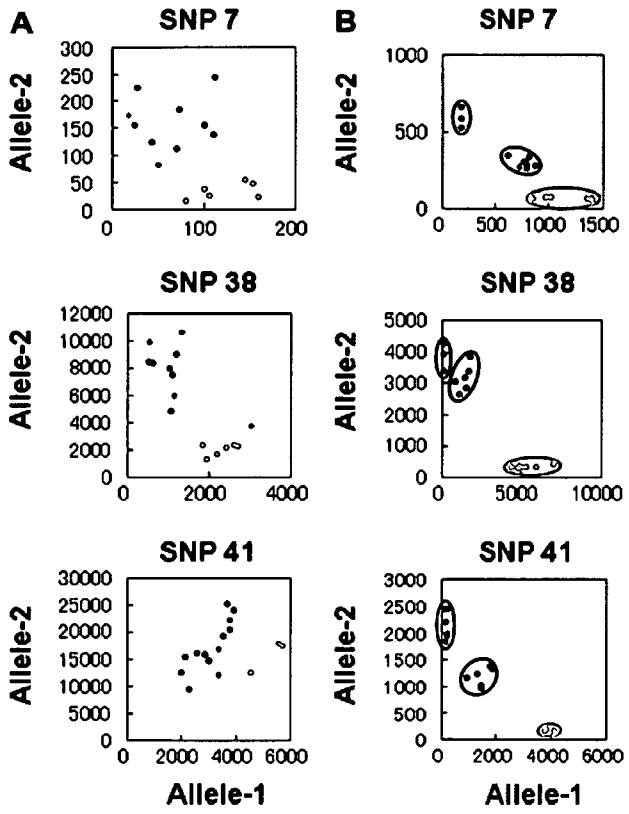


Fig. 3. Effects of D1 primer concentration in the labeling step. Green dots and circle show allele-1 homozygous samples, red dots and circle show allele-2 homozygous samples, and blue dots and circle show heterozygous samples. (A) Here 5 nM D1 primers was used with 500 nM fluorescent-labeled primers. (B) Here 2.5 nM D1 primers was used with 500 nM fluorescent-labeled primers.

match 5′ query probes (Fig. 2B). For these 5 SNPs, the mismatch base incorporated into the fourth position from the SNP site has a drastic effect on the hybridization stability between 5′ query probes and the target multiplex PCR products in the encoding step and leads to signal loss on microarray detection. Therefore, we performed a 96-plex oligonucleotide ligation assay with the mismatch-induced 5′ query probes for 91 target SNPs in combination with perfect match 5′ query probes for these 5 target SNPs.

To incorporate the fluorescent label into the ligated products, we performed asymmetric PCR with fluorescent-labeled primers and D1 primers (mixture of D1s). Using the mixture of all D1s, instead of the single primer pair mentioned in the DigiTag assay [20], would make it possible to uniformly acquire all target fragments. However, the concentration of D1 primers used in the labeling step was found to exert an influence on the cluster distribution in scatter diagrams. When we used 5 nM D1 primers with 500 nM fluorescent-labeled primers, dispersed and/or indiscrete clusters were observed for several SNPs (Fig. 3A). However, the dispersed and/or indiscrete clusters became convergent and/or discrete clusters when we used 2.5 nM D1 primers with 500 nM fluorescent-labeled primers (Fig. 3B). The D1 primers share the fluorescent-labeled primers in the labeling step, and the ratio of each D1 primer to fluorescent-labeled primer was approximately 1:2 at 2.5 nM and 1:1 at 5 nM. When the amount of each D1 primer was greater than the amount of fluorescent-labeled primer, strong false-positive signals were observed on microarray detection, leading to indiscrete clusters on scatter diagrams (data not shown). On the other hand, when the amount of each D1 primer was less than the amount of fluorescent-labeled primer, insufficient amplification occurred in a number of target SNPs, leading to weak signal intensities on microarray detection (data not shown). We found that the optimal ratio of D1 primer to fluorescent-labeled primer is approximately 1:2, irrespective of the multiplicity of the assay (number of SNPs to be analyzed).

