CrossMark

RESEARCH PAPER

Pyrosequencing with di-base addition for single nucleotide polymorphism genotyping

Dan Pu[1] · Chengguang Mao[1] · Lunbiao Cui[2] · Zhiyang Shi[2] · Pengfeng Xiao[1]

Abstract We develop color code-based pyrosequencing with di-base addition for analysis of single nucleotide polymorphisms (SNPs). When a di-base is added into the polymerization, one or several two-color code(s) containing the type and the number of incorporated nucleotides will be produced. The code information obtained in a single run is useful to genotype SNPs as each allelic variant will give a specific pattern compared to the two other variants. Special care has to be taken while designing the di-base dispensation order. Here, we present a detailed protocol for establishing sequence-specific di-base addition to avoid nonsynchronous extension at the SNP sites. By using this technology, as few as 50 copies of DNA templates were accurately sequenced. Higher signals were produced and thus a relatively lower sample amount was required. Furthermore, the read length of per flow was increased, making simultaneous identification of multiple SNPs in a single sequencing run possible. Validation of the method was performed by using templates with two SNPs covering 37 bp and with three SNPs covering 58 bp as well as 82 bp. These SNPs were successfully genotyped by using only a sequencing primer in a single PCR/sequencing run. Our results demonstrated that this technology could be potentially developed into a powerful methodology to accurately determine SNPs so as to diagnose clinical settings.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-016-9359-5) contains supplementary material, which is available to authorized users.

✉ Pengfeng Xiao
xiaopf@seu.edu.cn

[1] State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, Nanjing, Jiangsu 210096, China

[2] Key Laboratory of Enteric Pathogenic Microbiology (NHFPC), Jiangsu Provincial Center for Disease Control and Prevention, Nanjing, Jiangsu 210009, China

Introduction

Genetic variation is the basis for human diversity and plays an important role in human diseases. Single nucleotide polymorphisms (SNPs) in human genomes compose the majority of genetic variation. Therefore, it may largely determine the differences among individuals [1]. In the case of investigating SNPs, DNA sequencing technologies are likely to be the most reliable method for this purpose. An alternative technique for DNA sequencing, known as pyrosequencing, is extensively used to genotype SNPs. This technique adds limiting amount of deoxynucleotide triphosphate (dNTP) bases one at a time to control DNA synthesis [2] and yields detectable light by a cascade of enzymatic reactions [3, 4]. Because the detectable light is proportional to the amount of incorporated nucleotides, it is extensively used to measure microbial species [5, 6], secondary structure [7], allele frequency [8, 9], or single nucleotide polymorphisms [10, 11]. However, it has two main limitations including short read length due to nonsynchronized extension [12] and limited detection sensitivity owing to high background signal [13]. Increasing the read length will reduce the cost of DNA analysis for most applications [12]. In addition, improving the sensitivity will enable to detect relatively lower amount of DNA samples. Therefore, any technologies that could increase the read

length and improve the sensitivity of pyrosequencing are welcome.

Generally, to interrogate each variation, a pair of biotinylated and non-biotinylated PCR primers and a sequencing primer are needed, which will increase the material cost as the number of characterized SNPs increases. To reduce the cost, one kind of multiplexing pyrosequencing has been proposed, in which more than one polymorphisms on a single template were analyzed by using multiple sequencing primers [9, 14, 15]. Another kind of multiplexing used only one sequencing primer to genotype several SNPs at different positions. It has been reported that two variations located within a 15-base or 20-base segments on the same template were genotyped by a single sequencing primer in a single run [7, 16]. Also, a single sequencing primer complementary to DNA sequence flanking three closely located SNPs was hybridized to the template, and pyrosequencing was performed for allele calling [17]. This technique enables analysis of multiple SNPs on the same template, thereby decreasing the cost for PCR, template preparation, and genotyping. However, it should be mentioned that this multiplex genotyping technique is currently limited to the determination of variations located within 25 bases and to the determination of three SNPs at most with singleplex pyrosequencing [7, 10, 17].

Recently, our group has proposed a decoding sequencing technique. The technique has the potential in increasing the read length and amplifying the signal intensities [18]. Here, we focus on improvement of pyrosequencing for the analysis of SNPs by targeting the aforementioned two limiting factors—short read length and limited detection sensitivity. Pyrosequencing with di-base addition was developed, and only a single sequencing run was required to genotype several SNPs located in the uridine diphosphoglucuronosyl transferase 1A1 (UGT1A1) gene in different positions.

Material and methods

DNA extraction

Genomic DNA was extracted from 500 μL of whole blood using DNA Isolation Kit for Blood/Bone Marrow/Tissue (Roche). Extracted DNA was stored at −20 °C until analyzed.

Primer design and PCR amplification

Primers for the initial PCR amplification of UGT1A1-specific fragments were designed using conventional standard criteria. Table 1 showed different sets of primers used to amplify UGT1A1-specific fragments. These fragments contained SNPs in the regions of exon 1 (rs4148323, rs56059937), 3′-conserved region (rs4148329, rs17862881, rs6717546), and intron 1 (rs6742078, rs3771342, rs4148324).

The PCR amplification was performed in a total volume of 50 μL, using 10 ng human genomic DNA, 0.2 μM each of the forward and reverse primers, 10× Taq Buffer, TaqGold (Applied Biosystems, NJ, USA), and 200 μM of each dNTP. PCR amplifications were performed with an initial denaturation for 4 min at 98 °C, followed by 45 cycles of denaturation for 30 s at 95 °C, primer annealing for 30 s at 58 °C, and extension for 1 min at 72 °C, followed by a final extension for 5 min at 72 °C.

Template preparation

PCR product (45 μL) was transferred into a new PCR plate, and then for each sample, 50 μL of binding buffer (10 mM Tris–HCl, 2 M NaCl, 1 mM EDTA, and 0.1 % Tween 20, pH 7.6) and 5 μL of streptavidin Sepharose beads (Amersham Biosciences AB, Uppsala, Sweden) were added and mixed at room temperature, 1500 rpm for 15 min (Thermomixer Comfort, Eppendorf). Single-stranded DNA was obtained by using the standard protocol for the Vacuum Prep Tool (Biotage AB, Sweden). The single-stranded DNA was hybridized to 10 pmol sequencing primer dissolved in 45 μL 1× annealing buffer (200 mM Tris-acetate and 50 mM MgAc$_2$, pH 7.6) at 80 °C for 2 min and then cooled to room temperature.

