Ming Chuan-Health Tech Journal, 2010, Vol. 1, No. 1, e2

Genotyping of multi-copy genes by pyrosequencing

Yih-Horng Shiao *

Laboratory of Comparative Carcinogenesis, National Cancer Institute at Frederick, MD 21702

Received 17 November 2009/ Published 1 January 2010

ABSTRACT

Pyrosequencing is a sequencing-by-synthesis method and provides a quantitative assay for sequence variants, including single nucleotide polymorphisms (SNPs). For single-copy genes, SNP frequency is dichotomous, either 50% for heterozygotes or 100% for homozygotes. For genotyping of a multi-copy gene, such as the 45S ribosomal RNA gene, the frequency of a specific SNP can be theoretically continuous, ranging from 0 to 100%. Pyrosequencing is a method of choice to produce high-resolution quantitative data to distinguish frequency differences of less than 5%. This provides a platform for identification of unique features of the diverse genome, which may link to disease processes. The advantages and current challenges of pyrosequencing are discussed.

INTRODUCTION

Genotyping of single nucleotide polymorphisms (SNPs) or sequence variations is critical for several types of investigation, such as whole-genome association study [1], pharmacogenetic surveillance [2,3], forensic science [4], and differentiation of microorganisms [5,6]. Many technologies are available for high-throughput genotyping [2,3]. They can be categorized according to chemistries for allele

discrimination, such as primer extension, hybridization, ligation, and enzymatic cleavage, and for allele detection, for examples, fluorescence, mass spectrometry, and chemiluminescence. Their principles, applications, advantages, and limitations have been described to some extent [2,3]. In this review, utility of the pyrosequencing method for genotyping of SNPs in multi-copy genes is discussed. For single-copy genes, a SNP presents in paternal and/or maternal alleles, giving a SNP frequency of 50% for heterozygotes or 100% for homozygotes. However, a multi-copy gene can have various permutations with allele frequency ranging 0-100% (Fig. 1). As exemplified in the figure, some genotypes with 50% SNP frequency are homozygotic just like those of 100% SNP frequency. The permutation of possible genotypes increases when the copy and SNP numbers elevate. Examples of multi-copy genes are the rRNA, tRNA, repetitive sequences, pseudogenes, and the C4 complement gene [7,8]. In addition, growing evidence of copy number variation (CNV) in mammalian genomes suggests that genes having more than one copy are not rare [9]. If a SNP is present in CNV, the diversity of CNV can be detected by SNP frequency. Since the SNP frequency is a continuous value for multi-copy genes, a method with high resolution to detect small frequency differences, such as pyrosequencing, is needed for accurate genotyping.

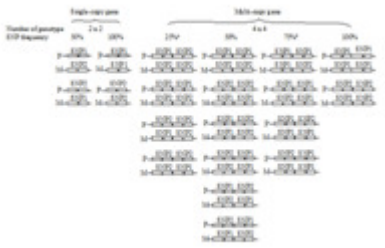

FIG. 1. Possible number of genotypes, equivalent to the number of paternal (P) alleles multiplied by the number of maternal (M) alleles, and SNP frequency (percentage for either SNP1 or SNP2 except a and b representing SNP1 only).

THE CHEMISTRY OF PYROSEQUENCING

Pyrosequencing is a solid-phase sequencing-by-synthesis method. The DNA sequence is a readout of serial light intensity generated through a cascade of 4-enzyme cocktail reactions with no need for gel electrophoresis [10]. Since the reaction takes place at 28 ^{o}C, a specific primer and a template free of self-priming hairpin structures are key parameters to ensure sequencing specificity. The Assay Design software, provided with the pyrosequencer, was developed to provide scoring for various parameters, such as self-priming, hairpin formation, false priming, and homopolymers, to assist primer selection for polymerase chain reaction and pyrosequencing. Other software capable of performing the same analyses can also be used.

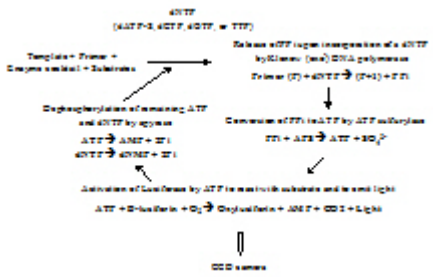

FIG. 2. Pyrosequencing reaction in the automatic pyrosequencer. It starts from additions of the 4-enzyme cocktail and subsequent substrates into a 96-well plate containing immobilized DNA template and primer. Individual nucleotides are dispensed sequentially and initiate cascade of enzymatic reactions if a nucleotide is incorporated. dNTP: deoxyribonucleotide; APS: Adenosine phosphosulfate.

