A computational framework for the identification of genes modules epigenetic regulated and targetable by drugs at the level of single patient.

Tagliazucchi Malagoli G(1)[†] , Taccioli C(2)

(1) Azienda Ospedaliero Universitaria di Parma, Department of Research and Innovation, 43126 Parma, Italy
(2) Dipartimento di Medicina Animale, Produzione e Salute, Università di Padova, 35020 Legnaro, Padova, Italy.

[†] Email: `guidantonio.mt@gmail.com`

Motivation

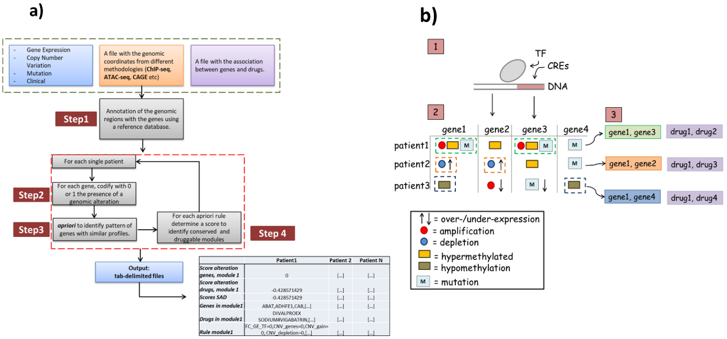
The integration of "omics" data with data-fusion methods [1] from Next Generation Sequencing (NGS) technologies provides a great potential to better understand the molecular basis of tumors and stratify different cancer subtypes [2]. Moreover, the increasing amount of NGS data, introduces the possibility to further develop precision and personalized-medicine. The USNRC define the personalized medicine as "the situation in which therapeutics are synthesized for specific individuals" [3]. In this context, the integration of omics data can also provide an important step for the identification of altered biological mechanisms at the level of single patient. Diseases can arise from multiple steps that involved also epigenetic mechanisms. However, at the present moment, few tools able to integrate omics data and study the role of epigenetic components (e.g. histone marks) on groups of genes at the level of single patient. Here, we describe a framework, GMIEC, that integrates omics data (gene-expression, copy number variation, mutation, methylation) with data containing genomic coordinates of cis and/or trans epigenetic regulators (e.g. ChIP-seq). Then, at the level of the single patient, GMIEC identify groups of genes (modules) that share common genomic features. Then, using an external database, the genes in each module are associated with their own target drugs. Therefore, GMIEC allows the identification, at the level of the single patient, groups of genes that are regulated by the same epigenetic regulators (e.g. TFs), sharing common genomic features and associated with different number of drugs.

Methods

The input of GMIEC are: i) a file with the genomic regions of the regulators of interest; ii) omics data; ii) a file that includes the association between genes and drugs. The four principal steps (Figure 1A) of analysis are: i) the annotation of the epigenetic components (ECs) derived from a catalogue (ReMap [3]) with their target genes using ChIPpeakAnno[4]; ii) a step of discretization used to mark the presence of one alteration (1) or not (0) for all genes of single patient, considering all datasets (e.g. gene-expression data, copy-number alterations data, mutations data); iii) the genes with the similar genomic features grouped using Association Rules Discovery (ARD) technique. This is a data mining method used to discover associations between subsets of items and large dataset extensively used in bioinformatics [5], [6]. In our case, (Figure 1B) if two genes, target of the ECs, share similar common genomic features (e.g. amplification, hypermethylation), then they are grouped together. Finally, a combined score (SAD) is computed to verify if a specific group of genes are altered or not and if they are targets of drugs or not. GMIEC is provided as an R-script (https://github.com/guidmt/gmiec) and depends on the R-package “arules”. To test GMIEC we used The Cancer Genome Atlas BRCA datasets.

Results

We applied GMIEC on a dataset of Breast Cancer TCGA data and considering RUNX1 as regulator. RUNX1 was identified as strongly expressed in breast cancer epithelia and dysregulated during tumorigenesis. Here, GMIEC was set to select only the group of genes with few genomic alterations. Considering only the samples of the basal-like subtype we identified 18 patients with only 1 gene in their module, 23 patients with more than 2 genes. Nine of these samples contained modules with more than 16 genes (range from 16 to 131). Our analysis indicates that some of the patient modules were also associated with anti-neoplastic agents. These results were confirmed by further analyses performed using oncoScore and DrugPattern [7]. Thus, our strategy proved its effectiveness in the identification of subtle differences along samples and patients. In summary, GMIEC is able to: i) stratify samples and patients; ii) identify patient specific genes; iii) plan new treatment strategies personalized for cancer patients.


a)
- Gene Expression
- Copy Number Variation
- Mutation
- Clinical
A file with the genomic coordinates from different methodologies (ChIP-seq, ATAC-seq, CAGE etc)
A file with the association between genes and drugs.
Step1
Annotation of the genomic regions with the genes using a reference database.
For each single patient
Step2
For each gene, codify with 0 or 1 the presence of a genomic alteration
Step3
apriori to identify pattern of genes with similar profiles.
For each apriori rule determine a score to identify conserved and druggable modules
Step 4
Output: tab-delimited files
b)
1
TF
CREs
DNA
2
gene1
gene2
gene3
gene4
patient1
patient2
patient3
M
3
gene1, gene3
drug1, drug2
gene1, gene2
drug1, drug3
gene1, gene4
drug1, drug4
↑↓ = over-/under-expression
● = amplification
● = depletion
▬ = hypermethylated
▬ = hypomethylation
M = mutation

	Patient1	Patient 2	Patient N
Score alteration genes, module 1	0	[...]	[...]
Score alteration drugs, module 1	-0.428571429	[...]	[...]
Scores SAD	-0.428571429	[...]	[...]
Genes in module1	ABAT,ADHFE1,CA8,[...]	[...]	[...]
Drugs in module1	DIVALPROEX SODIUM#VIGABATRIN,[...]	[...]	[...]
Rule module1	FC_GE_TF=0,CNV_genes=0,CNV_gain=0,CNV_depletion=0,[...]	[...]	[...]

References

1. Methods for the integration of multi-omics data: mathematical aspects. Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, Luciano Milanesi BMC Bioinformatics. 2016; 17(Suppl 2): 15.

2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70.

3. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. Grif-

fon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., Ballester, B. Nucleic Acids Research, Volume 43, Issue 4, 27 February 2015.

4. Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR.ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010 May 11;11:237.

5. Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics. 2006 Feb 7;7:54.

6. Park SH, Lee SM, Kim YJ, Kim S. ChARM: Discovery of combinatorial chromatin modification patterns in hepatitis B virus X-transformed mouse liver cancer using association rule mining. BMC Bioinformatics. 2016 Dec 13;17.

7. Piazza R, Ramazzotti D, Spinelli R, Pirola A, De Sano L, Ferrari P, Magistroni V, Cordani N, Sharma N, Gambacorti-Passerini C. OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. Sci Rep. 2017 Apr 25.