Interpreting pathways to discover cancer driver genes with Moonlight

Antonio Colaprico [1,2,3,19,*], Catharina Olsen [1,2,4,5,19], Matthew H. Bailey [6,7], Gabriel J. Odom [3,8], Thilde Terkelsen [9], Tiago C. Silva [3,10], Andre Vidas Olsen [9], Laura Cantini [11,12,13,14], Andrei Zinovyev [11,12,13], Emmanuel Barillot [11,12,13], Houtan Noushmehr [10,15], Gloria Bertoli [16], Isabella Castiglioni [16], Claudia Cava [16], Gianluca Bontempi [1,2,20], Xi Chen [3,17,20], Elena Papaleo [9,18,20]

[1]Interuniversity Institute of Bioinformatics in Brussels (IB)², Brussels, Belgium [2]Machine Learning Group, ULB, Brussels, Belgium [3]Department of Public Health Sciences, University of Miami, Miller School of Medicine, Miami, FL 33136, USA
[4]Center for Medical Genetics, Reproduction and Genetics, Reproduction Genetics and Regenerative Medicine, Vrije Universiteit Brussel, UZ Brussel,Laarbeeklaan 101, 1090 Brussels, Belgium
[5]Brussels Interuniversity Genomics High Throughput core (BRIGHTcore), VUB-ULB, Laarbeeklaan 101, 1090 Brussels, Belgium. [6]Division of Oncology, Department of Medicine, Washington University in St. Louis, St. Louis, MO 63110, USA
[7]McDonnell Genome Institute, Washington University, St. Louis, MO 63108, USA [8]Department of Biostatistics, Stempel College of Public Health, Florida International University, Miami, FL 33199, USA
[9]Computational Biology Laboratory, and Center for Autophagy, Recycling and Disease, Danish Cancer Society Research Center, Strandboulevarden 49, 2100, Copenhagen, Denmark [10]Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, RibeirSão Preto, Brazil
[11]Institut Curie, 26 rue d'Ulm, F-75248 Paris France [12]INSERM, U900, Paris, F-75248 France [13]Mines ParisTech, Fontainebleau, F-77300 France
[14]Computational Systems Biology Team, Institut de Biologie de lâEcole Normale Supérieure, CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, Paris Sciences et Lettres Research University, 75005 Paris, France.
[15]Department of Neurosurgery, Brain Tumor Center, Henry Ford Health System, Detroit, MI, USA [16]Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), Milan, Italy
[17]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL 33136, USA [18]Translational Disease System Biology, Faculty of Health and Medical Science, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
[19]These authors contributed equally: Antonio Colaprico and Catharina Olsen [20]These authors jointly supervised this work: Gianluca Bontempi, Xi Steven Chen, Elena Papaleo. email (Corresponding Authors): *mailto*axc1833@med.miami.edu

Abstract

We present Moonlight, a tool that uses multiple -omics data to discriminate between oncogenes, tumor suppressors, and dual-role genes by leveraging context-specific gene programs. We applied Moonlight to over 8,000 tumors from 18 cancer types and predicted 160 dual-role genes that contribute most to this effect-switching phenomenon.
We discovered that tissue type and molecular subtype indicate dual-role status.
Moonlight elucidated the underlying biological mechanisms of these genes.

Introduction

Cancer is an extremely complex disease, hallmarked by the poor regulation of critical growth, proliferation, and apoptotic pathways. Over the last 12 years, The Cancer Genome Atlas (**TCGA**) has explored the heterogeneous nature of this disease using many high-throughput technologies.
In order to better understand the hallmarks of cancer, such as biological processes (**BPs**) (e.g. proliferation, apoptosis, invasion of cells, etc.), it is critical to accurately identify cancer-driver genes (**CDGs**) and their roles in specific tissues. CDGs are traditionally classified as either oncogenes (**OCGs**) or tumor suppressor genes (**TSGs**), depending on their role in cancer development. The gain-of-function of OCGs together with the loss-of-function of TSGs determine the processes that control tumor formation and development. However, certain CDGs may exhibit OCG or TSG behavior depending on biological context, which we define as dual-role genes. In particular, we and others within the TCGA Pan-Cancer Atlas initiative, employed 26 computational tools to identify 299 CDGs and more than 3,400 driver mutations, which represent potentially actionable oncogenic events [1].
Although all these methods were demonstrated as effective, it remains critical to clarify the consequences of each mutation and their link with possible underlying biological interpretation as well as downstream effects.