Genotyping results

Multiplex PCR products, including the 96 SNP sites, showed similar band patterns as 48 individual DNA samples, although it was difficult to clearly discern all 96 PCR products due to limitations in electrophoretic resolution (Fig. 4A). We then performed a multiplexed oligonucleotide ligation assay using the multiplex PCR products as targets. To incorporate the fluorescent label into the ligated products, asymmetric PCR was performed using the fluorescent-labeled and D1 primers. DNA microarray revealed hybridization images of 24 individual samples from each of the 24 separated areas having 100 spots (96 probes for 96 SNPs and 4 probes for validation controls) (Fig. 4B). The hybridization image was analyzed using a DNA chip scanner, and the Cy3 and Cy5 signal intensities of each spot were plotted to produce a scatter diagram. The SNP genotypes of 16 genomic DNA samples, randomly selected from the 48 samples, were alternatively determined by direct sequencing and were used as reference data.

As a result of 96-plex genotyping under optimal labeling conditions using the mismatch-induced 5′ query probes in combination with the perfect match 5′ query probes for 5 SNPs, three distinct clusters corresponding to two homozygous genotypes and one heterozygous genotype were observed from 84 SNPs (exceptions were SNP 31, SNP 37, SNP 60, SNP 61, and SNP 87) (Fig. 4C). The remaining 7 SNPs (SNP 12, SNP 22, SNP 27, SNP 33, SNP 67, SNP 88, and SNP 91) were found to be monomorphic in 48 genomic DNA samples and were excluded from further analysis. For SNP 37, SNP 60, and SNP 87, drastically attenuated signal intensities were observed on microarray detection (Fig. 4D). Signal loss was caused by insufficient amplification of the target fragments on multiplex PCR because no amplified products were observed on singleplex PCR, even when the second candidate primer pairs were used. There were some structural obstacles in the target region, although we could not identify any characteristic structures. SNP 31 and SNP 61 were found to have strong false-positive signals, leading to indistinct clusters on scatter diagrams (Fig. 4E). The false-positive signals would be caused by the misligation in the encoding step that was

