EBioMedicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response

Magali Champion [a,1], Kevin Brennan [a,1], Tom Croonenborghs [b,c], Andrew J. Gentles [d], Nathalie Pochet [b], Olivier Gevaert [a,*]

[a] Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine & Biomedical Data Science, Stanford University, United States
[b] Program in Translational Neuropsychiatric Genomics, Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Broad Institute of Harvard and Massachusetts Institute of Technology, United States
[c] Advanced Integrated Sensing Lab, Campus Geel, Department of Computer Science, University of Leuven, Belgium
[d] Department of Medicine, Center for Cancer Systems Biology, Stanford University, United States

ARTICLE INFO

Article history:
Received 9 August 2017
Received in revised form 23 November 2017
Accepted 29 November 2017
Available online xxxx

Keywords:
Data fusion
Cancer driver gene discovery
Module network

ABSTRACT

The availability of increasing volumes of multi-omics profiles across many cancers promises to improve our understanding of the regulatory mechanisms underlying cancer. The main challenge is to integrate these multiple levels of omics profiles and especially to analyze them across many cancers. Here we present AMARETTO, an algorithm that addresses both challenges in three steps. First, AMARETTO identifies potential cancer driver genes through integration of copy number, DNA methylation and gene expression data. Then AMARETTO connects these driver genes with co-expressed target genes that they control, defined as regulatory modules. Thirdly, we connect AMARETTO modules identified from different cancer sites into a pancancer network to identify cancer driver genes. Here we applied AMARETTO in a pancancer study comprising eleven cancer sites and confirmed that AMARETTO captures hallmarks of cancer. We also demonstrated that AMARETTO enables the identification of novel pancancer driver genes. In particular, our analysis led to the identification of pancancer driver genes of smoking-induced cancers and 'antiviral' interferon-modulated innate immune response.

Software availability: AMARETTO is available as an R package at https://bitbucket.org/gevaertlab/pancanceramaretto

© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the last two decades, advances in high-throughput experimental technologies have produced an abundance of molecular data. An increasing number of large multi-omics projects have launched and provide millions of data points for thousands of biological samples. For example, The Cancer Genome Atlas (TCGA) project (Hoadley et al., 2014; Cancer Genome Atlas Research Network, 2013; Yuan et al., 2014) was launched to improve our ability to diagnose, treat and prevent cancer and has produced an enormous amount of multi-omics data. Interpreting these high dimensional datasets to identify novel cancer driver genes represents an outstanding challenge. True cancer driver genes are those whose perturbation pushes a cell towards a malignant phenotype. Within this study, we define cancer driver genes as genes that fulfill all of the following criteria: (1) genes that are genetically and/or epigenetically deregulated in cancer, (2) genes whose genetic and epigenetic aberrations have a direct impact on their own functional gene expression levels, and (3) genes that are predicted to play regulatory roles high in the causal hierarchy of the origin of tumors. These include, for example, transcription factors, cell cycle genes or epigenetic modifying enzymes, whose altered state in cancer results in deregulation of downstream target genes; as well as upstream signaling molecules. They typically hide amongst a large number of passenger genes that are only by chance genetically or epigenetically altered (Eifert and Powers, 2012).

Previously, several computational methods have been developed to integrate multi-omics data. For example, Ciriello et al. used a method based on mutual exclusivity of copy number and mutation events to identify driver genes in glioblastoma (Ciriello et al., 2012). Similarly, Vandin et al. developed a method to identify driver genes in cancer, but focused on finding pathways with a significant enrichment of mutually exclusive genes (Vandin et al., 2012). In addition, Akavia et al. built further on this work and used copy number data to identify potential cancer driver genes in a modified Bayesian module network analysis called CONEXIC (Akavia et al., 2010). More recently, other groups are focusing on identifying driver genes through network analysis of copy

* Corresponding author at: Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine and & Department of Biomedical Data Science, Stanford University, 1265 Welch Road, Stanford, CA 94305, United States.
E-mail address: ogevaert@stanford.edu (O. Gevaert).
[1] Equally contributing first authors.

https://doi.org/10.1016/j.ebiom.2017.11.028
2352-3964/© 2017 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article as: Champion, M., et al., Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Respon..., EBioMedicine (2017), https://doi.org/10.1016/j.ebiom.2017.11.028

number data to identify potential drivers using a Bayesian module network analysis (Ray et al., 2014).

We have previously developed AMARETTO, an algorithm that integrates copy number, DNA methylation and gene expression data to identify a set of driver genes altered by DNA methylation or DNA copy number alterations, and constructs a gene expression network to connect them to clusters of co-expressed genes, defined as modules (Gevaert and Plevritis, 2013; Gevaert et al., 2013). These gene expression modules are subsequently ascribed biological pathways using gene set enrichment analysis (GSEA), revealing the pathways affected by cancer driver gene regulation. AMARETTO is thus a data driven pathway approach, using genomic, epigenomics and transcriptomics data as inputs, and produces modules and cancer driver genes associated with these modules as output. Integration of epigenomics data is essential to comprehensive analysis of cancer genomic analysis, as DNA methylation is a major mechanism of transcriptional deregulation in virtually all cancers. For example, cancer driver genes such as BRCA1 and MLH1, which are often altered by mutation in cancer, are also frequently deregulated by DNA methylation in other patients, with similar downstream consequences (Simpkins et al., 1999; Das and Singal, 2004; Catteau and Morris, 2002). Our data-driven pathway approach contrasts with previous work that relies upon use of known cancer pathways and networks such as PARADIGM, an algorithm that uses human-curated pathways and estimates their activity using DNA copy number and mRNA expression data (Vaske et al., 2010).

