The International Journal of Biochemistry & Cell Biology xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

The International Journal of Biochemistry & Cell Biology

journal homepage: www.elsevier.com/locate/biocel

Medicine in focus

Dissecting cancer heterogeneity – An unsupervised classification approach

Q1 Xin Wang [a,*], Florian Markowetz [a], Felipe De Sousa E Melo [b], Jan Paul Medema [b], Louis Vermeulen [a,b]

[a] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK
[b] Laboratory for Experimental Oncology and Radiobiology, Center for Experimental Molecular Medicine, Academic Medical Center, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 28 July 2013
Received in revised form 20 August 2013
Accepted 22 August 2013
Available online xxx

Keywords:
Cancer subtypes
Consensus clustering
Gene expression
Stratified medicine
Personalized medicine

ABSTRACT

Gene-expression-based classification studies have changed the way cancer is traditionally perceived. It is becoming increasingly clear that many cancer types are in fact not single diseases but rather consist of multiple molecular distinct subtypes. In this review, we discuss unsupervised classification studies of common malignancies during the recent years. We found that the bioinformatic workflow of many of these studies follows a common main stream, although different statistical tools may be preferred from case to case. Here we summarize the employed methods, with a special focus on consensus clustering and classification. For each critical step of the bioinformatic analysis, we explain the biological relevance and implications of the technical principles. We think that a better understanding of these ever more frequently used methods to study cancer heterogeneity by the biomedical community is relevant as these type of studies will have an important impact on patient stratification and cancer subtype-specific drug development in the future.

1. Introduction

Cancers, even within the same organ, can no longer be understood as single disease entities. In many cancers heterogeneity exists both within the same primary tumour (intra-tumour) and across individual patients (inter-tumour) of the same histopathological type (De Sousa E Melo et al., 2013a). The complexities resulting from both intra- and inter-tumour heterogeneity impinge on the selection of patients that may benefit from adjuvant therapy. Understanding the biological properties distinguishing tumours is key to more individualized therapy and targeted drug design. An efficient strategy to dissect inter-cancer heterogeneity is to classify tumours into molecular subgroups. The relevance of identifying cancer subtypes is evident as it improves the biological understanding of the disease and is an important step towards stratified medicine.

Traditionally, tumours are categorized according to histopathological features, tumour size, grade and disease stage. Although these classification methods contributed to a reduction of cancer mortality, they provide limited understanding of underlying biology and are in many cases not sufficient for decision-making about adjuvant therapy. In addition, molecular classifications have employed a handful of "classic" molecular biomarkers such as mutations in individual genes or chromosomal aberrations, epigenetic markers, or protein expression. These methods have some known pitfalls and for example similar mutations may lead to highly various cancer phenotypes: the same BRAF mutation (V600E) occurs both in colorectal cancers (CRCs) and melanomas, however, BRAF inhibitors only have a demonstrated benefit in melanoma patients harbouring this mutation and are inactive in CRC indicating a context dependent effect of this mutation (Prahallad et al., 2012). Conversely, different mutations may lead to very similar biological behaviour. For instance, Gastro Intestinal Stromal Tumours (GIST) resulting from activating mutations in the receptor tyrosine kinases c-KIT or PDGFRA display very similar biological and clinical properties (Corless et al., 2011). Consequently, these low-dimensional features, which are used in current classifications, only partially capture the biomolecular patterns of cancer subtypes.

Here we review unsupervised classification approaches to identify cancer subtypes. This approach incorporates expression profiles at the whole transcriptome level and therefore draws a more comprehensive landscape of the intrinsic molecular characteristics across distinct subgroups. This is different from supervised classification methods, which incorporate a priori variables such as therapy response or survival information. Unsupervised classification

Q2 * Corresponding author at: Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. Tel.: +44 7534364129.
E-mail addresses: xin.wang@cruk.cam.ac.uk (X. Wang), louis.vermeulen@amc.uva.nl, Louis.Vermeulen@cruk.cam.ac.uk (L. Vermeulen).

1357-2725/$ – see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.biocel.2013.08.014