Conventional pyrosequencing

A 96-well plate containing single-stranded DNA template with the annealed sequencing primer was placed inside the pyrosequencer instrument, PSQ 96MA system (Biotage AB, Uppsala, Sweden). The cartridge was filled with an enzyme mixture (DNA polymerase, ATP sulfurylase, luciferase, apyrase), a substrate mixture (luciferin, adenosine 5′-phosphosulfate), and nucleotides (dATPαS (α-thio-dATP), dCTP, dGTP, dTTP) contained in SQA PyroMark Gold Q96 Reagents 5 × 96 (QIAGEN, Hilden, Germany). The volumes required were indicated in the run setup. After the run, the peaks were evaluated according to the expected pattern by referring to the dispensation order.

Pyrosequencing with di-base addition

As for the cartridge, it contains six reagent vials. Substrates and enzymes are separately added into a reagent vial. With respect to the nucleotide addition, there are six available di-base combinations (AG, CT, AC, GT, AT, CG), and each combination can be applied to genotype SNPs. However, there are only a total of four reagent vials instead of six ones to delivery nucleotides in the cartridge. Thus, for each assay, any four from the six di-base combinations (AG, CT, AC, GT, AT, CG) can be chosen and added into the four reagent vials. Take di-base dispensation orders AG/AT/AC/CT as an

Table 1 PCR and sequencing primers

dbSNP	Location	Nucleotide change	PCR primer
rs4148323	Exon 1	G>A	F[a]-ACCTGACGCCTCGTTGTACATC
rs56059937		T>C	R[b]-B[c]-TGCCCGAGACTAACAAAAGACT
			SP[d]-TCGTTGTACATCAGAGA
rs6717546	3′-conserved region	A>G	F-B-ACTGGTGTCGCATTTATTTC
rs17862881		A>G	R-GAGTAGAAGAAAGAGTGTTAGC
rs4148329		C>T	SP-GTTATTTCTGGGCGAT
rs4148324	Intron 1	G>T	F-B-GAGCAGGGAAGGTTGAGATATATG
rs3771342		G>T	R-CCAGCCAAAATGAGACATTTAACA
rs6742078		G>T	SP-TTTTAGTTTAGGCTTTTAAT

[a] Forward primer

[b] Reverse primer

[c] Biotin-labeled primer

[d] Sequencing primer

example to demonstrate it. A mixture of dATPaS and dGTP (di-base AG), a mixture of dATPaS and dTTP (di-base AT), a mixture of dATPaS and dCTP (di-base AC), and a mixture of dCTP and dTTP (di-base CT) were placed into each reagent vial. The sequencing procedures were carried out by stepwise adding AG/CT/AC/GT, respectively. As the elongation of the primer strand proceeded, a set of the encodings was sequentially obtained. The pyrosequencing reactions were carried out with the di-base dispensation orders shown in Table S1 in the Electronic Supplementary Material (ESM). After the run, the peaks were translated into two-color codes and evaluated according to the expected color code patterns by referring to the dispensation order.

Stepwise verification of the method

The amplified fragments were examined according to length, after gel electrophoresis in a 2 % agarose gel for 45 min at 110 V. The verification also included an assay that genotyped these SNPs with Sanger DNA technique. The interest PCR amplicons were sequenced using the BigDye Terminator Reaction Kit (ABI, USA) on an ABI automated DNA sequencer (3130 DNA Analyzer).

Reproducibility

Each sample used for analysis variants was also reanalyzed in triplicate to confirm the reproducibility of the method.

Results and discussion

The principle of this method for SNP genotyping

In general, nucleotides A, G, C, and T can form 16 combinations of di-bases. These include AA, CC, GG, TT, AC, CA, AG, GA, AT, TA, CG, GC, CT, TC, TG, and GT, six of which are different di-base combinations (AG, CT, AC, GT, AT, CG). The six combinations can form three sets of dual mononucleotide additions, AG/CT, AC/GT, and AT/CG. In our previous work, a decoding method was proposed, in which AG/CT, AT/CG, or AC/GT was cyclically added to interrogate the queried sequences, and the queried template was reconstructed by decoding two sets of encodings from two sequencing runs [18]. Here, pyrosequencing with di-base addition is established, in which AG, CT, AC, GT, AT, and CG are selectively added into the reactions instead of cyclically adding AG/CT, AC/GT, or AT/CG, and SNP genotyping is performed by using only a single sequencing run. We use two-color code(s), which contain(s) the information about the possible type of incorporated base, to represent the sequencing signal in each cycle (a reaction is defined as a cycle in this assay). The number of two-color code(s) obtained represents the number of incorporated nucleotide(s) in each cycle. Four color codes (blue, green, yellow, red) and 12 two-color codes are applied to encode for the 16 possible di-bases (Fig. 1a). The color codes, blue, green, yellow, and red, represent bases A, C, G, and T, respectively. Among the 16 di-bases, six are different di-bases (AG, CT, AC, GT, AT, CG) and applied for pyrosequencing with di-base addition. When a di-base AC is added into the reaction, one or several nucleotide(s) would be incorporated into the template, and thus a peak is produced in the pyrogram. The resultant peak is translated into two-color code(s) which is half blue and half green. Concerning to the color codes, there are several issues needed to be stated. First, two different di-bases that have the same first (or the same second) base get different two-color codes, for example, two-color codes (AC) ≠ (AT) and two-color codes (AC) ≠ (TC). Second, a di-base and its reverse get the same color code, for example, two-color codes (AC) = (CA). All the same two-color codes are shown in Fig. 1b. The system can reconstruct the queried base sequences based on a scheme that
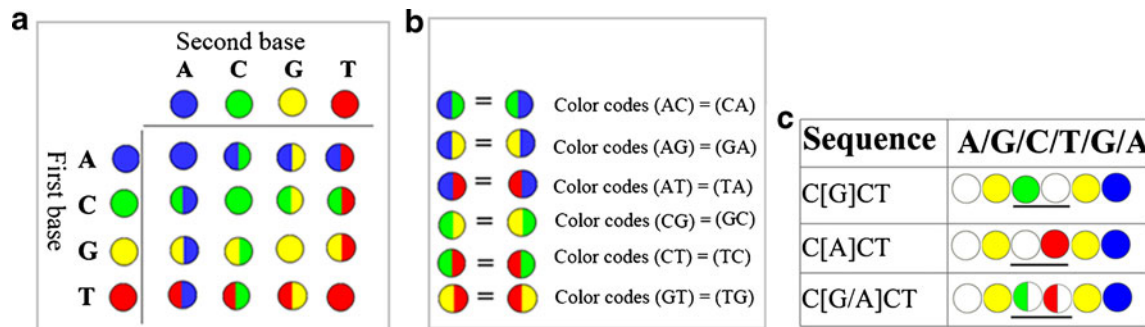
Fig. 1 The color coding scheme of this technology and possible color codes predicted from rs6717546. (**a**) The 12 two-color codes and four-color codes applied to represent 16 possible di-bases (AC, CA, AG, GA, AT, TA, CG, GC, CT, TC, GT, TG, AA, CC, GG, TT). (**b**) All the same two-color codes. (**c**) Color codes from conventional pyrosequencing. Sequence C[G/A]CTAA is interrogated by sequentially adding A/G/C/T/G/A

