European Journal of Cancer 61 (2016) 61–68

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Review

Oncogenic viruses: Lessons learned using next-generation sequencing technologies

Ronan Flippot [a], Gabriel G. Malouf [a,*], Xiaoping Su [b], David Khayat [a], Jean-Philippe Spano [a]

[a] *University Hospital Pitié Salpêtrière, Department of Medical Oncology, University Pierre and Marie Curie, Paris, France*
[b] *Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, U.S.A*

Received 29 November 2015; received in revised form 25 March 2016; accepted 30 March 2016

KEYWORDS
Cancer;
Integration;
Next-generation sequencing;
Virus;
NGS

Abstract Fifteen percent of cancers are driven by oncogenic human viruses. Four of those viruses, hepatitis B virus, human papillomavirus, Merkel cell polyomavirus, and human T-cell lymphotropic virus, integrate the host genome. Viral oncogenesis is the result of epigenetic and genetic alterations that happen during viral integration. So far, little data have been available regarding integration mechanisms and modifications in the host genome. However, the emergence of high-throughput sequencing and bioinformatic tools enables researchers to establish the landscape of genomic alterations and predict the events that follow viral integration. Cooperative working groups are currently investigating these factors in large data sets. Herein, we provide novel insights into the initiating events of cancer onset during infection with integrative viruses. Although much remains to be discovered, many improvements are expected from the clinical point of view, from better prognosis classifications to better therapeutic strategies.

1. Introduction

Cancers are the leading cause of mortality in developed western countries. In a variety of situations, pathogens like bacteria, viruses and parasites are suspected to be the key event for cancer development. Cancer related bacteria are mostly represented by *Helicobacter pylori* in gastric carcinoma [1], or *Fusobacterium* in colorectal

* *Corresponding author*: Hôpital Salpêtrière, département d'oncologie médicale, 47-83 Boulevard de l'hôpital, 75013, Paris, France.
E-mail address: gabriel.malouf@aphp.fr (G.G. Malouf).

http://dx.doi.org/10.1016/j.ejca.2016.03.086
0959-8049/

cancer [2]. Parasites such as *Fasciola hepatica* or *Schistosoma haematobium* are responsible for cholangiocarcinomas [3] and bladder cancer [4], respectively. Seven viruses have been reported to be associated with human cancers [5]. Of those, hepatitis C virus (HCV), Epstein-Barr virus (EBV) and human herpes virus 8 (HHV8) are non-integrative viruses, while hepatitis B virus (HBV), human papillomavirus (HPV), Merkel cell polyomavirus (MCPV), and human T-cell lymphotropic virus (HTLV) integrate the host genome.

The causal association between viruses and cancers has been established in about 10–15% of malignant tumours worldwide [6]. Mechanistically, most virus-associated tumours are thought to follow a multistep carcinogenesis model, with a continuum between dysplasia and invasive cancer [7]. Cell transformation mechanisms are diverse [8], ranging from induced expression of viral and cellular oncogenes [5], to epigenetic modifications [9], miRNA regulations [9] and pro-tumoural inflammation [10].

The incidence of virus-associated tumours has long been underestimated, as next-generation sequencing (NGS) of large tumour data sets was not available. Furthermore, the oncogenic mechanisms have been poorly understood, but we have observed that tumours caused by integrative viruses remain unique in their natural history, prognosis, and response to treatment [6]. Thus, exploring viral integration patterns and host–pathogen interactions is crucial for the understanding and management of virus-associated tumours.

The development of high-throughput DNA and RNA sequencing, in conjunction with new bioinformatic tools, allows for deep, whole-genome sequencing of the host and pathogen. Recent improvements in genomics and transcriptomics have made it possible to precisely identify viral sequences in tumour genomes, which has led to the discovery of new tumour-associated viruses [11]. The identification of redundant integration sites and recurrent changes in gene expression is essential to the comprehension of this type of oncogenesis. Further extending these techniques to large-scale studies can establish the genomic landscapes of virus and tumour associations with great precision.

This article aims to provide an updated perspective on NGS techniques and software for the study of integrative tumour-associated viruses. We describe their contribution to the understanding of the oncogenic mechanisms that occur during viral integration and the clinical relevance of such developments.

2. Impact of NGS and bioinformatics

In the last decade, new techniques have led to a drastic improvement in the comprehension of genetics and epigenetics, all of which can be applied to tumour-associated viruses. Foremost, NGS has provided a giant leap towards whole-genome sequencing and the comprehensive characterization of modifications in gene expression.

2.1. Charting genomic and epigenomic architecture

The use of RNA microarrays to study gene expression became possible in the 1990s [12,13]. Probe-based methods use fluorescent markers of RNA hybridization to homologous cDNA to compare the differences in gene expression between a study genome and a reference genome, in which the relative changes in gene expression are determined by the change in fluorescence. This method was the first to directly compare the genomes of the host and pathogen *in vitro* and *in vivo*. However, it requires the physical separation of both host and pathogen prior to processing, which can alter the evidence of host–pathogen interactions in gene expression changes. In addition, it does not cover non-coding RNAs nor study mRNA splicing variants [14].

NGS is a powerful tool for studying the changes in gene expression between cell lineages from different tissues or in different environments. Indeed, RNA sequencing allows for the study of the whole transcriptome through standardized techniques. Nucleic acids are extracted from fresh frozen or formalin-fixed paraffin-embedded tissues, and nucleic acid sequences are fragmented into about 200-nucleotide inserts for sequencing. The most common among the diverse sequencing methods is synthesis sequencing, in which deoxynucleotide triphosphates labelled with fluorescent markers are incorporated into the nucleic acid chain [15]. Other base detection techniques have appeared in the last decade, such as pH-based methods or transmembrane channels that transport sequences with specific templates [16,17]. The gene expression level is correlated with the number of reads per exon, normalized by the length of said exon. Thanks to deep and parallel sequencing, these NGS techniques can cover virtually every transcription variant. Such advancements have made it easier and more affordable to sequence virus and cellular lineages of interest such as tumour samples.

These advances in sequencing also benefit the study of epigenetics and chromatin remodelling processes. ChiP-sequencing combines chromatin immunoprecipitation and NGS to define the state of the chromatin and the level of gene expression: fragmented DNA associated with a transcription factor or a specific histone variant is immunoprecipitated and sequenced [18]. Using antibodies directed against markers of repression, such as H3K27, or active transcription, such as H3K4me3, it is possible to identify repressed genes or transcriptionally active genes in a tumour sample [19].