A schematic representation of the sequential enzymatic reactions is depicted in Fig.2. Biotin-labeled single-stranded DNA fragments are immobilized on sepharose beads and a sequencing primer with high specificity is annealed to a sequence immediately upstream region of a SNP site. In some cases, multiple SNPs can be interrogated by a single primer. During the reaction, one nucleotide is dispensed at a time. If the nucleotide is incorporated by the Klenow (exo^-) DNA polymerase, pyrophosphate (ppi) is released. The amount of ppi is proportional to the number of nucleotides that are added into the nascent strand. The recombinant ATP sulfurylase, of yeast origin, converts ppi along with adenosine phosphosulfate to ATP. ATP subsequently activates firefly luciferase to act on D-luciferine with resultant emission of light. The remaining unused nucleotide is finally digested by apyrase. This enzymatic cascade is kinetically optimized to carry out serial reactions in the same tube. The volume increases after addition of individual nucleotides. If the dispensed nucleotide is not complementary to the template, incorporation and release of ppi do not occur. This 4-enzyme sequential reaction takes about 1 minute. The next cycle begins with addition of a new nucleotide. The light from each cycle is captured by a charge-coupled device (CCD) camera and is registered as a peak in a pyrogram (Fig 3). The peak height is equivalent to the number of nucleotides incorporated at each dispensation. A homopolymer of 5 consecutive nucleotides is the accepted limit for accurate estimation of the number of the same nucleotide in sequence. For learning the methodology in motion pictures, readers can access the reference 11 and the pyrosequencing technology web site, http://www.pyrosequencing.com/DynPage.aspx?id=7454, to view the demonstration.

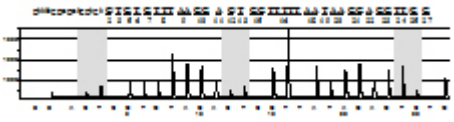

FIG. 3. Example of a pyrogram (E: enzyme mix; S: substrates; number: dispensation order) from incorporated nucleotides (letters in capital) extended from a primer (letters in lower case) for 3 C/T variant sites.

For determination of a specific SNP or mutation, the sequence for interrogation should be known. Two nucleotides at a SNP site are automatically highlighted to indicate the location. A newer model is equipped with software to differentiate more than 2 sequence variants at the same SNP site. A computer algorithm calculates the percentage of SNP and provides quality evaluation, either Pass, Check, or Failed, using reference peaks from an internal sequence. The user is also allowed to set criteria manually for data analysis. The nucleotide A peak is calculated with 86% signal by default to adjust for the noise coming from direct interaction of dATP with luciferase. Pyrosequencing with a 96-well format facilitates the SNP quantification. It also provides excellent resolution for separation of less than 5% frequency difference; this is essential for detection of small differences that may be of significance for particular diseases.

HIGH-RESOLUTION GENOTYPING BY PYROSEQUENCING

Many techniques are available for determination of SNPs and their frequencies [12,13]. Either PCR- or hybridization-based approaches rely on a sophisticated algorithm to estimate SNP frequency based on a single-copy gene as depicted in Fig. 1. Such a methodology can be high-throughput, but it is not suitable for genotyping of multi-copy genes that require a high-resolution technique to differentiate diverse SNP frequencies other than just 50% and 100%. It is extremely challenging to quantify SNP frequency from multiple copies and possibly more than 2 types of SNPs at the same site. The largest probable SNP frequency can be calculated as copy number multiplied by the number of SNP types. For a 100-copy gene from a single allele with 2 SNPs at one site, the frequency for a SNP, in theory, can be 0/(100 x 2), 1/(100 x 2), 2/(100 x 2), …, or (100 x 2)/(100 x 2). The real genome may not carry SNPs in every copy.

For example, the rRNA gene is highly abundant in prokaryote and eukaryote organisms. The major rRNA gene in mammalian cells is about 45-kb in each copy and generates nearly 13.6 kb precursor transcript for subsequent processing to produce mature 18S, 5.8S, and 28S rRNAs. It is tandemly repeated >200 times in mice and >400 times in humans across 5 chromosomes [14-16]. SNPs have been observed in transcribed regions of the rRNA gene [17]. SNPs in the promoter regions have regulatory functions related to gene transcription [18]. In a previous study, a SNP in the mouse rRNA spacer promoter was identified by cloning and DNA sequencing [19]. However, the estimation of SNP frequency from about 20 clones of each amplified product may not be accurate because of insufficient representation of over 200 copies of the gene in the mouse genome. Pyrosequencing was used to quantify the SNP frequency directly from all amplified products [20]. An example of the result is shown as a pyrogram in Fig. 4. A large discrepancy was observed between pyrosequencing and estimation from cloning/sequencing. The frequency of the T SNP at the -2214 site of the rRNA was in general below 50% and the difference between experimental samples was as low as a few percentages. The run-to-run variation of pyrosequencing is less than 5% on average. The advantage of the high-degree resolution of pyrosequencing has been also reviewed by others [12].