Moonlight data integration and functionalities

We developed the tool **Moonlight** [2] which detects CDG events specific to the tumor and tissue of origin including potential dual-role genes but also elucidates their downstream impact. In order to accomplish this, Moonlight distills information from literature, pathways, and multiple -omics data into a comprehensive assessment of a gene's role and function.

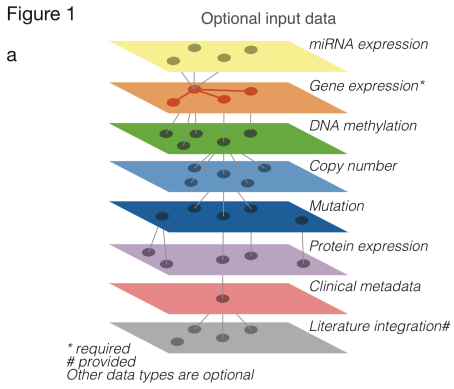

Figure 1: **(a)** Possible input datatypes. **(b)** Moonlight pipeline for discovery of TSGs, OCGs, and dual-role genes

Moonlight [2] offers two main approaches (expert based or machine learning) to identify CDGs. Also, this tool potentially shows the impact of the predicted CDGs on multiple -omics data which may represent the mechanisms for gene inactivation or activation.
The two approaches share three initial steps: (i) Moonlight identifies a set of Differentially Expressed Genes (**DEGs** between two conditions, then (ii) the gene expression data is used to infer a Gene Regulatory Network (**GRN**) with the DEGs as vertices, and (iii) using Functional Enrichment Analysis (**FEA**) quantifies the DEG-BP association with a Moonlight Z-score. Finally, we input DEGs and their GRN to Upstream Regulatory Analysis (**URA**), yielding upstream regulators of BPs mediated by the DEG and its targets.

In the first approach, Pattern Recognition Analysis (**PRA**) takes in two objects: (i) URA's output and (ii) selection of a subset of the BP provided by the end-user. In contrast, if the BPs are not provided, their selection is automated by a machine learning method (e.g. random forest model) trained on gold standard OCGs-TSGs in the second approach. In addition, Dynamic Recognition Analysis (**DRA**) detects multiple patterns of BPs when different conditions are selected

Moonlight case studies and published applications

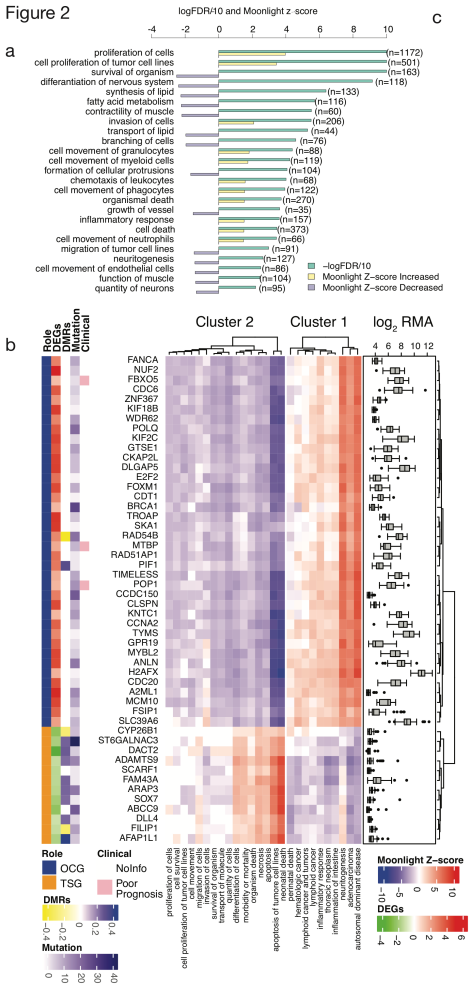

Figure 2: **(a)** Barplot from FEA showing the BPs significantly enriched in breast cancer. **(b)** Heatmap showing top 50 predicted TSGs and OCGs in breast cancer along with BPs regulated.

