Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Big data in cancer genomics

Ana-Teresa Maia[1,2,3], Stephen-John Sammut[4,5], Ana Jacinta-Fernandes[1,2,3] and Suet-Feung Chin[4,5]

Abstract

Advances in genomic technologies in the last decade have revolutionised the field of medicine, especially in cancer, by producing a large amount of genetic information, often referred to as Big Data. The identification of genetic predisposition changes, prognostic signatures, and cancer driver genes, which when mutated can act as genetic biomarkers for both targeted treatments and disease monitoring, has greatly advanced our understanding of cancer. However, there are still many challenges, such as more sophisticated analysis tools and higher processing capacity, along with cheaper storage, faster and more efficient data transfer, that must be overcome before personalised medicine finally becomes a reality.

Addresses

[1] Department of Biomedical Sciences and Medicine (DCBM), University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
[2] Centre for Biomedical Research (CBMR), University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
[3] Algarve Biomedical Center (ABC), University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal
[4] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK
[5] Department of Oncology, University of Cambridge, Cambridge CB2 2QQ, UK

Corresponding author: Suet-Feung Chin (suet-feung.chin@cruk.cam.ac.uk)

Current Opinion in Systems Biology 2017, **4**:78–84

This review comes from a themed issue on **Big data acquisition and analysis (2017)**

Edited by **Pascal Falter-Braun and Michael A. Calderwood**

For a complete overview see the Issue and the Editorial

Available online 24 July 2017

http://dx.doi.org/10.1016/j.coisb.2017.07.007

Introduction

Cancer is a genetic disease. Finding cancer-causing genetic events was painstakingly slow until the introduction of Sanger sequencing [1], comparative genomic hybridisation [2], microarray and more recently, next generation sequencing (NGS) technologies (reviewed in Ref. [3]). These technological advances increased our knowledge of the human genome and its role in diseases, shifting from "single gene-single disease" with limited sample numbers, to millions of data points collected from up to thousands of samples. This led to not only the publication of the human genome by the Human Genome Project (HGP) [4,5], but also discoveries of copy number events and transcriptomic signatures that have refined tissue-specific signatures, aiding diagnosis, prognosis and therapeutic decisions.

The reference genome took 13 years and US$3 billion to produce. It highlighted the need for more sophisticated sequencing technologies, resulting in several NGS or 2nd generation sequencing technologies allowing for rapid large-scale genomic sequencing (reviewed in Ref. [3]). The US$1000 genome challenge, introduced by NHGRI back in 2004, has now been met as Veritas Genomics recently advertised complete genome sequence for a single individual with interpretation for US$999.

In this review, we will consider the explosion of data generated from the advent of both microarrays and the NGS era as "Big Data". We will discuss the importance of "Big Data" in refining our understanding of cancer biology, its potential and challenges.

Types of data in cancer genomics

DNA and RNA are the two most utilised biological materials in cancer genomics. DNA sequencing includes whole genome sequencing (WGS) or targeted sequencing (TS), where PCR primers or oligonucleotides are designed for enrichment of specific genomic regions, such as whole exome (WES) or smaller custom panels. WGS is the most comprehensive and can identify all types of genomic alterations, but it is time consuming and cost prohibitive, whereas enrichment methods detect predominantly single nucleotide alterations. RNA sequencing reveals not only gene expression, but also alternative splicing events, gene fusions, post-transcriptional modifications and allelic expressions. Amongst the less studied are the proteome, non-coding transcriptome and epigenome, although these fields are rapidly gaining interest.

Big data achievements

For almost a decade, microarrays were indispensable tools for the understanding of cancer genomics. Expression arrays were pivotal in providing an insight into the biology of cancer. Indeed, the use of expression arrays radically altered our classification of breast cancer, which had been previously thought to be one disease. Five subtypes emerged, with very different clinical prognoses [6]. By layering on copy number alterations

from copy number arrays, breast cancer was further refined into 10 integrative clusters (IntClust), identifying, for the first time, ER positive tumours with poor survival [7]. However, microarray data is limited to gross expression, copy number changes, and single nucleotide polymorphism (SNP) identification. It is not able to provide base pair level information across the whole genome. On the other hand, NGS technologies could provide this, and a paradigm shift occurred, with preferential usage of NGS rather than microarrays. International consortia such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) have spearheaded sequencing efforts into common cancer types. TCGA have sequenced 12 types of cancer (https://cancergenome.nih.gov/) and ICGC have sequenced 21 tissue types (http://icgc.org/).

