RESEARCH ARTICLE

Parent-child signals identify candidate cancer driver genes

Emilie Ann Ramsahai¹, Vrijesh Tripathi¹, Melford John²

¹Department of Mathematics and Statistics, The University of the West Indies, St Augustine, Trinidad and Tobago
²Department of Preclinical Sciences, The University of the West Indies, St Augustine, Trinidad and Tobago

First published: 03 Feb 2021, 10:67
https://doi.org/10.12688/f1000research.22391.1

Abstract

Background: The DREAM Challenge evaluated methods to identify molecular pathways facilitating the detection of multiple genes affecting critical interactions and processes. Dysregulation of pathways by well-known driver genes is often found in the development and progression of cancer. We used the gene interaction networks provided and the scoring rounds to test disease module identification methods to nominate candidate driver genes in these modules.

Method: Our algorithm calculated the proportion of the whole network accessible in two steps from each node in a combined network, which was defined as a 2-reach gene value. Genes with high 2-reach values were used to form the center of star cover clusters. These clusters were assessed for significant modules. Within these modules we identified novel candidate driver genes, by considering the parent-child relationship of well-known driver genes. Disturbance to such driver genes or their upstream parents, can lead to disruption of highly regulated signals affecting the normal functions of cells. We explored these parents as a potential source for candidate driver genes.

Results: An initial list of 57 candidate driver genes was identified from 13 significant modules. Analysis of the parent-child relationships of well-known driver genes in these modules prioritized PRKDC, YWHAB, GSK3B, and PPP1CB.

Conclusion: Our method incorporated the simple m-reach topology metric in disease module identification and its relationship with known driver genes to identify candidate genes. The four genes shortlisted have been highlighted in recent publications in the literature, which supports the need for further wet lab experimental investigation.

Keywords
driver gene, cancer, module identification, pathway analysis, network biology, topology, reach metric, signaling network
This article is included in the DREAM Challenges gateway.

Corresponding author: Emilie Ann Ramsahai (emilieannramsahai@gmail.com)

Author roles: Ramsahai EA: Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation; Tripathi V: Supervision; John M: Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the University of the West Indies, independent of grants.

Copyright: © 2021 Ramsahai EA et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ramsahai EA, Tripathi V and John M. Parent-child signals identify candidate cancer driver genes [version 1; peer review: 1 approved with reservations] F1000Research 2021, 10:67 https://doi.org/10.12688/f1000research.22391.1

First published: 03 Feb 2021, 10:67 https://doi.org/10.12688/f1000research.22391.1
**Introduction**

Cancer is a disease of uncontrolled cell proliferation. Genetic mutations alter operations inside normal cells in ways that promote tumorigenesis. Receptors at the cell surface gather the signals from other cells, and funnel them into the cell. Signals are transmitted from upstream proteins and passed to downstream effector proteins. These mappings are represented in signaling networks, one of which was provided as a weighted directed graph by the DREAM Challenge. Dysregulation of these pathways by driver mutations is often found in the development and progression of cancer. Known driver genes, as listed in the Cancer Genome Census (CGC), within this signaling network provided points of reference. Disturbance to these driver genes or their upstream parents by mutations in either parent or child, can lead to disruption of highly regulated signals affecting the normal functions of cells. We explored these parents as potential candidate driver genes. We also considered driver gene parent protein products that were functionally related in physical protein-protein interaction (PPI) networks and co-expression networks as candidate driver genes.

Module or community detection is a classical problem in social and computer network science. Different methods yield communities that are either too large or too small to be easily understood. Biologically relevant communities and their relevance to a disease are poorly understood. With the increased availability of biological network data, scientists are no longer solely relying on fixed pathways. They are now seeking to expand on known pathways and discover novel pathways in the analysis of diseases. The challenge assessed our modules using a number of genome-wide association studies (GWAS). Each of our modules were scored for enrichment against 104 GWAS, using their PASCAL tool. We then classified these novel disease pathways with CGC known driver genes as cancer modules. The signaling parent genes, which formed part of these novel cancer pathways, were nominated as candidate driver genes.

**Methods**

**Construction of combined network**

The co-expression and physical Protein-Protein Interaction Network provided in Subchallenge 2 were combined. The networks were maintained as interaction lists, which proved to be more efficient than the memory demanding matrix manipulation. The network combination method employed was based on the system developed previously. This network combination process considered the interaction existence and weighting in each of the individual networks before accepting it as part of the combined network. Linear regression was used in the calculation of the weighting in the combined network.

The individual networks are summarized in Table 1. These anonymized networks prevented any bias in the module identification process. Edges in the combined network were weighted and during the leaderboard stages this weight and its cutoff values were optimized to construct a final combined network with high-scoring edges. The clustering technique was employed on this combined network, while the signaling network was used to determine the parent-child relationships.

**The 2-reach center cluster formation**

The number of nodes accessible within two steps as a proportion of the total nodes in the network, is defined as the 2-reach of that node. This 2-reach value was calculated for each node in the combined network. By repeatedly selecting two genes with the highest 2-reach values from the set of genes not yet assigned to a cluster, the complete network was decomposed. These two genes were the center of two separate clusters and their immediate neighbors chosen as members. To prevent overlapping, clusters were removed from the network, resulting in further fragmentation; such fragments give rise to clusters themselves as they may be within a range of 3–100 members. The largest remaining fragment was decomposed by repeating the 2-reach cluster formation process.