the same color code between the two compared two-color codes is right the incorporated nucleotide (see ESM Fig. S1).

Conventional pyrosequencing is based on the fact that each allelic variant will give a specific pattern to differentiate allelic variants. As for pyrosequencing with di-base addition, although a sequencing reaction provides two-color code(s) or encoding(s) instead of concrete base sequences, specific pattern (color codes or encodings) will be produced for each variant under a given di-base addition, thus making it suitable to genotype SNPs. Generally, when a di-base is added into a polymerization, one or several nucleotide(s) will be incorporated in the template. It thus may lead to nonsynchronized extension. To make sure that each variant gives specific pattern and to keep synchronized extension after SNP positions, a
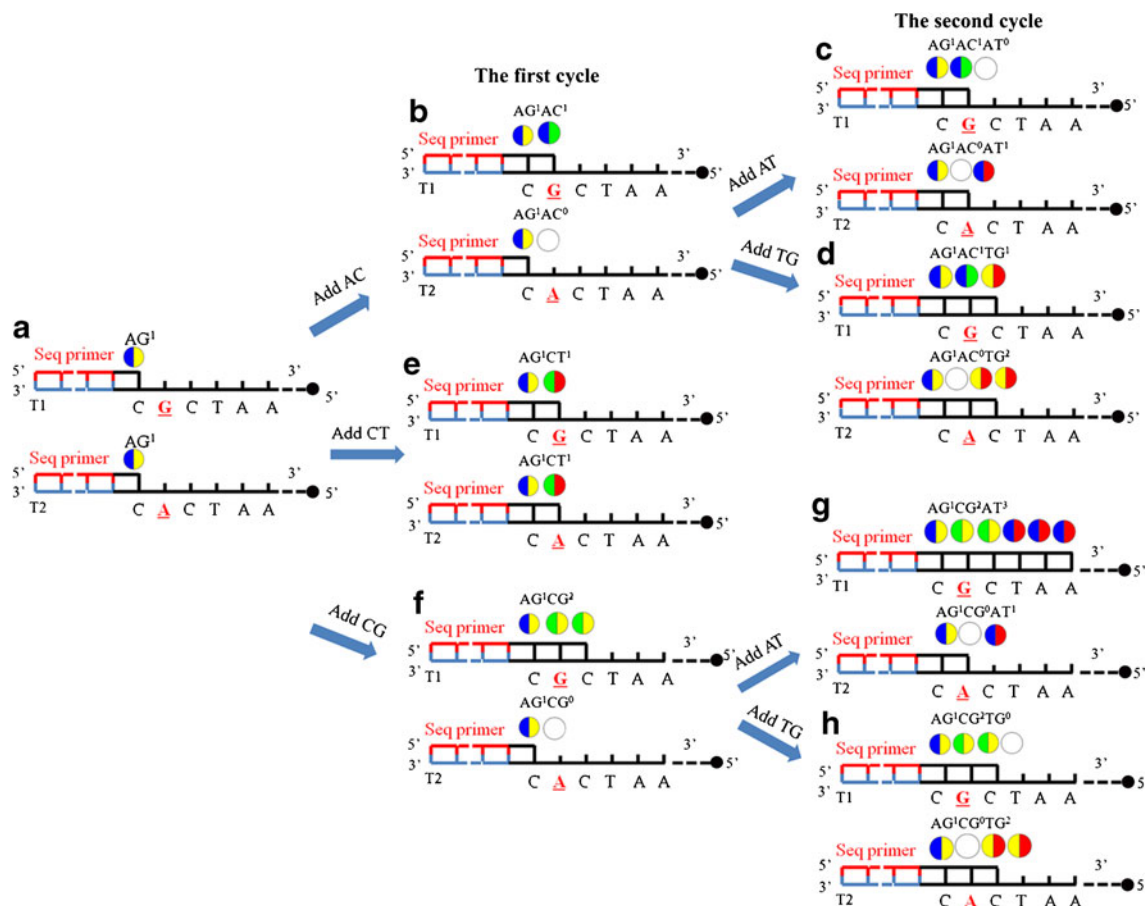
Fig. 2 The procedure illustrates how to design sequence-specific di-base dispensation orders for analysis of SNP rs6717546 (its flanking sequence is C[G/A]CTAA) via pyrosequencing with di-base addition. (**a**) Di-base AG is added to obtain a reference counterpart AG[1] before the SNP site.

(**b**) Di-base AC is added in the first cycle. (**c**, **d**) After the addition of AC, di-base AT or TG is added in the second cycle. (**e**) Di-base CT is added. (**f**) Di-base CG is added. (**g**, **h**) After the addition of CG, di-base AT or TG is added in the second cycle

new sequence-specific dispensation order is required. The following items introduce how to design the sequence-specific dispensation order at SNP sites (we assume the observed SNPs are biallelic in this assay). First, a di-base containing a base complementary to (or equal to) one of the possible SNP variants is firstly added on purpose to make sure that part of DNA templates is extended (whether the added di-base is complementary to one of the possible SNP variants or not is related to the sequencing direction). Second, another di-base that contains a base complementary to (or equal to) the other possible SNP variant is subsequently added to extend the remaining templates. The sequence-specific dispensation order will make sure that one allelic variant will be interrogated at a time and obtain an estimate of primer elongation synchronization. Once the sequence at the SNP sites is synchronized, the subsequent sequence will be interrogated with alternative addition of any four of the six di-bases. Take a SNP site G/T as an example to demonstrate it. We assume that template fragment is interrogated in the forward direction. Thus, di-base GX containing base G (base G is equal to one of the possible SNP variants) is firstly added on purpose to make sure that part of DNA templates containing variant G is extended. Afterward, another di-base TX that contains base T equal to the other possible SNP variant T is added to extend the remaining templates. Thus, the available di-base dispensation order could be GX/TX or TX/GX. Similarly, if the template fragment is interrogated in the reverse direction, the available di-base dispensation order could be CX/AX or AX/CX. In conclusion, as for SNP site G/T, the available di-base dispensation orders could be GX/TX, TX/GX, CX/AX, or AX/CX.