For the very first time, whole genome aberrations (copy number aberrations, single nucleotide variations, small insertions-deletions) could be identified for each patient in a single experiment making personalised medicine a reality. Common alterations per tumour type were identified, as well as subtle differences within a single tumour subtype. For instance, three separate NGS studies on lobular breast cancers (ILC) identified only *CDH1* and *PIK3CA* as frequently mutated, regardless of where the studies were conducted, suggesting that these are *bona fide* cancer driver genes in ILC [8–10]. More importantly, meta-analyses of combined data from multiple sites can increase sample numbers, especially for rare disease subtypes.

Surprisingly NGS data revealed that very few genes were mutated at high frequencies across different cancer types, including breast cancer [8,11–13], suggesting that there is no one gene that drives carcinogenesis in all cancer types. Hence, few novel drugs targeting single gene alterations would result from current sequencing efforts. Instead, existing drugs might work better if we know the patients' genetic code. For example, PI3K inhibitors have not shown much efficacy in breast cancer patients with *PIK3CA* mutations [14]. Whilst *PIK3CA* mutations do not confer worse prognosis in all ER positive (+) breast cancer patients, they do in specific ER+ IntClusts, suggesting that PI3K inhibitors might be more beneficial for these women [12]. In addition, some genes can function as tumour suppressors (TSG) or oncogenes in different types of tumours, such as *NOTCH1* [15], or only in specific contexts, e.g. *SMAD4* is a putative TSG in ER+ breast cancers only [12]. Identifying genetic biomarkers for each patient will allow residual disease and the efficacy of each treatment cycle to be monitored using non-invasive cell-free DNA in plasma [16].

The availability of data from different tumour types has allowed for the first time an innovative meta-analysis combining all "omics" data in a "pan-cancer" effort to study the similarities and disparities in the genomic changes. TCGA combined data across 12–14 cancer types and found that different tumour types could be classified into copy number or mutation driven groups. *APOBEC3B* was shown to be associated with widespread mutagenesis in many cancer types [17,18]. Aggregating rare mutations occurring in all cancers allowed some to be implicated as drivers when integrated into known biological pathways hence, identifying novel drug targets that would never have been significant in a single cancer analysis [19,20]. Similarly, ICGC did a "pan-cancer" analysis across 40 cancer types using data from 10,952 exomes and 1048 whole-genomes (http://cancer.sanger.ac.uk/cosmic/signatures), and found that mutations occur in specific patterns that represent the different biological processes the cell has undergone, such as smoking, ageing and DNA damage [21,22].

With deep sequencing, variant frequencies can be quantified even for rare alleles. This has led to the development of statistical methods that seek to measure clonal heterogeneity within the tissue, or intra-tumour heterogeneity (ITH) [23–25], and has expanded our knowledge of cancer evolution by identifying clonal and subclonal mutations [26,27]. Investigating ITH has led studies into the temporal dynamics of cancer growth, but it is still inconclusive if it is a gradual accumulation or punctuated formation of driver mutations (reviewed in Ref. [28]).

Comprehensive mutation documentation and advances in epitope predicting algorithms have aided in the identification of patient specific neoantigens [29–31]. These cancer neoantigens could make personalised immunotherapy a possibility in the not too distant future.

The era of big data has also transformed the field of cancer susceptibility, not only in terms of larger acquisition capability, but data handling and analysis. The development of better genotyping technologies and the establishment of international consortia have produced over 100 new genome-wide association studies (GWAS) in the last two years alone, yielding over 1200 associations for a range of cancer types (The NHGRI-EBI Catalog of published GWAS is available at: www.ebi.ac.uk/gwas. Accessed February 7th 2017) [32].

However, NGS has not been a major contributor to the discovery of new risk loci, mostly because of costs involved in analysing the necessary number of individuals. Nevertheless, there are already some examples of its application in association studies, as well as follow-up fine mapping studies [33–35]. Additionally, GWAS consortia have taken associations studies to the tens of thousands of subjects, and with that have allowed a much more detailed dissection of cancer aetiology, such as identifying cancer sub-type specific risk loci, and cross-cancer analysis risk loci with