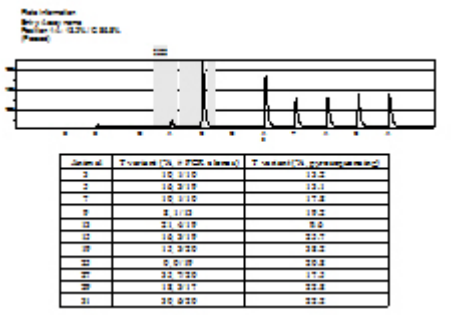

FIG. 4. Example of a pyrogram in antisense direction from the genotyping of the T/G SNP at -2214 site of the rRNA spacer promoter in the Cr:NIH Swiss mouse strain. The frequencies of the T variant determined by both cloning/sequencing and pyrosequenceing were compared in sperm DNA of 11 animals.

PYROSEQUENCING FOR MULTI-COPY GENE DISCOVERY

I In addition to the rRNA, many multi-copy genes have been observed in mammalian genomes, including GAPDH, tRNA, repetitive sequences, and pseudogenes [7]. The C4 complement and CYP2D6 in humans, and Dominant white/KIT in pigs are also multi-copy genes [8,21,22]. Pyrosequencing protocols have been developed to examine SNP frequency in the CYP2D6 gene [23]. Copy number variation (CNV) in the

human genome is also a growing area of research. Combinations of SNPs and copy number generate highly diversified genomes and distinguish individuals. This high-degree variation may play a fundamental role in individual susceptibility to diseases. Detection of CNV is commonly carried out using high-throughput genome-wide array-CGH and SNP array [25]. Pyrosequencing has been also used to determine the size variants of small tandem repeats [26,27]. If a specific SNP is present in different copies, pyrosequencing has the potential to identify intra-genome duplication when the percentage of SNP is not either 50% or 100%, except when the gene carrying equal number of SNPs as depicted in Fig. 1. Pyrosequencing has been used to examine SNPs in CNV to differentiate number, such as Dominant white/KIT in pigs [22,28].

CHALLENGES AND CONCLUSION

Pyrosequencing surely offers great discriminating power for genotyping of multi-copy genes. Several inherent challenges of the current platform may have discouraged some users. The chemiluminescence signal saturates for homopolymers of over 5 consecutive nucleotides. The peak after addition of 2'-deoxyadenosine 5'-thiotriphosphate is estimated by multiplying raw intensity with 86% to deduct the background contributed by direct interaction of the nucleotide with luciferase. The other 3 nucleotides (2'-deoxyguanosine 5'-triphosphate, 2'-deoxycytidine 5'-triphosphate, and thymidine 5'-triphosphate) do not yield noticeable background. Only about 100 nucleotides can be interrogated without losing the signal and resolution. The run-to-run reproducibility decreases for the SNP at locations away from the first dispensation. These inherent limitations are primarily caused by the increase of volume as a result of consecutive dispensations. Maintaining of the volume and activities of enzymes are the subjects of further improvement for the pyrosequencing technology.

ACKNOWLEDGEMENT

This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

FOOTNOTES

* Correspondence: Yih-Horng Shiao, Ph.D., Building 538, Room 205A, NCI-Frederick, National Institutes of Health, West 7th Street, Frederick, MD 21702, USA. Tel: 301-846-1246; E-mail: shiaoy@mail.nih.gov