Here, we take variant rs6717546 (G>A) as an example to demonstrate how to design sequence-specific dispensation orders in detail. Its flanking sequence is C[G/A]CTAA (Fig. 1c). There are two kinds of templates in the reaction, one is CGCTAA (we define it as T1) and the other is CACTAA (we define it as T2). The fragment is interrogated in the reverse direction, so that di-base dispensation order could be CX/TX or TX/CX. Generally, a peak equal to one base incorporation before the SNP site is functioned as a reference counterpart. Thus, AG is added at first to obtain reference peak AG^1 (Fig. 2a). In the SNP position, we can choose to extend genotype G (or A) in the first cycle and then to extend genotype A (or G) in the second cycle. Here, di-base containing base C is chosen to extend genotype G in the first cycle. Since three di-bases (AC, CT, and CG) which contain base C are available to extend genotype G in the first cycle. Thus, di-base added in the first cycle can be discussed in three cases. Firstly, if AC is added in the first cycle, the encodings from T1 and T2 are AC^1 and AC^0, respectively (Fig. 2b). Di-base containing base T is further added to extend another part of templates containing genotype A. In such case, there are two di-bases (AT or TG) available. If AT is added in the second cycle, the encodings from T1 are AC^1AT^0 while the encodings from T2 are AC^0AT^1. The extensions are

synchronized (Fig. 2c). If GT is added in the second cycle, the encodings from T1 are AC^1TG^1, while the encodings from T2 are AC^0TG^2. As shown in Fig. 2d, the extensions are also synchronized. Thus, the di-base dispensation orders AC/AT and AC/GT are suitable to genotype this SNP site. Secondly, if CT is added in the first cycle, the encodings from T1 and T2 are CT^1 and CT^1, respectively. Thus, the SNP genotypes will not be differentiated (Fig. 2e). Thirdly, if CG is added in the first cycle to extend templates containing variant G, the encodings from T1 and T2 are CG^2 and CG^0, respectively (Fig. 2f). We further add
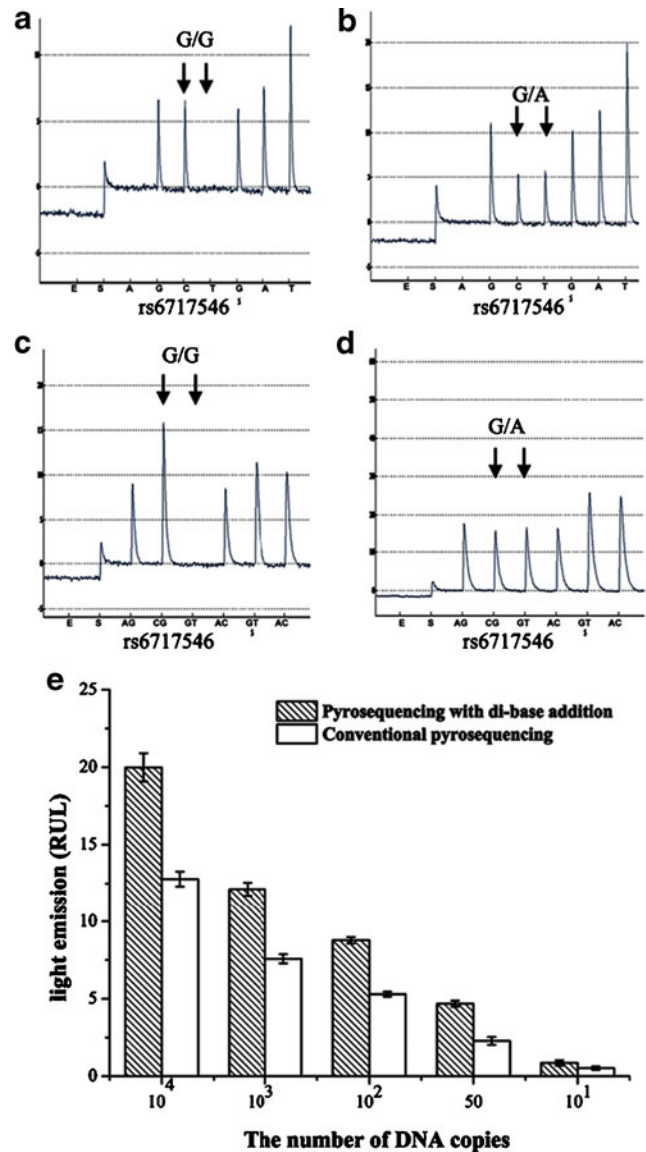
Fig. 3 Identification of rs6717546 by conventional pyrosequencing and by pyrosequencing with di-base addition. Each SNP is indicated by two *arrowhead droplets*. Homozygous G/G (a) and heterozygous G/A (b) genotypes of rs6717546 are analyzed by conventional pyrosequencing. Homozygous G/G (c) and heterozygous G/A (d) of rs6717546 are analyzed by pyrosequencing with di-base addition. (e) Comparison of relative light units (RUL) equal to one nucleotide incorporation of conventional pyrosequencing to that of pyrosequencing with di-base addition

di-bases (AT or TG) containing base T to extend another part of the templates in the second cycle. If AT is added, the encodings from T1 are CG^2AT^3, while the encodings from T2 are CG^0AT^1. The extensions are not synchronized (Fig. 2g). If TG is added, the encodings from T1 are CG^2TG^0, while the encodings from T2 are CG^0TG^2. The extensions are synchronized (Fig. 2h). Thus, CG/GT is suitable to genotype the SNP site.