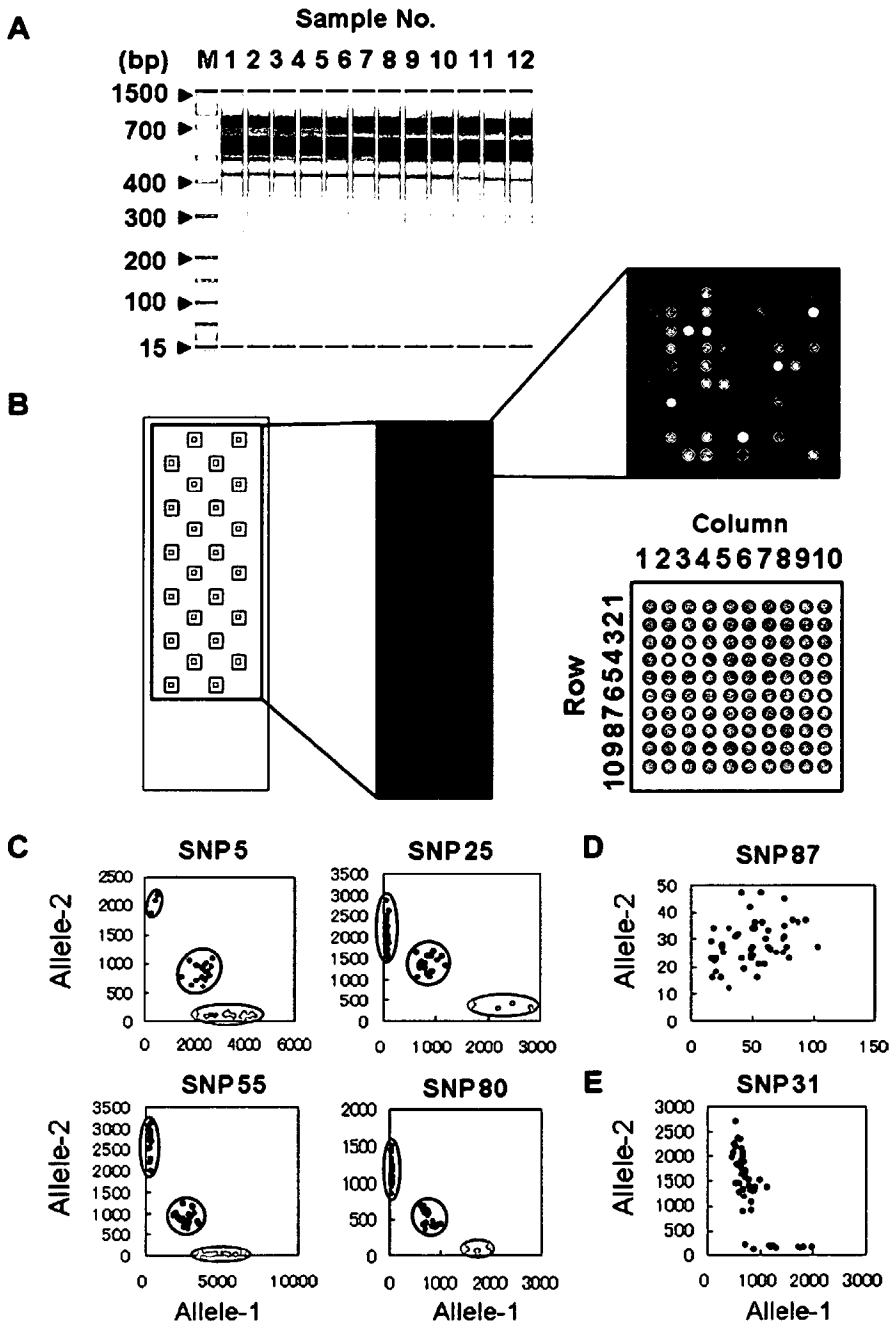


Fig. 4. Multiplex SNP typing for 96 SNPs using 48 individual genomic DNA samples. (A) Gel images of multiplex PCR products with different samples. In all sample lanes, sample bands were observed between two inner markers: 15 and 1500 bp. (B) Hybridization images of DNA microarray. (C) Scatter plot diagrams for 4 randomly selected SNPs from 84 working SNPs. Green dots and circle show allele-1 homozygous samples, red dots and circle show allele-2 homozygous samples, and blue dots and circle show heterozygous samples. (D) Example of the typing-failed SNP caused by insufficient amplification of target fragment in multiplex PCR. (E) Example of the typing-failed SNP that was found to have strong false-positive signals.

reported to be prone to occur when the mismatched pairs are G–T, G–A, G–G, A–G, and T–G [23,24]. Of the 2 misligated SNPs, 1 had an A–G mismatch (SNP 31) and the other had a T–G mismatch (SNP 61) between the 5′ query probe and the target fragment. These 2 SNPs were undetectable, even when 5′ query probes with the mismatched base incorporated into different positions were used (data not shown). Although there were other G–T, G–A, G–G, A–G, and T–G mismatches within the set of 84 working SNPs, we consider that these mismatches would increase the likelihood of misligation in some cases. In the future, we will be able to search for the cause of misligation by

accumulating data from failed analyses of numerous target SNPs.

Conversion rate, call rate, accuracy, and reproducibility

We investigated the feasibility of the DigiTag2 assay by performing 96-plex SNP typing using 48 human genomic DNA samples, and we found that the DigiTag2 assay has the potential to analyze all types of SNP with high accuracy and reproducibility. Here we excluded 7 SNPs from further analysis because they were revealed to be monomorphic in 48 samples. The DigiTag2 assay was found to have a 94.4% (84/89) conversion rate, which is defined as the proportion of successfully genotyped SNPs among the total number of SNPs examined. The call rate, which is defined as the number of genotype calls among the total number of samples examined, was 99.95% (4030/4032) The typing results were 100% identical to the results of direct sequencing. The reproducibility of this assay was examined by duplicate experiments, and it was found that genotype calls were 100% identical between duplicate experiments.

