Mutation Research 776 (2015) 136–143

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

journal homepage: www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Review

High-throughput sequencing in mutation detection: A new generation of genotoxicity tests?

Alexander Y. Maslov*, Wilber Quispe-Tintaya, Tatyana Gorbacheva[1], Ryan R. White, Jan Vijg**

Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Ave., Bronx, NY 10461, USA

ARTICLE INFO

Article history:
Received 30 August 2014
Received in revised form 7 March 2015
Accepted 27 March 2015
Available online 20 April 2015

Keywords:
Genetic toxicology
Next generation sequencing
Mutagenicity
Mutation
Genome rearrangement

ABSTRACT

The advent of next generation sequencing (NGS) technology has provided the means to directly analyze the genetic material in primary cells or tissues of any species in a high throughput manner for mutagenic effects of potential genotoxic agents. In principle, direct, genome-wide sequencing of human primary cells and/or tissue biopsies would open up opportunities to identify individuals possibly exposed to mutagenic agents, thereby replacing current risk assessment procedures based on surrogate markers and extrapolations from animal studies. NGS-based tests can also precisely characterize the mutation spectra induced by genotoxic agents, improving our knowledge of their mechanism of action. Thus far, NGS has not been widely employed in genetic toxicology due to the difficulties in measuring low-abundant somatic mutations. Here, we review different strategies to employ NGS for the detection of somatic mutations in a cost-effective manner and discuss the potential applicability of these methods in testing the mutagenicity of genotoxic agents.

Contents

* Corresponding author. Tel.: +1 718 678 1153.
** Corresponding author. Tel.: +1 718 678 1152.
E-mail addresses: alex.maslov@einstein.yu.edu (A.Y. Maslov), Jan.Vijg@einstein.yu.edu (J. Vijg).
[1] Present address: Department of Genetics, Cytology, and Bioengineering Voronezh State University, Voronezh, Russia.

http://dx.doi.org/10.1016/j.mrfmmm.2015.03.014
0027-5107/

1. Introduction

Evaluating the hazardous effects of chemicals, such as pharmaceutical, environmental, and industrial compounds, or other agents, such as ionizing radiation, on human health is among the most important problems facing humankind in the modern world. Human contact with these toxic agents is growing exponentially, and even low-level exposures to environmental toxins/pollutants pose serious long-term health risks.

The genome is considered the central governor of all cellular processes and any interference that affects genome integrity may lead to serious health consequences [1]. As such, DNA lesions caused by genotoxic agents may have two different outcomes, i.e., cell death, either actual (apoptosis) or functional (senescence), and acquisition of mutations, due to erroneous DNA replication or repair. The second outcome is arguably more important. The induced mutations, e.g., base-pair substitutions, small insertions and deletions (indels), genome rearrangements and chromosomal events, such as numerical chromosome changes, are generally considered to be a cause of many congenital diseases [2] and the multi-step process of malignant transformation [3]. Also the process of aging has been considered to be ultimately caused by the accumulation of mutations [4,5]. Thus, an assessment of the somatic mutation frequency in cells after treatment with potentially genotoxic agents or in biopsied tissues of individuals potentially exposed to such agents is a critical step in hazard evaluation (Fig. 1).

Historically, short-term tests (STTs) for genotoxic chemicals were established and validated decades ago. STTs include the Ames bacterial mutagenesis test [6], in vitro cytogenetics tests [7,8], and the in vitro and in vivo micronucleus assays [9,10]. More recently, transgenic animal models have been generated that enable testing for spontaneous or induced mutations in any target organ or tissue using reporter genes introduced into various loci of animal genomes [11–14]. However, these tests are indirect and do not provide information on the sequence integrity of the entire genome. Indeed, the field of genetic toxicology has always been based on surrogate markers and has never been able to assess human health risks based on systematic analysis of the entire genome in primary human cells or tissues. Now that the next-generation sequencing (NGS) era is well underway, new methods have been developed to directly analyze genetic material in a genome-wide manner with single nucleotide resolution. Moreover, there is no dependency on any particular gene or cell line and genetic material derived from any cell or tissue can be analyzed. This makes NGS-based mutagenicity assays particularly suitable for use in genetic toxicology. However, there are some serious obstacles that have thus far essentially constrained the application of NGS in genotoxicity testing.

