Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Collection, integration and analysis of cancer genomic profiles: from data to insight

Jianjiong Gao, Giovanni Ciriello, Chris Sander and Nikolaus Schultz

The recent deluge of cancer genomics data provides a tremendous opportunity for the discovery of detailed mechanisms of tumorigenesis and the development of therapeutics. However, identifying the functionally relevant genomic alterations ('*drivers*') among the many non-oncogenic events ('*passengers*') presents a major challenge. Several new methods have been developed over the past few years that identify recurrently altered genes. Mapping the recurrent genomic alterations, such as somatic mutations and focal DNA copy-number alterations, onto individual tumor samples as tumor-specific event calls facilitates the identification of altered processes and pathways. The resulting reduction in complexity makes cancer genomics data more easily interpretable by cancer researchers and is now driving the development of powerful yet intuitive web-based analysis tools.

Addresses
Computational Biology Center, Memorial Sloan-Kettering Cancer Center, Box 460, New York, NY 10065, USA

Corresponding author: Schultz, Nikolaus (schultz@cbio.mskcc.org)

Current Opinion in Genetics & Development 2014, 24:92–98

This review comes from a themed issue on **Cancer genomics**

Edited by **David J Adams** and **Ultan McDermott**

For a complete overview see the Issue and the Editorial

Available online 27th February 2014

0959-437X/$ – see front matter, © 2014 Elsevier Ltd. All rights reserved.

http://dx.doi.org/10.1016/j.gde.2013.12.003

Introduction

Large-scale cancer genomics projects such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and several efforts led by individual institutions, have recently generated an unprecedented amount of genomic data on tumor samples [1–6,7•,8–22]. While early cancer genomics projects initially focused on array-based mRNA expression and then DNA copy-number data, most projects now employ some form of high-throughput sequencing, for example, all RNA, whole exome, and/or whole genome sequencing. To date, these projects have explored somatic mutations in the coding regions of all genes in more than 15,000 tumors from more than 30 tumor types, and many have also generated detailed maps of DNA copy-number alterations, DNA methylation changes, and mRNA expression changes.

These efforts have led to the discovery of novel cancer genes, such as isocitrate dehydrogenase (IDH1) [23] and polymerase epsilon (POLE) [24••], and they have elucidated the involvement of a number of biological processes and signaling pathways in tumor initiation and progression [25••,26,27,28•]. Many of these pathways tend to be altered in the majority of tumors, but the exact genomic mechanisms of dysregulation differ. As a result, we now have a better understanding of the heterogeneity of alterations within tumor types as well as an appreciation of similarities across tumor types [28•].

However, many challenges remain. As the field is moving away from array-based technologies and Sanger sequencing, new software and algorithms for sequence analysis need to be developed (reviewed in [29,30]). While sequence alignment and mutation calling methods are evolving rapidly, their performance is further complicated by varying degrees of tumor heterogeneity, tumor purity and uneven sequence coverage.

This review covers the current state of downstream data collection, integration and analysis. One of the main challenges is to make these complex data easily accessible and interpretable. Experience from the last few years has shown that this is best achieved by first *distilling* the genomic data to a set of likely functional alteration events (mutations, copy-number changes, methylation events, significant overexpression and underexpression) (Figure 1). These candidate functional events can be mapped onto individual tumor samples. The resulting simplified maps of genomic alterations (event maps) can be more easily used to identify commonly targeted pathways and to identify potential treatment options down to the level of a single patient (Figure 1).