Advantages of DigiTag2 assay

The DigiTag2 assay performs multiplex PCR to excise target regions, including SNP sites from genomic DNA, prior to oligonucleotide ligation assay. Reducing the complexity of the genome by selectively collecting target SNP sites from the genome would lead to successful genotyping [25]. In designing the genotyping probes for oligonucleotide ligation assay, there are no alternatives because the SNP sequence is included in the probe sequences. Therefore, multiplex PCR prior to oligonucleotide ligation assay has an important role in analyzing SNPs that have highly homogeneous regions in the genome. Based on 96-plex SNP typing using 48 individual genomic DNA samples, the DigiTag2 assay has the potential to analyze all types of SNP with high accuracy and reproducibility. Moreover, the DigiTag2 assay uses unmodified primers and probes for target SNPs, thereby reducing assay cost, and requires only simple assay protocols without specialized equipment. We estimated that the running cost for the DigiTag2 assay (for oligonucleotides, reagents, DNA microarrays, etc.) is less than $0.06/genotype. The DigiTag2 assay can use the same set of D1s and EDs for any set of target SNPs, thereby enabling 96-plex genotyping with the same assay protocols and the same microarray having the same set of probes. However, hybridization products, which are prepared in the labeling step, may cross-hybridize to the wrong D1 probes on DNA microarray due to SNP-specific sequences being introduced into the hybridization products. With regard to the 96 target SNPs selected in this study, there is no evidence of cross-hybridization between D1 probes and SNP-specific sequences. Cross-hybridization may be avoided by predicting the interaction between D1 probes and SNP-specific sequences. In the future, we will attempt to predict cross-hybridization by accumulating data from failed analyses of numerous target SNPs.

Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research on Priority Areas and the New Energy and Industrial Technology Development Organization.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ab.2007.02.005.

References

[1] L. Kruglyak, D.A. Nickerson, Variation is the spice of life, Nat. Genet. 27 (2001) 234–236.
[2] M.W. McBride, D. Graham, C. Delles, A.F. Dominiczak, Functional genomics in hypertension, Curr. Opin. Nephrol. Hypertens. 15 (2006) 145–151.
[3] M. Ochi, H. Osawa, H. Onuma, A. Murakami, T. Nishimiya, F. Shimada, K. Kato, I. Shimizu, K. Shishino, M. Murase, Y. Fujii, J. Ohashi, H. Makino, The absence of evidence for major effects of the frequent SNP +299 G > A in the resistin gene on susceptibility to insulin resistance syndrome associated with Japanese type 2 diabetes, Diabetes Res. Clin. Pract. 61 (2003) 191–198.
[4] B.I. Freedman, D.W. Bowden, M.M. Sale, C.D. Langefeld, S.S. Rich, Genetic susceptibility contributes to renal and cardiovascular complications of type 2 diabetes mellitus, Hypertension 48 (2006) 8–13.
[5] R. Yamada, S. Tokuhiro, X. Chang, K. Yamamoto, SLC22A4 and RUNX1: identification of RA susceptible genes, J. Mol. Med. 82 (2004) 558–564.
[6] K. Yamamoto, R. Yamada, Genome-wide single nucleotide polymorphism analyses of rheumatoid arthritis, J. Autoimmun. 25 (2005) 12–15.
[7] N. Tsuchiya, J. Ohashi, K. Tokunaga, Variations in immune response genes and their associations with multifactorial immune disorders, Immunol. Rev. 190 (2002) 169–181.
[8] J. Ohashi, K. Tokunaga, The power of genome-wide association studies of complex disease genes: Statistical limitations of indirect approaches using SNP markers, J. Hum. Genet. 46 (2001) 478–482.
[9] A. Wille, J. Hoh, J. Ott, Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers, Genet. Epidemiol. 25 (2003) 350–359.
[10] C.S. Carlson, M.A. Eberle, L. Kruglyak, D.A. Nickerson, Mapping complex disease loci in whole-genome association studies, Nature 429 (2004) 446–452.
[11] N. Hu, C. Wang, Y. Hu, H.H. Yang, C. Giffen, Z-Z. Tang, X-Y. Han, A.M. Goldstein, M.R. Emmert-Buck, K.H. Buetow, P.R. Taylor, M.P. Lee, Genome-wide association study in esophageal cancer using GeneChip Mapping 10 K Array, Cancer Res. 65 (2005) 2542–2546.
[12] D.J. Schaid, J.C. Guenther, G.B. Christensen, S. Hebbring, C. Rosenow, C.A. Hilker, S.K. McDonnell, J.M. Cunningham, S.L. Slager, M.L. Blute, S.N. Thibodeau, Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility loci, Am. J. Hum. Genet. 75 (2004) 948–965.
[13] T. Arinami, T. Ohtsuki, H. Ishiguro, H. Ujike, Y. Tanaka, Y. Morita, M. Mineta, M. Takeichi, S. Yamada, A. Imamura, K. Ohara, H. Shibuya, K. Ohara, Y. Suzuki, T. Muratake, N. Kaneko, T. Someya, T. Inada, T. Yoshikawa, T. Toyota, K. Yamada, T. Kojima, S. Takahashi, O. Osamu, T. Shinkai, M. Nakamura, H. Fukuzako, T. Hashiguchi, S. Niwa, T. Ueno, H. Tachikawa, T. Hori, T. Asada, S. Nanko, H. Kunugi, R. Hashimoto, N. Ozaki, N. Iwata, M. Harano, H. Arai, T. Ohnuma, I. Kusumi, T. Koyama, H. Yoneda, Y. Fukumaki, H. Shibata, S. Kaneko, H. Higuchi, N. Yasui-Furukori, Y. Numachi, M. Itokawa, Y. Okazaki, Japanese Schizophrenia Sib-Pair Linkage