Here, we discuss problems and pitfalls in the implementation of NGS in genetic toxicology. We will first explain why the application of NGS in measuring low-abundant somatic mutations is not straightforward, then describe how this obstacle can be overcome, albeit at high cost, by taking a single cell approach and, finally, review various NGS approaches for assessing mutations, both point mutations and genome structural variations, in small amounts of DNA at low cost.

2. Direct mutation assessment by next generation sequencing

Unlike conventional Sanger sequencing [15], next-generation sequencing is capable of processing hundreds of millions of DNA fragments in parallel, providing the previously unprecedented opportunity to decode the entire genome within days. Due to the relatively simple nature of genetic material, all possible mutations are, in principle, amenable to detection by direct sequencing. However, this is only true for mutations that are present in most or all cells in a given tissue or populations. Indeed, in genetic toxicology the mutations one wishes to detect are typically random, de novo mutations, turning the cell population under study into a mixture of genomes. In such genome mosaics each cell harbors hundreds if not thousands of unique, de novo mutations.

In principle, cellular heterogeneity in genome sequence integrity can be addressed by NGS in a straightforward way by sequencing at great depth. Sequence variants, even at very low abundance, should then be identifiable among the sequence reads at each locus. However, the reliable identification of mutations in this way is constrained by errors associated with each step of the NGS workflow (Fig. 2). Detection of different types of mutation, i.e., point mutations (base substitutions and small indels) and large structural variation (translocations, inversions, large insertions and deletions) is affected in different ways by these errors, which is why we will discuss each mutation type separately.

2.1. Assessment of point mutations and small indels

In principle, somatic point mutations and small indels that occur at low frequencies, i.e., down to 1×10^{-6} per locus, can be detected easily enough by sequencing the entire genome or part of it. However, straightforward detection of somatic mutations as variant reads after sequencing at great depth is essentially precluded by sequencing errors and artifacts introduced during library preparation (Fig. 2). For example, errors may result from base misincorporation during PCR amplification, which is often part of the library preparation protocol. PCR errors stem from less than absolute fidelity of polymerase. If they occur during the first round of amplification (the worst case scenario) they will be propagated and inherited by 50% of the daughter molecules of the starting template [16]. PCR errors may be exacerbated by the presence of damaged bases in the template molecule, which may readily lead to mis-incorporation of bases in the nascent strand, e.g., G → T mutations at 8-oxo-G lesions, which favor insertion of adenosine [17] or C → T at deaminated cytosines [18]. Sequencing errors, i.e., erroneous base calls, missed bases, or homopolymer-length errors, occur during sequencing. These types of errors are usually randomly distributed along the reads and will differ between the two strands (it's highly unlikely that the same error will occur when sequencing the opposing strands). The frequency of sequencing errors for contemporary platforms such as Illumina and Ion Torrent is estimated at 0.1–0.7 insertions/deletions and substitutions per 100 nucleotides of sequencing data [19]. Combined, these sources of error could result in an artifactual mutation frequency of up to 1% [20,21] efficiently masking true mutations, which usually occur at a much lower frequency.

To address the issue of errors, all variant-calling algorithms utilize a consensus model. That is each analyzed region of the genome must be represented by several independent sequencing reads, i.e., independently sequenced fragments representing the same loci but originating from genomes of different cells. Randomly occurring errors are filtered out, while true mutations can be identified based on their presence in 50% of the reads (a hetereozygous mutation affects only one allele). This strategy works well if the same mutations are present in all cells, e.g., germline mutations or clonally amplified mutations in tumor tissue. However, ultra-low-abundant somatic mutations, often unique for each cell are discarded because like sequencing errors they are present in one read [22,23] (Fig. 3). To address these issues several approaches have been developed with the core idea to identify and verify the