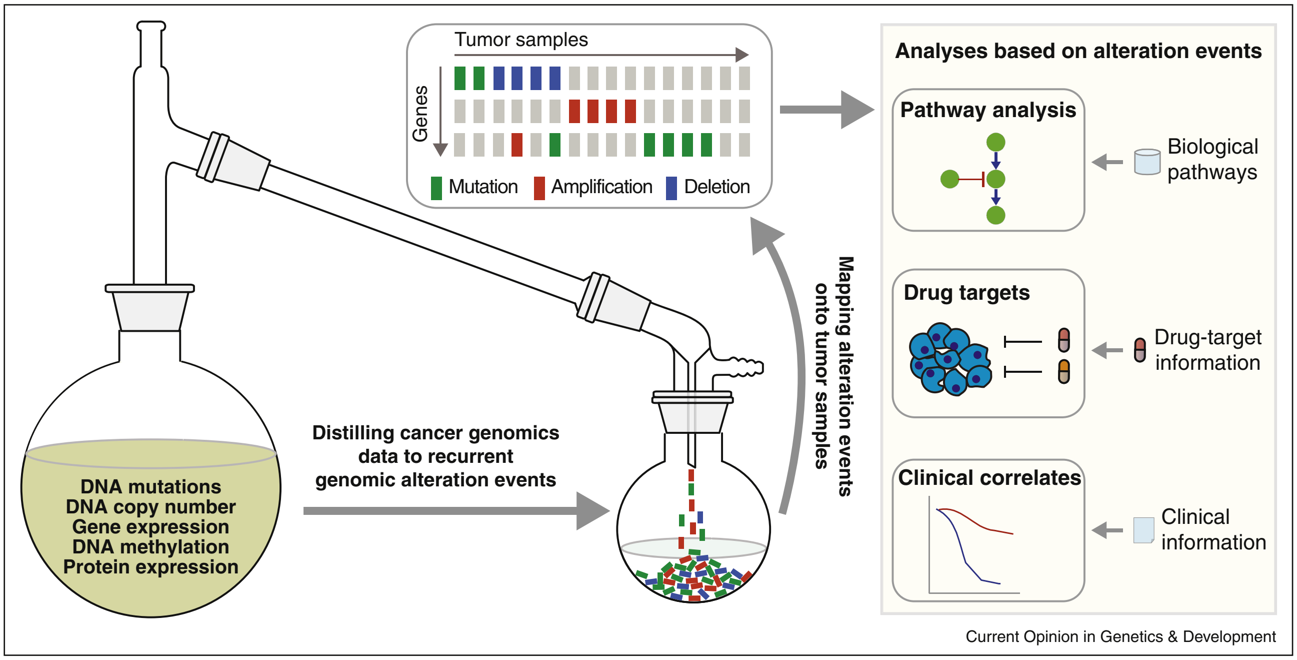
Repositories for cancer genomics data

There are several online databases that host cancer genomics data. For reasons of practicability and access control, normalized gene-level data and raw sequencing data are usually stored in separate repositories. All these data are freely available to the public, but access to raw sequence data requires authorization by the individual projects' data access committees. A full list of all available resources that serve TCGA, the ICGC, the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative, the Cancer Cell Line Encyclopedia (CCLE) and other projects is shown in Table 1.

Detecting recurrent genomic alterations to find cancer drivers

Part of the art in interpreting complex genomic data from tumor samples is to separate the signal from the noise,

Figure 1



Current Opinion in Genetics & Development

Cancer genomics data processing and analysis: From raw data to biological insight. A key step towards extracting biological insights from complex cancer genomic data is the identification of the genomic alterations that contribute to tumorigenesis. The most likely candidates are the (statistically re-ranked) recurrent events, which, when mapped onto individual tumor samples, can be used to identify commonly altered pathways and potentially inform treatment decisions.

that is, identify specific genomic alterations that contribute to the development and growth of a tumor (so-called *drivers*) within a background of a large number of alterations that do not confer a selective advantage for the tumor (*passengers*). Several methods have been developed for the identification of somatic mutations or DNA copy-number alterations that, across a set of tumors, occur at a higher rate than expected by chance (recurrent events).

The methods that identify recurrently mutated genes typically take into account factors such as the number and types of mutations in a gene, the length of the gene, the background mutation rate of a tumor and gene, DNA sequence conservation and recurrence at specific positions (hotspots). The most commonly used methods are MutSig [31••], MUSIC [32], and InVex [33]. More recently, the functional impact of mutations, as predicted by tools such as SIFT [34], PolyPhen-2 [35], and MutationAssessor [36], has also been considered (OncodriveFM [37]) as well as the clustering of mutations along the protein sequence of a gene (MUSIC [32] and OncodriveCLUST [38]). However, since these methods rely on recurrence, they cannot identify rare driver mutations. Some of these mutations may be common in certain cancer types, but others may be so rare that they cannot be detected by even the most sophisticated recurrence-based methods.

Recurrence-based methods have also been developed to identify genes that are altered by copy-number changes, for example GISTIC2.0 [39] and RAE [40]. These methods include amplitude and focality. Many of the recurrently altered regions (referred to as Regions of Interest, ROIs) contain no known oncogenes or tumor suppressors [41•], and most contain multiple genes. Correlation with mRNA expression can be used to exclude from downstream analyses the genes that are not expressed or not sensitive to changes in DNA copy number. The impact of copy number changes on expression has been considered for driver genes in Oncodrive-CIS [42].

Similar methods can be applied to DNA methylation data to identify recurrently silenced genes, especially when coupled to mRNA expression data. Outlier expression analysis has been successfully applied to identify ETS family gene fusions in prostate cancer [43], but these methods are not yet commonly used to identify genes with unusual (e.g., bimodal) expression patterns in cancer genomics data sets. Expression data from RNA sequencing now make it possible to detect fusion genes. Several software tools for fusion detection exist, like DeFuse [44], FusionSeq [45], or BreakFusion [46], but no consensus method has emerged yet. The role of aberrant splicing events can also be explored, for example by using JuncBASE [47].

Download English Version:

https://daneshyari.com/en/article/5893630

Download Persian Version:

https://daneshyari.com/article/5893630

Daneshyari.com