All in all, for analysis of any SNP, there are two items need to be emphasized. Firstly, a di-base in which the first and second bases are the same as (or complementary to) two allelic variants are not suitable to determine the given SNP. For example, if the genotype of a SNP is G/A, di-bases GA and AG (or di-bases CT and TC) addition does not work (Fig. 2e). Secondly, if the number of two-color codes obtained from two different kinds of templates at the SNP sites is equal under a given order of di-base dispensation, the extensions would be synchronized. For example, when AC/GT is added at the SNP site G/A, the encodings from part of the templates are AC^1GT^1, while the encodings from the remaining templates are AC^0GT^2. The number of two-color codes at the SNP site from the two different kinds of templates is 2 and 2, respectively. Thus, this order of di-base dispensation would be suitable to genotype the SNPs.

When proper order of di-base addition is chosen to analyze the same amount of DNA template, much higher signal intensities will be produced in the SNP position compared to conventional pyrosequencing. Take variant rs6717546 as an example, in the case of heterozygous G/A, two codes equal to 0.5 peak height of C and T were obtained in the SNP position by conventional pyrosequencing (Fig. 1c). In contrast, two two-color codes equal to 1 peak height of CG and TG were obtained when di-bases AG/CG/TG are sequentially added (Fig. 2h). Obviously, the signal intensities of pyrosequencing with di-base addition are twice that of conventional pyrosequencing.

Verification of the method for SNP genotyping

To evaluate the feasibility of this method for SNP genotyping, one SNP (rs6717546) of the UGT1A1 gene was genotyped. The coding single nucleotide polymorphism of rs6717546 involved either a G or A (Fig. 3a, b). According to the principle described above, di-base dispensation order AG/CG/CT/AC was applied and the predicted two-color codes for each allelic variant were shown in ESM Table S2. The first peak equal to one base incorporated was functioned as a reference counterpart and set to correspond to one peak equivalent. To estimate the precision of SNP genotyping, mean values and the relative standard deviation (RSD) of ratios between peaks of the variable position and the reference counterpart were calculated (Table 2). In the case of homozygous G (Fig. 3c), the mean values of ratio between peak CG and reference counterparts were 1.882. In the heterozygous case (Fig. 3d), the mean values of ratios between the peak CG and reference counterparts as well as between the peak GT and reference counterparts were 0.925 and 0.946, respectively. The signal intensities of this method were consistent with the expected values. Furthermore, mean values of ratio between peaks of homozygous G in pyrosequencing with di-base addition and in conventional pyrosequencing were 2.32 (16.84/7.27). Mean values of ratios between peaks of heterozygous G and

Table 2 Accuracy and precision of SNP genotyping by pyrosequencing with di-base addition

dbSNP	Variant	Homozygotes (B1s/B1)[a]			Heterozygotes (B1a/B1)[b]		
		ER	AM	RSD	ER	AM	RSD
rs4148329	C>T	2	1.882	0.079	1	0.925	0.038
		0	0.034	0.002	1	0.946	0.001
rs4148323	G>A	2	1.896	0.097	0.5	0.552	0.024
		0	0.085	0.004	1.5	1.451	0.063
rs56059937	T>C	2	1.848	0.105	n.d.	n.d.	n.d.
		0	0.225	0.023	n.d.	n.d.	n.d.
rs6717546	A>G	1	0.921	0.005	0.5	0.554	0.024
		0	0.043	0.004	0.5	0.533	0.054
rs17862881	A>G	2	2.055	0.035	n.d.	n.d.	n.d.
		2	1.954	0.048	n.d.	n.d.	n.d.
rs4148329	C>T	1	0.932	0.085	0.5	0.578	0.038
		0	0.143	0.024	0.5	0.641	0.066
		1.5	1.371	0.102	1.5	1.459	0.044

ER expected ratio, *AM* arithmetic mean, *RSD* relative standard deviation (the number limit for this operation was set to three), *n.d.* not determined

[a] Ratio between a homozygous SNP and a reference counterpart

[b] Ratio between the base of the respective SNP (in the heterozygous samples) and a non-allelic counterpart

A in pyrosequencing with di-base addition and in conventional pyrosequencing were 3.05 (16.16/5.29) and 3.02 (16.54/5.48), respectively. Therefore, in both allelic cases, the peak heights were twice higher than that of traditional pyrosequencing. It pointed out that this method would enhance the signal intensity thus improving sensitivity. Additionally, in most instances, RSD values for the ratio between key peaks of the respective SNPs and reference counterparts were about 0.1 or lower. It was consistent with the other report in the literature [16]. Sanger DNA technique was applied to further validate the results (see ESM Fig. S2). As shown, the similarities between raw data and the two sets of predicted two-color codes patterns (see ESM Table S2) allowed for direct and reliable SNP determination.

To further verify that this method could enhance the signal intensity so as to improve the detection sensitivity, tenfold serial dilutions of one DNA template were interrogated by conventional pyrosequencing and by pyrosequencing with di-base addition, respectively. The mean signal intensities from the first peak corresponding to one nucleotide incorporation in each assay were shown in Fig. 3e. A decrease in the amount of DNA templates led to decreasing signal intensities. The proportions of pyrosequencing with di-base addition to conventional pyrosequencing in the dilutions 10^4, 10^3, 10^2, 50, and 10^1 were 1.56 (19.99/12.78), 1.60 (12.12/7.58), 1.66 (8.77/5.29), 2.09 (2.68/1.28), and 1.62 (0.86/0.52), respectively. The peaks of pyrosequencing with di-base addition were about 1.6 times the height of those in conventional pyrosequencing. Therefore, when relatively low amount of DNA templates was analyzed, this technology would enhance the sensitivity due to much higher signal intensities.

Simultaneous genotyping of multiple SNPs on a single template

The capacity that conventional pyrosequencing allows to read through up to 20–25 consecutive bases enhances the possibility to apply this methodology for SNP identification.