Here, we present an extension of AMARETTO to a pancancer application using multi-omics data of eleven cancer sites from TCGA. We show that AMARETTO captures modules enriched in major pathways of cancers and modules that accurately predict molecular subtypes. Next, we connect the modules of co-expressed genes in a pancancer module network. We show that this allows the identification of major oncogenic pathways and cancer driver genes involved in multiple cancers. More specifically, we identified a pancancer driver gene that is involved in smoking induced cancers and a pancancer driver gene that is involved in antiviral IFN modulated immune response. Overall, our results show the potential of pancancer multi-omics data fusion to identify cancer drivers that are high within the causal hierarchy of cancer development and associated with common pathways across different types of tumors that eventually can lead to the identification of pancancer drug targets. The AMARETTO algorithm and its pancancer application are publicly available.

2. Materials and Methods

2.1. Data Preprocessing

We used gene expression, copy number and DNA methylation data from TCGA for 11 cancer sites, namely bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon and rectal adenocarcinoma (COADREAD), glioblastoma (GBM), head and neck squamous cell carcinoma (HNSC), clear cell renal carcinoma (KIRC), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV) and uterine corpus endometrial carcinoma (UCEC) (Table 1). All data sets are available at the TCGA data portals.

The gene expression data were produced using Agilent microarrays for GBM and OV cancers, and RNA sequencing for all other cancer sites. Preprocessing was done by log-transformation and quantile normalization of the arrays. The DNA methylation data were generated using the Illumina Infinium Human Methylation 27 Bead Chip. DNA methylation was quantified using β-values ranging from 0 to 1 according to the DNA methylation levels. We removed CpG sites with more than 10% of missing values in all samples. We used the 15-K nearest neighbor algorithm to estimate the remaining missing values in the data set (Troyanskaya et al., 2001). Finally, the copy number data we used are produced by the Agilent Sure Print G3 Human CGH Microarray Kit 1Mx1M platform. This platform has high redundancy at the gene level, but we observed high correlation between probes matching the same gene. Therefore, probes matching the same gene were merged by taking the average. For all data sources, gene annotation was translated to official gene symbols based on the HUGO Gene Nomenclature Committee (version August 2012). TCGA samples are analyzed in batches and significant batch effects were observed based on a one-way analysis of variance in most data modes. We applied Combat to adjust for these effects (Johnson et al., 2007).

2.2. AMARETTO: Multi-omics Data Fusion

Our approach for analyzing TCGA cancer data is based on AMARETTO, a novel algorithm devoted to construct modules of co-expressed genes through the integration of multi-omics data (Gevaert and Plevritis, 2013; Gevaert et al., 2013). More precisely, AMARETTO is a three-step algorithm that (i) identifies tumor specific DNA copy number or DNA methylation changes, (ii) identifies a set of potential cancer driver genes by integrating DNA copy number, DNA methylation and gene expression data, (iii) connects these cancer driver genes to modules of co-expressed target genes that they control using a penalized regulatory program. AMARETTO, consists of three steps (Fig. 1).

2.2.1. Step 1

Identification of candidate cancer driver genes with tumor-specific DNA copy number or DNA methylation alterations compared to normal tissue: we first restrict the list of candidates to genes that have either copy number or DNA methylation alterations. These alterations are detected using the GISTIC (Taylor et al., 2008; Mermel et al., 2011) and MethylMix (Gevaert, 2015; Gevaert et al., 2015) algorithms for copy number and DNA methylation data respectively. GISTIC separately models arm-level and focal alterations, identifying amplified and

Table 1
Number of samples and number of genes for each of the data modalities (gene expression, DNA copy number and DNA methylation) and for the eleven studied cancer sites.

TCGA cancer site	TCGA cancer code	Gene expression		GISTIC		MethylMix	
		Samples	Genes	Samples	Genes[a]	Samples	Genes[b]
Bladder urothelial carcinoma	BLCA	181	15.432	178	1.974	123	472
Breast invasive carcinoma	BRCA	985	16.02	968	1.523	887	890
Colon and rectum adenocarcinoma	COADREAD	589	15.533	578	2.523	570	522
Glioblastoma multiforme	GBM	501	17.811	481	1.561	321	395
Head and neck squamous cell carcinoma	HNSC	371	15.828	365	2.184	308	753
Kidney renal clear cell carcinoma	KIRC	509	16.123	501	3.052	497	567
Acute myeloid leukemia	LAML	173	14.296	166	1.681	170	613
Lung adenocarcinoma	LUAD	489	16.092	487	3.585	367	678
Lung squamous cell carcinoma	LUSC	490	16.219	487	2.592	355	679
Ovarian serous cystadenocarcinoma	OV	541	17.814	528	1.499	540	510
Uterine corpus endometrial carcinoma	UCEC	508	15.706	500	2.074	496	821

[a] Number of significant genes found after running GISTIC (data available in the TCGA data portal).
[b] Number of genes with significant methylation patterns identified using MethylMix.

Please cite this article as: Champion, M., et al., Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Respon..., EBioMedicine (2017), https://doi.org/10.1016/j.ebiom.2017.11.028

Download English Version:

https://daneshyari.com/en/article/8437602

Download Persian Version:

https://daneshyari.com/article/8437602

Daneshyari.com