Cancer Letters 403 (2017) 152–158

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Original Article

A novel method to quantify base substitution mutations at the 10^{-6} per bp level in DNA samples

Satoshi Yamashita [a], Naoko Iida [a], Hideyuki Takeshima [a], Naoko Hattori [a], Masahiro Maeda [a], Takayoshi Kishino [a], Reiko Nagano [a], Taichi Shimazu [b], Shoichiro Tsugane [b], Toshikazu Ushijima [a, *]

[a] Division of Epigenomics, National Cancer Center Research Institute, Tokyo, Japan
[b] Epidemiology and Prevention Group, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan

ARTICLE INFO

Article history:
Received 7 April 2017
Received in revised form
5 June 2017
Accepted 9 June 2017

Keywords:
Point mutation
Next-generation sequencing
Target sequencing
Neutral theory

ABSTRACT

Somatic base substitution mutations of frequencies at the 10^{-6}/bp level are expected to be present in many biomedical samples, such as tissues exposed to carcinogenic factors and exhausted stem cells. However, measurement of such rare mutations has been very difficult in human DNA samples. Here, we invented the use of 100 copies of genomic DNA as a template for amplicon deep sequencing so that a real mutation in a single DNA molecule would be detected at a variant allele frequency of 1% while sequencing errors have less frequency. In addition, we selected 15,552 error-resistant base positions whose mutation frequency was expected to reflect that of base positions that can drive carcinogenesis or potentially even of the entire genome. The validity of the method was first confirmed by the successful detection of mutations premixed at the frequency of 0.1%. Second, increasing mutation frequencies ($4–60 \times 10^{-6}$/bp) were successfully detected in cells treated with increasing doses of one of two mutagens, and their signature mutations were detected. The ratio of non-synonymous mutations to synonymous mutations time-dependently decreased after treatment with a mutagen, supporting the neutral theory of molecular evolution for somatic mutations. Importantly, gastric mucosae exposed to *Helicobacter pylori* infection was shown to have significantly higher mutation frequency than those without. These results demonstrated that our new method can be used to measure rare base substitution mutations at the 10^{-6}/bp level, and is now ready for a wide range of applications.

Introduction

Quantification of mutations present in a tiny fraction of a DNA sample is important for a variety of biomedical research. Most prominently, it is considered that small amounts of mutations have already accumulated in normal-appearing tissues after exposure to carcinogenic factors and also by aging [1,2], and that accumulation levels of mutations in tissues are associated with cancer risk [3,4]. This concept may be applicable to a wide variety of cancers and even to risk estimation of therapy-induced secondary cancers. Also for stem cell research and regenerative medicine, stem cell quality, reflecting a variety of factors, including culture conditions and induction conditions, especially for induced pluripotent stem cells [5,6], may be assessed by the use of accumulation levels of rare somatic mutations. Generation of immunogenic mutations may also be associated with the level of background somatic mutations [7,8], and this may be important to predict responses to immune checkpoint inhibitors.

Such mutations with low frequencies have been analyzed using specialized technologies. Transgenic animals with a marker gene, such as *cII*, *gpt*, *lacI* and *lacZ*, have been used for the analysis of their tissues and cells [9,10]. Endogenous marker genes, such as *HPRT* and *TK*, have been used to select cells with their mutations in cell culture [11]. By the use of these systems, extremely low mutation frequencies in normal and mutagen-treated tissues and cells ($<10^{-4}$/bp) have been reported [12–14]. However, such approaches are not applicable to human tissue samples. Therefore, the use of massively parallel sequencing is promising, but, in practice, rare mutations cannot be distinguished from sequencing errors. To overcome this issue, molecular barcode-based methods, in which

* Corresponding author. Division of Epigenomics, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan.
E-mail address: tushijim@ncc.go.jp (T. Ushijima).

http://dx.doi.org/10.1016/j.canlet.2017.06.010
0304-3835/

individual DNA molecules are uniquely tagged, were established [15–17]. This strategy effectively distinguishes somatic mutations in single molecules from sequencing errors, but needs a large number of sequencing reads, and analysis of a large number of samples is prohibitive due to its high cost.