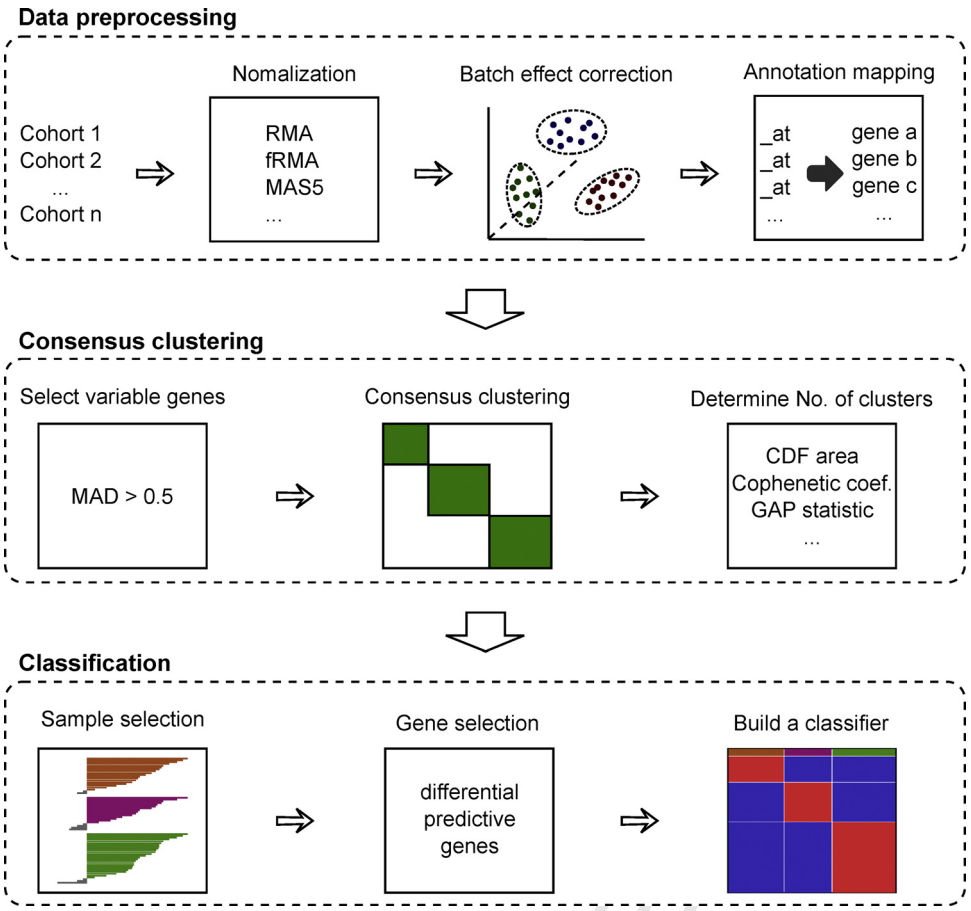

Fig. 1. A typical workflow to identify and validate cancer subtypes. The workflow involves three main steps: data pre-processing, consensus clustering (or NMF) and classification, that each consistent of multiple steps.

strategies have been successfully employed to identify subtypes in a vast number of malignancies. From these studies, we distilled a common bioinformatic workflow including the main steps of data curation and pre-processing, consensus clustering, classification, subtype characterization and validation (Fig. 1). In particular we will focus on the biological relevance of some key statistical analyses involved in these studies. We believe a basic understanding of these methods is crucial for a critical evaluation of classification studies involving gene expression data, which will be of increasing relevance in the future for both biologists and clinicians. The bioinformatic workflow we describe in this review is also the basis of more advanced methodologies that integrate multi-level genomic data (methylation, copy number, gene expression etc.).

2. Gene expression-based unsupervised classification studies

2.1. Discovery patient cohort

The first step in identification of cancer subtypes is a well-characterized patient cohort including histopathological data and clinical outcome. The size of the cohort significantly influences the outcome and rare subtypes are likely not to be picked up in smaller patient series. Most recent subtype discovery studies employ discovery sets ranging from 100 to 500 patients, some of which are combinations of newly generated gene expression profiles with publicly available datasets. The latters are either derived from individual academic projects or cancer consortia such as the Cancer Genome Consortium (TCGA, http://cancergenome.nih.gov/). Additional characteristics of the discovery set will also influence the outcome: when a clinically homogenous set is used, for example one particular disease stage, the observed heterogeneity is more likely to reflect biological variation. On the other hand cancer subtypes not frequently found within that specific patient selection will not be identified.

2.2. Data generation, curation and pre-processing

Microarrays are the most commonly employed technique to generate a comprehensive overview of the transcriptome of individual malignancies. First, raw microarray data of an individual cohort are background corrected, normalized and transformed from a linear to a logarithmic scale, using standard tools such as RMA, MAS5 or fRMA (McCall et al., 2010). This step adjusts the individual hybridization intensities to ensure a meaningful comparison across samples. Integrating normalized gene expression profiles of different cohorts can be challenging, as they may be collected from different hospitals, and generated using different platforms and conditions. These systematic non-biological differences can confound true biological differences and are often insufficiently corrected using the classic normalization tools mentioned above (Luo et al., 2010). Crucial at this stage is to merge multiple cohorts such that non-biological batch effects are minimized while biological variations are retained. This can be achieved by using simple linear models, factor analysis (Verhaak et al., 2010), ComBat (Johnson et al., 2007), SVA (Leek and Storey, 2007) or distance-weighted discrimination approaches (Benito et al., 2004), that each have their unique merits. For instance, ComBat is

Please cite this article in press as: Wang X, et al. Dissecting cancer heterogeneity – An unsupervised classification approach. Int J Biochem Cell Biol (2013), http://dx.doi.org/10.1016/j.biocel.2013.08.014