C. R. Biologies xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Comptes Rendus Biologies

www.sciencedirect.com

Trajectories of genetics, 150 years after Mendel/Trajectoires de la génétique, 150 ans après Mendel

Using large-scale genome variation cohorts to decipher the molecular mechanism of cancer

Étude à grande échelle de variations génétiques pour déchiffrer les mécanismes moléculaires de l'oncogenèse

Nina Habermann [a], Balca R. Mardin [a], Sergei Yakneen [a], Jan O. Korbel [a,b,*]

[a] *European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany*
[b] *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, CB10 1SD Cambridge, UK*

ARTICLE INFO

Article history:
Received 5 April 2016
Accepted after revision 30 April 2016
Available online xxx

Keywords:
Genetic variation
Chromothripsis
Cancer genomics
Structural variation
Big data
Hybrid cloud

ABSTRACT

Characterizing genomic structural variations (SVs) in the human genome remains challenging, and there is a growing interest to understand somatic SVs occurring in cancer, a disease of the genome. A havoc-causing SV process known as chromothripsis scars the genome when localized chromosome shattering and repair occur in a one-off catastrophe. Recent efforts led to the development of a set of conceptual criteria for the inference of chromothripsis events in cancer genomes and to the development of experimental model systems for studying this striking DNA alteration process *in vitro*. We discuss these approaches, and additionally touch upon current "Big Data" efforts that employ hybrid cloud computing to enable studies of numerous cancer genomes in an effort to search for commonalities and differences in molecular DNA alteration processes in cancer.

Mots clés :
Variation génétique
Chromothripsie
Génomique du cancer
Variation structurale
Mégadonnées
Informatique en nuages hybrides

RÉSUMÉ

Caractériser les variations structurelles génomiques (SVs) dans le génome humain reste difficile et comprendre ces variations somatiques survenant dans le cancer, une maladie du génome, fait l'objet d'un intérêt croissant. Un processus de SV ravageur, connu sous le nom de chromothripsie, endommage le génome par un phénomène de fragmentation-réparation localisé en une seule étape, qui devient soudainement catastrophique. Des efforts récents ont conduit à l'élaboration d'un ensemble de critères conceptuels pour révéler des événements de chromothripsie dans les cancers et pour développer des systèmes modèles expérimentaux pour étudier in vitro ce processus d'altération de l'ADN. Nous discuterons ces approches et de plus, nous évoquerons les efforts actuels qui, mettant en œuvre des ressources informatiques à

* Corresponding author. European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany.
E-mail address: korbel@embl.de (J.O. Korbel).

http://dx.doi.org/10.1016/j.crvi.2016.05.008
1631-0691/

l'échelle mondiale de type « Big Data » sur de nombreux génomes, recherchent des points communs et des différences dans les processus d'altération moléculaire de l'ADN dans le cancer.

1. Introduction

Genetic variation is a fundamental reason why humans differ from one another and why some individuals are more susceptible to diseases such as cancer than others, and there is a growing interest in investigating the mechanisms and the phenotypic consequences of genetic variation. A sharp decrease in the costs of DNA sequencing has enabled the sequencing of numerous human genomes and their mining using "Big Data" analytical approaches for unraveling molecular disease processes [1]. One example for such studies is the presently ongoing Pan-Cancer Analysis of Whole Genomes (PCAWG) project (https://dcc.icgc.org/pcawg), a forerunner project in Big Data analytics of genomes from patients, which aims to integrate data from somatic and germline whole genomes, DNA methylomes, transcriptomes, and clinical data from more than 2800 cancer patients amounting to nearly a Petabyte of sequencing data [1]. The objective of PCAWG is to unravel commonalities and to distinguish factors between cancer types and subtypes at the molecular level, to facilitate the molecular classification of malignancies with impact on diagnostics and treatment, and to uncover causalities linking genotype, environment, and phenotype. The unprecedented resource developed through PCAWG will enable standardized analysis of cancer genomes and associated datasets including transcriptomes, DNA methylomes, and clinical data to obtain insights into molecular disease processes relevant to cancer.

The objective of this paper is to briefly review methodologies for analyzing disease processes, with a specific focus on complex DNA rearrangements, in cancer genomes. Additionally, we will provide an outlook to coming Big Data efforts – with one example being the PCAWG project – which will facilitate the understanding of basic processes as well as translational research (Box 1).

2. Cancer genomes can evolve through catastrophes: the molecular process of chromothripsis

Cancer genome sequencing has enabled new insights into how tumors evolve, and has led to quite remarkable findings relating to the fact that cancer is not merely driven by stepwise alterations but can arise in conjunction with bursts of mutational events [2]. One particularly remarkable example for this is chromothripsis, a molecular process first described by Stephens et al. in 2011 based on cancer genome analysis, which can scar individual chromosome arms or one up to several chromosomes when localized chromosome shattering and repair occurs in a one-off catastrophe [3]. Rearrangement patterns associated with chromothripsis occur in approximately 2–3% of human cancers [3] and *TP53* germline mutations are linked with the occurrence of chromothripsis in pediatric medulloblastoma [4]. Chromothripsis is also abundant in other cancers, such as bladder [5], breast [6], melanoma [7], and in bone cancers [3]. While the prevalence of chromothripsis in diverse cancer cell lines and cancer genomes [3,4,8] suggest a crucial role of chromothripsis in cancer development, the reproducible inference of this process has remained challenging, requiring that cataclysmic one-off rearrangements can be distinguished from localized genetic lesions that occur in a stepwise fashion. We recently devised a set of conceptual criteria for the inference of complex DNA rearrangements suitable for rigorous statistical analyses, which included previously established [3] as well as novel criteria: clustering of breakpoints, regularity of oscillating copy-number states, interspersed loss and retention of heterozygosity, prevalence of rearrangements affecting a specific haplotype, randomness of DNA segment order and fragment joins, as well as the ability to walk the derivative chromosome [8]. These criteria attempt to reject the alternative hypothesis that DNA rearrangements have occurred in a stepwise (progressive) fashion. Further refinement of these criteria allow inferring chromothripsis events in conjunction with additional stepwise patterns of DNA alteration [9], and collectively, these criteria have begun to be used quite regularly to operationally define chromothripsis events based on cancer genome sequencing data.

Box 1. Challenges in genomics due to recent increase in data set sizes.

Opportunities emerging from large datasets are not without challenges. When using a typical university Internet connection, for example, it would take more than 15 months to transfer a data set such as the one amassed by the ICGC from its data repository into a researcher's local IT infrastructure [1]. And the hardware needed to store, let alone process the data, would produce costs at a million dollar level each year. So, consider a PhD student pursuing a 3-year thesis project at a European university. Before she/he could even consider to start pursuing analyses, 15 months of download time would need to be overcome – a considerable portion of the overall time frame anticipated for the thesis, and a million USD in storage would need to be spent before any research work could begin.

Given the establishment of new sequencing platforms that can now sequence over 10,000 human genomes a year at several locations around the world, the number of human genomes generated in the context of disease research is going to be ramped up dramatically in the coming 5 to 10 years. Genomics England Ltd, a government-driven initiative in the UK, for example, plans to sequence 100,000 human genomes – including 25,000 from cancer patients, by the end of 2017.

Please cite this article in press as: N. Habermann, et al., Using large-scale genome variation cohorts to decipher the molecular mechanism of cancer, C. R. Biologies (2016), http://dx.doi.org/10.1016/j.crvi.2016.05.008