Fig. 4 Assessment of two SNPs (rs4148323, rs56059937) by conventional pyrosequencing and by pyrosequencing with di-base addition. *Arrowhead doublets* indicated the location of SNPs. Analysis of samples containing homozygous G/G and T/T genotypes of rs4148323 and rs56059937 by conventional pyrosequencing (**a**) and by pyrosequencing with di-base addition (**b**). Analysis of samples containing heterozygous G/A and homozygous T/T genotypes of rs4148323 and rs56059937 of rs4148323 and rs56059937 by conventional pyrosequencing (**c**) and by pyrosequencing with di-base addition (**d**)
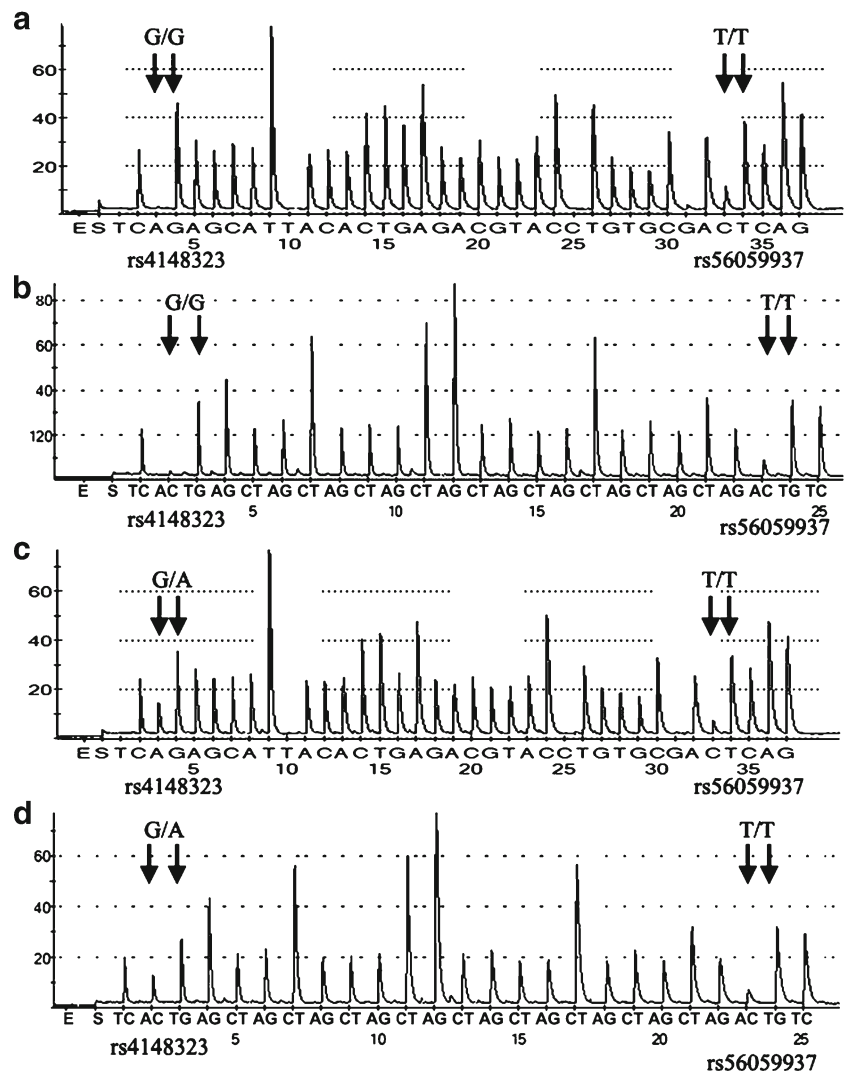
However, it only enables to identify several closely located SNPs in a single run. If several SNPs on a single template are genotyped in a single PCR reaction and a single 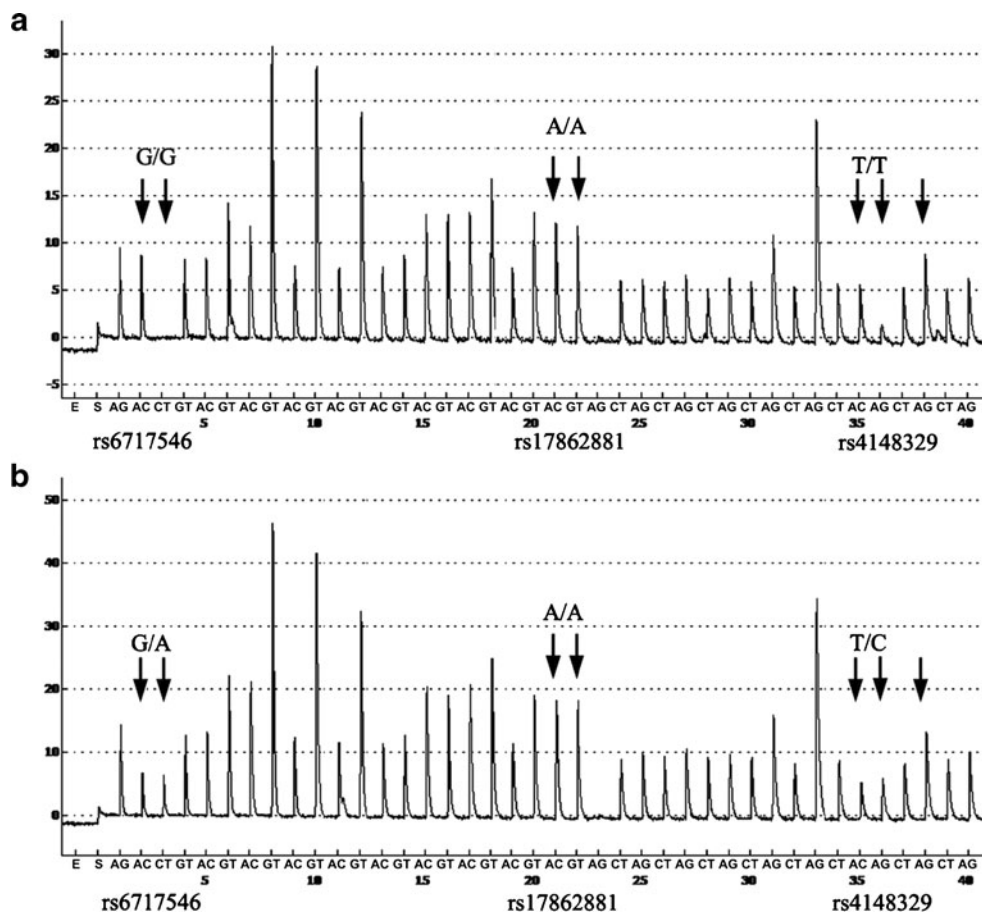sequencing run instead of multiple PCR reactions and multiple sequencing runs, the costs would be dramatically reduced. The technology was able to read 0.9 bp per flow, whereas conventional pyrosequencing was able to read ~0.5 bp per flow [18]. Therefore, it had the potential in simultaneously genotyping several SNPs on a single template. Two variants rs4148323 and rs56059937 which were associated with Gilbert syndrome [10] were genotyped by using pyrosequencing with di-base addition. Sanger DNA technique results were shown in Fig. S3 in the ESM. According to the principle described above, the optimal di-base addition order and the predicted color codes were shown in Table S3 in the ESM. As shown in Fig. 4, compared with the predicted two-color codes, the peak patterns in the programs were exactly as expected. In both the two methods, allele extensions were in phase. To read through the sequence, conventional pyrosequencing required 35 cycles (Fig. 4a, c), while pyrosequencing with di-base addition needed 25 cycles (Fig. 4b, d). Thus, one advantage of this technology is that SNP analysis will be finished nearly 1.4 times as fast compared to conventional pyrosequencing, providing an even more rapid interpretation. For variant