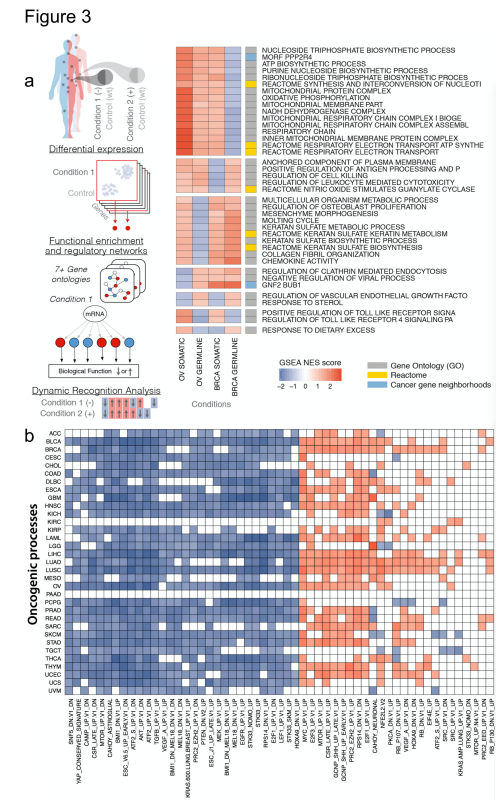

Figure 3: **(a)** Moonlight PRA-DRA analysis revelead pathways that are differentially expressed depending on whether the mutations affecting BRCA1 and/or BRCA2 are somatic (condition 1) or germline (condition 2) in comparison versus wild-type [3]. **(b)** Moonlight PRA-DRA analysis further validated gene-expression-based Stemness index and confirmed engagement of MYC and EZH2, along with E2F3, MTOR, and SHH in driving oncogenic dedifferentiation. [4]

Software Availability

Multiple -omics data were downloaded from Genomic Data Commons (GDC) legacy archive, normalized and analyzed using the R package **TCGAbiolinks** `http://bioconductor.org/packages/TCGAbiolinks/` [5]
Moonlight is freely available as an open-source R package within the Bioconductor project at `http://bioconductor.org/packages/MoonlightR/`. [2]

Funding

The project was supported by BridgeIRIS and GENGISCAN projects (Belgian FNRS PDR T100914F) to A.C., C.O. and Gi.B; I.C, C.C. and GI.B. were supported by INTEROMICS project (91J12000190001). E.P., T.T. and A.V.O were supported by a grant from LEO Foundation (LF17006) and InnovationFund Denmark (5189-00052B). A.C., G.J.O. and X. C. were supported by grants from NCI R01CA200987, R01CA158472 and U24CA210954.

References

Bailey, M. H. *et al.*
Comprehensive characterization of cancer driver genes and mutations.
Cell **173**, 371 – 385.e18 (2018).

Colaprico, A. *et al.*
Interpreting pathways to discover cancer driver genes with moonlight.
Nature Communications **11**, 69 (2020).
URL `https://doi.org/10.1038/s41467-019-13803-0`.

Ding, L. *et al.*
Perspective on oncogenic processes at the end of the beginning of cancer genomics.
Cell **173**, 305 – 320.e10 (2018).

Malta, T. M. *et al.*
Machine learning identifies stemness features associated with oncogenic dedifferentiation.
Cell **173**, 338 – 354.e15 (2018).

Colaprico, A. *et al.*
Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data.
Nucleic Acids Research **44**, e71 (2016).