To develop a cost-effective method, we here invented the use of an extremely small number of DNA copies as a sequencing template (Fig. 1), contrary to the common practice of massively parallel sequencers. When 100 copies of DNA are sequenced, a mutation, if present, is expected to be detected at a variant allele frequency of 1%, while it is reported that sequencing errors are observed at a frequency of ~1% [18]. There is no need to explain that SNPs are detected at frequencies of 50% or 100%. Therefore, when an appropriate cutoff variant allele frequency is set, for example at 1%, it may become possible to distinguish mutations in one of the 100 DNA copies from sequencing errors. Furthermore, if we analyze any mutations in the regions sequenced, they will include many silent mutations without functional selection, and are expected to reflect somatic mutation load in the regions, and potentially even of the entire genome. We will verify that the sequencing strategy can distinguish a somatic mutation in a single molecule from sequencing errors, and use it to show that non-synonymous mutations are selected against.

Materials and methods

Cell lines and drug treatment

The human TK6 lymphoblast cell line, the human N87 gastric cancer cell line, and the HeLa cell line were obtained from the American Type Culture Collection (Manassas, VA). Two human gastric cancer cell lines, MKN1 and NUGC3, were purchased from the Japanese Collection of Research Bioresources (Tokyo, Japan). Two human gastric cancer cell lines, HSC43 and HSC57, were kindly provided by Dr. K. Yanagihara. All cell lines were maintained in RPMI1640 containing 10% (v/v) FBS.

The TK6 cells were treated with a mutagen, *N*-nitroso-*N*-methylurea (MNU, Sigma-Aldrich St. Louis, MO) or 4-nitroquinoline 1-oxide (4-NQO, Sigma-Aldrich). The TK6 cells were seeded at a density of 1×10^5 cells/10-cm plate on day 0, treated with MNU or 4-NQO (0.1, 0.3, 1, 3, 10, and 30 μg/mL) on day 1, and the medium was changed to a normal medium on day 2. The cells were harvested on day 7.

The HeLa cells were synchronized using a double thymidine block protocol [19] before treatment with 4-NQO. Two mM thymidine was added to the culture medium for 18 h. Cells were released from the first block by replacement of the medium with a fresh one. After 9 h, cells were exposed again to 2 mM thymidine. After 15 h, 10 μg/mL 4-NQO was added, and 30 min later cells were released from the second block and 4-NQO by replacement of the medium with a fresh one. The cells at 0 h, 3 h, 9 h, and 24 h after the release were harvested. Genomic DNA was extracted by the phenol/chloroform method.

Human DNA samples

Genomic DNA was extracted from 20 gastric mucosa samples collected endoscopically from individuals who underwent cancer screening at the Research Center for Cancer Prevention and Screening, National Cancer Center, Japan, for a study published elsewhere [20]. The infection status by *Helicobacter pylori* (*HP*) was determined by histologic examination or the rapid urease test (Otsuka, Tokushima, Japan), and 10 were considered to have past *HP* infection. The study was approved by the Institutional Review Boards, and written informed consent was obtained from all participants.

Preparation of sequencing library and massively parallel sequencing

The copy number of genomic DNA in a sample was measured by real-time PCR using three genes (*ALB*, *ARHGEF4*, and *RAPGEFL1*) located outside of genomic regions with reported copy number variations [21]. The primer sequences are listed in Supplementary Table 1. A sequence library (291 regions of 55 cancer-related genes, covering 48,005 base positions; Supplementary Table 1) was prepared by multiplex PCR (22 cycles) using Ion AmpliSeq Library Kits 2.0 (Thermo Fisher Scientific, Waltham, MA) and 100 copies of genomic DNA as a template. Libraries from different samples were uniquely barcoded, and sequenced using an Ion PI chip, Ion PI Hi-Q Sequencing 200 Kit, and Ion Proton sequencing system (Thermo Fisher Scientific) with an average sequencing depth of at least 5000 reads.