rs4148323, the sequential di-base addition generated differences in two peak positions. In the case of homozygous G, the allelic peak was 1.986 times the height of the reference counterpart. In the heterozygous case G/A, the allelic peaks attained about 55 and 145 % of the height of the non-variable counterparts, respectively (Table 2). Allelic peaks were not evenly distributed since a nucleotide G was identical to one of the alterative bases of the variant rs4148323. For variant rs56059937, only homozygous cases were found in our tested samples. When homozygous T occurred, allelic peak was 1.848 times the height of the reference counterpart (Table 2). Although observed peak ratios slightly deviated from the expected ones in the second SNP position, both the two variations were accurately determined.

To further evaluate the technology for SNP differentiation, a 268-bp PCR fragment that harbored three SNPs (rs6717546, rs17862881, rs4148329) was investigated by using only a sequencing primer in a signal sequencing run (Fig. 5). The template was read through by up to 40 cycles, and each allelic variant had visual differences in the peak patterns. The genotype of rs17862881 in all the analyzed samples was homozygous A. In all the cases, RSD values commonly attained figures below 0.1; only a few scattered figures were above this level (Table 2). The predicted two-color codes for the three

Fig. 5 Assessment of three SNPs (rs6717546, rs17862881, rs4148329) using pyrosequencing with di-base addition. *Arrows* indicate SNP positions. (**a**) Pyrogram from homozygous G/G, A/A, and T/T genotypes of rs6717546, rs17862881, and rs4148329. (**b**) Pyrogram from heterozygous G/A, homozygous A/A, and heterozygous T/C genotypes of rs6717546, rs17862881, and rs4148329

variants (see ESM Table S4) corresponded to the peak patterns in Fig. 5. Sanger sequencing results of the three variants were also shown in ESM Fig. S4, and it further pointed out that the genotypes were consistent with the raw genotypes. Albeit with a slightly reduced conformity between expected and observed peak ratios at the distal end of this sequence context, these variations were accurately determined with high clarity.

Optimization for SNP analysis

In Fig. 5, to avoid the out of phase in the SNP positions, the read length per cycle is one base pair in most reactions although a di-base is added at a time. It thus may give rise to much shorter read length. To optimize this method for simultaneous analysis of multiple SNPs, we add sequence-specific di-bases into the incorporations to extend more than one base but no more than five bases at a time. Routinely, about 60 bases can be interrogated using the commercial instruments PSQ 96MA and HS 96 [12]. In order to further evaluate the potential of this method for genotyping multiple SNPs flanking more than 60 bp, three SNPs (rs4148324, rs3771342, rs6742078) covering 84 bp were investigated by pyrosequencing with di-base addition (Fig. 6a) and by conventional pyrosequencing (Fig. 6b). Sanger sequencing was applied to further validate the results (see ESM Fig. S5). For conventional pyrosequencing, as the incorporations took

place as expected, there was a continuous loss in signal intensities. After 45 cycles, the signals were indistinguishable. If these SNPs were interrogated, up to 67 cycles would be needed. As for pyrosequencing with di-base addition, the proper dispensation order was predicted according to the aforementioned principle and the resultant two-color codes were shown in Table S4 in the ESM. When di-bases were added to extend two to five bases at a time, the length of 84-bp covering three SNPs was successfully interrogated by only a total of 46 cycles. The peak patterns were consistent with the predicted two-color codes. Compared to traditional pyrosequencing, the peak heights were approximately twice the height of conventional pyrosequencing, even though the overall peak heights dropped with increasing read length. Furthermore, the read length of this technology was nearly 1.4 times as long as that of conventional pyrosequencing. However, most of the peaks were equal to one base from the 20th to 35th cycles. If the di-bases in these cycles were able to extend two to five bases, the sequencing cycles would be further reduced and read length would be further increased.

In generally, one challenge of this methodology is to avoid out-of-phase signals when examining heterozygous alleles. When the sequence is known, this problem can be overcome by sequence-specific di-base dispensation as aforementioned. To assess the accuracy, we calculated RSD and mean values of ratios between the base of a represent variant and a reference

Fig. 6 Analysis of three SNPs (rs4148324, rs3771342, and rs6742078) covering 84 bp by pyrosequencing with di-base addition (*upper*) and by conventional pyrosequencing (*lower*). Each SNP is indicated by *two arrowhead doublets*. Pyrograms from analysis of heterozygous G/T, homozygous G/G, and heterozygous G/T genotypes of rs4148324, rs3771342, and rs6742078 by pyrosequencing with di-base addition (**a**) and by conventional pyrosequencing (**b**)

counterpart and compared peak patterns to the expected two-color codes. A high concordance with pyrosequencing calls for most detected SNPs was observed in most cases. In Fig. 5a, b, a minus frame shift appeared. Peak AG in the 38th cycle was higher than peak AG in the 36th cycle because of the extra signals produced by the incomplete nucleotide incorporation reactions in the previous steps. The frame shift might be reduced by decreasing the apyrase concentration or by interrogating the fragments in the reverse direction.

When optimal amount of immobilized DNA was used to read relatively long bases by pyrosequencing, the decreasing signal intensities occurred. As shown in Fig. 6, signal intensities decreased gradually with read length increasing. This might be caused by multiple factors. First, the exonuclease activity of DNA polymerase would cause primer degradation, leading to out of phase sequencing [13], second an accumulation of by-product during sequencing [12], and third the dilution produced by continuous addition of nucleotides [12]. When the number of sequencing cycles was reduced by using this technology, these impacts might be relatively weakened.