REFERENCES

1. Gibbs J.R., and Singleton A. 2006. Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS Genet. **2**:e150.
2. Twyman, R.M., and Primrose, S.B. 2003. Techniques patents for SNP genotyping. Pharmacogenomics **4**:67-79.
3. Kim, S., and Misra, A. 2007. SNP genotyping: Technologies and biomedical applications. Annu. Rev. Biomed. Eng. **9**:289-320.
4. Budowle, B., and van Daal, A. 2008. Forensically relevant SNP classes. BioTechniques **44**:603-610.
5. Vernet, G. 2007. Use of molecular assays for the diagnosis of influenza. Expert Rev. Anti. Infect. Ther. **5**:89-104.
6. Crawford, J.T. 2003. Genotyping in contact investigations: a CDC perspective. Int. J. Tuberc. Lung Dis. **7**:S453-S457.
7. Richard, G.F., Kerrest, A., and Dujon, B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. **72**:686-727.
8. Stewart, C.A., Horton, R., Allcock, R.J.N., Ashurst, J.L., Atrazhev, A.M., Coggill, P., Dunharm, I., Forbes, S., Halls, K., Howson, J.M.M., Humphray, S.J., Hunt, S., Mungall, A.J., Osoegawa, K., Palmer, S., Roberts, A.N., Rogers, J., Sims, S., Wang, Y., Wilming, L.G., Elliott, J.F., de Jong, P.J., Sawcer, S., Todd, J.A., Trowsdale, J, and Beck, S. 2004. Complete MHC haplotype sequencing for common disease gene mapping. Genome Res. **14**:1176-1187.
9. Carter, N.P. 2007. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. **39**:S16-S21.
10. Ahmadian, A., Ehn, M., and Hober, S. 2006. Pyrosequencing: History, biochemistry and future. Clin. Chim. Acta **363**:83-94.

11. King, C., and Scott-Horton, T. 2008. Pyrosequencing: a simple method for accurate genotyping. J. Vis. Exp. (11) pii:630.
12. Isler, J.A., Vesterqvist, O.E., and Burczynski, M.E. 2007. Analytical validation of genotyping assays in the biomarker laboratory. Parmacogenomics **8**:353-368.
13. Ragoussis, J. 2009. Genotyping technologies for genetic research. Annu. Rev. Genomics Hum. Genet. **10**:5.1-5.17.
14. Elsevier, S.M., Ruddle, F.H. 1975. Location of genes coding for 18S and 28S ribosomal RNA within the genome of Mus musculus. Chromosoma **52**:219-228.
15. Grozdanov, P., Georgiev, O., Karagyozov, L. 2003. Complete sequence of the 45-kb mouse ribosomal DNA repeat: analysis of the intergenic spacer. Genomics **82**:637-643.
16. Worton, R.G., Sutherland, J., Sylvester, J.E., Willard, H.F., Bodrug, S., Dubé, I., Duff, C., Kean, V., Ray, P.N., Schmickel, R.D. 1988. Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5' end. Science **239**:64-68.
17. Leffers, H., and Andersen, A.H. 1993. The sequence of 28S ribosomal RNA varies within and between human cell lines. Nucleic Acids Res. **21**: 1449-1455.
18. Tseng, H., Chou, W., Wang, J., Zhang, X., Zhang, S., and Schultz, R.M. 2008. Mouse ribosomal RNA genes contain multiple differentially regulated variants. PLoS ONE **3**:e1843.
19. Shiao, Y.H., Crawford, E.B., Anderson, L.M., Patel, P., and Ko, K. 2005. Allele-specific germ cell epimutation in the spacer promoter of the 45S ribosomal RNA gene after Cr(III) exposure. Toxicol. Appl. Pharmacol. **205**:290-296.
20. Patel, P., Shiao, Y.H, and Fortina, P. 2007. Multiplex pyrosequencing for DNA variation analysis. Methods Mol. Biol. **373**:75-88.
21. Lundqvist, E., Johansson, I., and Ingelman-Sundberg M. 1999. Genetic mechanisms for duplication and multiduplication of the human CYP2D6 gene and methods for detection of duplicated CYP2D6 genes. Gene **226**:327-338.
22. Pielberg, G., Day, A.E., Plastow, G.S., and Andersson, L. 2003. A sensitive method for detecting variation in copy numbers of duplicated genes. Genome Res. **13**:2171-2177.
23. Zackrisson, A., and Lindblom, B. 2003. Identification of CYP2D6 alleles by single nucleotide polymorphism analysis using pyrosequencing. Eur. J. Clin. Pharmacol. **59**:521-526.
24. Henrichsen, C.N., Chaignat, E., and Reymond, A. 2009. Copy number variants, diseases, and gene expression. Hum. Mol. Genet. **18**:R1-R8.
25. Beaudet, A.L., and Belmont, J.W. 2008. Array-based DNA diagnostics: Let the revolution begin. Annu. Rev. Med. **59**:421-436.

26. Schentrup, A.M., Allayee, H., Lima, J.J., Johnson, J.A., and Langaee, T.Y. 2009. Genet. Test Mol. Biomarkers **13**:361-365.
27. Edlund, H., and Allen, M. 2008. Y chromosomal STR analysis using pyrosequencing technology. Forensic Sci. Int. Genet. **3**:119-124.
28. Lee, J.H., and Jeon, J.T. 2008. Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. Cytogenet. Genome Res. **123**:333-342.

Ming Chuan-Health Tech Journal, 2010, Vol. 1, No. 1, e2