Group, Genome-wide high-density SNP linkage analysis of 236 Japanese families supports the existence of schizophrenia susceptibility loci on chromosomes 1p, 14q, and 20p, Am. J. Hum. Genet. 77 (2005) 937–944.

[14] P.M. Holland, R.D. Abramson, R. Watson, D.H. Gelfand, Detection of specific polymerase chain reaction product by utilizing the 5′→3′ exonuclease activity of *Thermus aquaticus* DNA polymerase, Proc. Natl. Acad. Sci. USA 88 (1991) 7276–7280.

[15] N. Pourmand, E. Elahi, R.W. Davis, M. Ronaghi, Multiplex pyrosequencing, Nucleic Acids Res. 30 (2002) e31.

[16] K. Lindroos, U. Liljedahl, M. Raitio, A-C. Syvänen, Minisequencing on oligonucleotide microarrays: Comparison of immobilisation chemistries, Nucleic Acids Res. 29 (2001) e69.

[17] J. Tost, I.G. Gut, Genotyping single nucleotide polymorphisms by mass spectrometry, Mass Spectrom. Rev. 21 (2002) 388–418.

[18] M.S. Bray, E. Boerwinkle, P.A. Doris, High-throughput multiplex SNP genotyping with MALDI–TOF mass spectrometry: Practice, problems, and promise, Hum. Mutat. 17 (2001) 296–304.

[19] A.R. Tobler, S. Short, M.R. Andersen, T.M. Paner, J.C. Briggs, S.M. Lambert, P.P. Wu, Y. Wang, A.Y. Spoonde, R.T. Koehler, N. Peyret, C. Chen, A.J. Broomer, D.A. Ridzon, H. Zhou, B.S. Hoo, K.C. Hayashibara, L.N. Leong, C.N. Ma, B.B. Rosenblum, J.P. Day, J.S. Ziegle, F.M. De La Vega, M.D. Rhodes, K.M. Hennessy, H.M. Wenz, The SNPlex genotyping system: A flexible and scalable platform for SNP genotyping, J. Biomol. Tech. 16 (2005) 396–404.

[20] N. Nishida, T. Tanabe, K. Hashido, K. Hirayasu, M. Takasu, A. Suyama, K. Tokunaga, DigiTag assay for multiplex single nucleotide polymorphism typing with high success rate, Anal Biochem. 346 (2005) 281–288.

[21] H. Yoshida, A. Suyama, Solution to 3-SAT by breadth first search, DIMACS Ser. Discrete Math. Theor. Comput. Sci. 54 (2000) 9–22.

[22] F. Barany, Genetic disease detection and DNA amplification using cloned thermostable ligase, Proc. Natl. Acad Sci. USA 88 (1991) 189–193.

[23] J.N. Housby, E.M. Southern, Fidelity of DNA ligation: A novel experimental approach based on the polymerisation of libraries of oligonucleotides, Nucleic Acids Res. 26 (1998) 4259–4266.

[24] J. Luo, D.E. Bergstrom, F. Barany, Improving the fidelity of *Thermus thermophilus* DNA ligase, Nucleic Acids Res. 24 (1996) 3071–3078.

[25] H. Matsuzaki, H. Loi, S. Dong, Y-Y. Tsai, J. Fang, J. Law, X. Di, W-M. Liu, G. Yang, G. Liu, J. Huang, G.C. Kennedy, T.B. Ryder, G.A. Marcus, P.S. Walsh, M.D. Shriver, J.M. Puck, K.W. Jones, R. Mei, Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array, Genome Res. 14 (2004) 414–425.