The read length of conventional pyrosequencing limited it to determine the phase of SNPs provided they were closely located [19]. However, an inherent capacity of this method was to increase read length, making simultaneously examine several SNPs in a single run possible as aforementioned. In previously published report, the variants rs4148323 and rs56059937 were genotyped by using two sequencing primers in two sequencing runs [12]. In contrast, the two SNPs were differentiated in this work by only a sequencing primer in a single run. Besides, two SNPs rs4148329 and rs6717546 were genotyped by two sequencing runs [20], whereas only a single sequencing run was applied to genotype the two SNPs and one additional SNP here. Thus, it would decrease the cost for template preparation and SNP genotyping. As for the number of potentially genotyped SNPs, three SNPs were simultaneously genotyped here. However, more SNPs were likely to be genotyped as long as these SNPs were located within the read length of this method.

Conclusion

In summary, a novel pyrosequencing with di-base addition has been developed for SNP genotyping. Sequence-specific di-base dispensation order was proposed to avoid nonsynchronous extension. Data presented suggested that the di-base addition could be designed successfully to any SNP, and allelic variants of a SNP could easily be distinguished by pattern recognition (color codes and peaks). The advantage of much longer read length enabled to simultaneously genotype multiple SNPs in a single run. More importantly, the ability to amplify detectable signals could be exploited to develop alternative approaches to detect SNPs in rare DNA samples. We believe that it is a promising methodology for multiple SNPs genotyping in lower amount of DNA samples, expanding its application potential in medical diagnostics and forensic identification.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no competing interests.

References

1. Lin YS, Liu FG, Wang TY, Pan CT, Chang WT, Li WH. A simple method using Pyrosequencing™ to identify de novo SNPs in pooled DNA samples. Nucleic Acids Res. 2011;39(5):e28–38. doi:10.1093/nar/gkq1249.

2. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem. 1996;242(1):84–9. doi:10.1006/abio.1996.0432.

3. Ahmadian A, Ehn M, Hober S. Pyrosequencing: history, biochemistry and future. Clinica Chimica Acta; Int J Clin Chem. 2006;363(1–2):83–94. doi:10.1016/j.cccn.2005.04.038.

4. Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyren P, Uhlen M, et al. Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem. 2000;280(1):103–10. doi:10.1006/abio.2000.4493.

5. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323(5910):133–8. doi:10.1126/science.1162986.

6. Innings A, Krabbe M, Ullberg M, Herrmann B. Identification of 43 Streptococcus species by pyrosequencing analysis of the rnpB gene. J Clin Microbiol. 2005;43(12):5983–91. doi:10.1128/JCM.43.12.5983-5991.2005.

7. Rajatileka S, Luyt K, Williams M, Harding D, Odd D, Molnar E, et al. Detection of three closely located single nucleotide polymorphisms in the EAAT2 promoter: comparison of single-strand conformational polymorphism (SSCP), pyrosequencing and Sanger sequencing. BMC Genet. 2014;15:80–92. doi:10.1186/1471-2156-15-80.

8. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, et al. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 2010;472:431–55. doi:10.1016/S0076-6879(10)72001-2.

9. Banelli B, Morabito A, Laurent S, Piccioli P, Dozin B, Ghio M, et al. A novel multiplex pyrosequencing assay for genotyping functionally relevant CTLA-4 polymorphisms: potential applications in autoimmunity and cancer. Hum Immunol. 2014;75(8):730–9. doi:10.1016/j.humimm.2014.04.019.

10. Saeki M, Saito Y, Jinno H, Tohkin M, Kurose K, Kaniwa N, et al. Comprehensive UGT1A1 genotyping in a Japanese population by pyrosequencing. Clinical chemistry. 2003;49(7):1182–5. doi:10.1373/49.7.1182.

11. Mercke Odeberg J, Andrade J, Holmberg K, Hoglund P, Malmqvist U, Odeberg J. UGT1A polymorphisms in a Swedish cohort and a human diversity panel, and the relation to bilirubin plasma levels in

males and females. Eur J Clin Pharmacol. 2006;62(10):829–37. doi:10.1007/s00228-006-0166-3.

12. Mashayekhi F, Ronaghi M. Analysis of read length limiting factors in Pyrosequencing chemistry. Anal Biochem. 2007;363(2):275–87. doi:10.1016/J.Ab.2007.02.002.

13. Gong M, Foo SH, Lin L, Liu ET, Gharizadeh B, Goel S. Pyrosequencing enhancement for better detection limit and sequencing homopolymers. Biochem Biophys Res Commun. 2010;401(1):117–23. doi:10.1016/j.bbrc.2010.09.022.

14. Pourmand N, Elahi E, Davis RW, Ronaghi M. Multiplex pyrosequencing. Nucleic Acids Res. 2002;30(7):e31–e5. doi:10.1093/nar/30.7.e31.

15. Kim KA, Song WG, Lee HM, Joo HJ, Park JY. Multiplex pyrosequencing method to determine CYP2C9*3, VKORC1*2, and CYP4F2*3 polymorphisms simultaneously: its application to a Korean population and comparisons with other ethnic groups. Mol Biol Rep. 2014;41(11):7305–12. doi:10.1007/s11033-014-3617-4.

16. Alderborn A, Kristofferson A, Hammerling U. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res. 2000;10(8):1249–58. doi:10.1101/gr.10.8.1249.

17. Langaee T, Ronaghi M. Genetic variation analyses by Pyrosequencing. Mutat Res. 2005;573(1–2):96–102. doi:10.1016/j.mrfmmm.2004.07.023.

18. Pu D, Qi Y, Cui L, Xiao P, Lu Z. A real-time decoding sequencing based on dual mononucleotide addition for cyclic synthesis. Anal Chim Acta. 2014;852:274–83. doi:10.1016/j.aca.2014.09.009.

19. Fakhrai-Rad H, Pourmand N, Ronaghi M. Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. Hum Mutat. 2002;19(5):479–85. doi:10.1002/humu.10078.

20. Hong AL, Huo D, Kim HJ, Niu Q, Fackenthal DL, Cummings SA, et al. UDP-glucuronosyltransferase 1A1 gene polymorphisms and total bilirubin levels in an ethnically diverse cohort of women. Drug Metab Dispos. 2007;35(8):1254–61. doi:10.1124/dmd.106.014183.