Four libraries were prepared from 100 copies of normal human tissue DNA, and sequenced using both an Ion Proton and a HiSeq 2500 (Illumina, San Diego, CA) by paired-end 100-bp sequencing (PE-100).

Calculation of mutation frequencies

The sequences obtained by Ion Proton and HiSeq 2500 were mapped to the human reference genome Hg19 using Ion Suite 5.0 (Thermo Fisher Scientific) and CLC genomics workbench 8.5.1 (Qiagen, Hilden, Germany), respectively. To avoid detecting false mutations, we removed base positions with insufficient amplification or high error frequencies. Among the 48,005 base positions amplified in the library, 15,313 base positions (90 of 291 amplicons) showed insufficient amplification (<2,500x) in three independent analyses of untreated TK6 cells, and the remaining 32,692 base positions (201 amplicons) were used. Error-prone positions were determined as those with a variant allele frequency ≥0.2% in any of the three analyses, and 17,140 positions were removed from the 32,692 positions to obtain 15,552 error-resistant positions.

The cutoff variant allele frequency for the removal of error-prone positions was based upon the Poisson distribution with 0.325% frequency (average base substitution error rate for Ion PGM [22]). Variants with allele frequencies ≥0.3%, 0.2%, and 0.1% were expected to include 56%, 93%, and 99% of frequent errors, respectively, in a 5,000x depth analysis. Experimentally, among the 32,692 base positions, 17,140 and 23,865 positions had variant allele frequencies ≥0.2% and 0.1%, respectively. To balance the number of remaining error-resistant positions (32,692−17,140 = 15,552 for 0.2%; and 32,692−23,685 = 9007 for 0.1%) and sensitivity (93% with 0.2%; and 99% with 0.1%), we adopted a cutoff variant allele frequency of 0.2% (Supplementary Table 2).

The mutation frequency was estimated as the number of somatic mutations within the 15,552 base positions divided by 1,555,200 (= 15,552 positions × 100 molecules). A variant was counted as a somatic mutation only when (i) its frequency was ≥0.8%; (ii) its read count was ≥4; and (iii) it was found in at least two forward

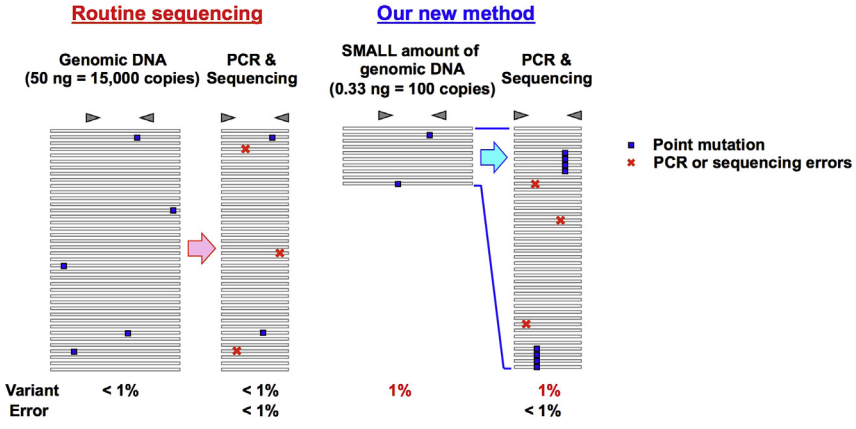

Fig. 1. Schematic presentation of a novel method to measure rare base substitution mutations. In a routine sequencing procedure (left), a sufficient amount of genomic DNA, for example 50 ng (15,000 copies), is used as a template. A rare mutation in the template DNA is expected to be detected with a low allele frequency, for example at 0.007% (1/15,000). However, such a variant with a low allele frequency cannot be distinguished from a sequencing error. In contrast, in our new method (right), 100 copies of genomic DNA are used as a template (for simplicity purposes, 10 molecules are illustrated here). When sequenced, a mutation, if present, is expected to be detected with an allele frequency of 1% (1/100), and can be distinguished from sequencing errors.