**DREAM Challenge scores**

The DREAM Challenge assessed modules based on a collection of GWAS, which is superior to matching the predicted modules against well-known pathway databases or annotated information. The unique GWAS used in the assessment was not part of the construction of the networks provided. The Challenge used these GWAS and our modules as input to the PASCAL scoring tool, which provided 104 p-values for each of our modules in the final submission. With these multiple p-values, for each of our modules, we were then required to perform a level of correction to control the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure, as some of the p-values less than 0.05 may have been by chance.

**Cancer modules with parent-child drivers**

Significant modules (SM) with at least one driver gene from the CGC list were classified as cancer modules (CM). These cancer modules were a subset of the set of significant disease modules as depicted in Figure 1A. The list of genes in these cancer modules (CMgenes), were a source of potential driver genes. Additionally, from the signaling network in Table 1,
Figure 1. Overlapping modules and genes. (A) The set of significant modules (SM) with well-known driver genes from the Cancer Genome Census list were identified as cancer modules (CM). (B) The intersection of the genes from the cancer modules CMgenes and the parents from the signaling network Pgenes highlight our first list of candidate driver genes.

we identified a list of parent genes (Pgenes) to well-known driver genes as another source of potential driver genes. Figure 1B highlights the intersection of these two lists. The CMgenes were based on the network module identification and GWAS scoring, while the Pgenes were based on the parent-child relationships with well-known driver genes in the signaling network. The overlap of these two lists provided the list of candidate driver genes based on a consensus.

The scripts and source code used to perform the analysis in this study are available as Extended data1.

Results

Networks

The network formed by combing the co-expression network and the protein-protein interaction (PPI) network consisted of 14,200 genes and 824,528 edges. The removal of low-confidence edges resulted in the loss of 1,583 genes, 199 of which were unique to the co-expression network, 1,201 of which were unique to the PPI network, and 183 of which were common to both. The overlap of these two DREAM Challenge networks and the combined network is illustrated in Figure 2A.

The signaling network included 5,254 nodes and 21,826 edges. Of the 616 CGC driver genes 470 were present in this network. These driver genes were children of 1,721 parents, with TP53 having 154 (the highest number of parents); 63 other driver genes were orphans. All of the orphan driver genes were themselves parents in this signaling network. Figure 2B shows the overlap of the signaling network, the CGC driver genes and their parents. We also see that 306 of the CGC driver genes are both parents and children of other driver genes.

Modules

Our 2-reach clustering algorithm produced 237 non-overlapping modules, ranging in size from 3 to 100, as shown in Figure 3A. Many of our modules were 50 genes in size, as this was the default module size in the cluster formation process. Our method also produced modules that were at the upper bound of 100 genes and the lower bound of 3 genes in size, a result of fragmentation of the network. The 237 modules extracted from the combined network included 4,682 genes, which was less than 33% of the total network, since only the largest fragment was further assessed for clusters.

On examination of the 104 multiple p-value scores provided by the DREAM Challenge for each of our modules, we found 27 significant modules existed, all shown in Figure 3B. The 13 solid circles are the CM, which contain known driver genes. It is difficult to get highly significant enrichment p-values for modules of a small size. Consider a module of size 4 that includes two significant genes. Even though 50% is certainly a clear enrichment, this could also occur by chance. In contrast, a module of size 100 that contains 50% significant genes would get an extremely small p-value.

Cancer pathways and driver genes

The 35 driver genes in these 13 cancer modules contained 84 unique parents from the signaling network. In total, 27 of the 84 parents were also driver genes, which left 57 as candidate driver genes. Many of the driver genes were also parents, but module 20, module 55, module 109, and module 143 contained driver genes that were not parents, since the “#Drivers in module” column value was larger than the number of red genes provided in Table 2. In the case of module 55, with as many as nine known driver genes, only five were parents in the signaling network. Module 143 has a unique condition of containing the single CGC driver gene HLA-A, which was not a parent, and therefore does not appear in Table 2. The presence of this driver gene, qualified module 143 to be a cancer module, but only the parents of other known driver genes were considered as candidate drivers. Each of the cancer
modules CM and the candidate driver genes they contain are listed in Table 2; candidate genes are listed in black.

Additional analysis of the cancer modules further prioritized our initial list; we assessed the specific parent child relationships in the subgraphs using the parent genes for each of the modules in the signaling network. On examination of the 13 subgraphs, only three (modules 15, 19 and 55) had direct parent-child relationships, where the parent was listed in the module and the child was a CGC well known driver gene. In the case of module 15, there were five driver genes \( \text{FBXW7}, \text{THRAP3}, \text{ABL1}, \text{FANCD2} \) and \( \text{YWHAE} \), with 13 parent genes. The five driver genes were themselves parent genes, so we were left with eight parents to consider as potential candidate driver genes as listed in Table 2. Of the five driver genes in this module, only \( \text{ABL1} \) had a parent from this same module in the signaling network – \( \text{PRKDC} \), as seen in Figure 5. Similarly, module 19 showed the \( \text{RAF1} \) driver gene had three parents in the...
Figure 4. Team results. The y-axis represents the number of significant (NS) modules identified in the final phase of subchallenge 2.

Table 2. Candidate Driver Gene Selection.

| Module Id | # Drivers In module | Parent Genes in Cancer Modules- Pgenes in CM | # Candidate |
|-----------|---------------------|---------------------------------------------|-------------|
| 15        | 5                   | UBR2 NUP153 EP400 KPNA4 FBXW7 THRAP3 ABL1 PRKDC CDC25A FNTB APK8IP1 FANCD2 YWHAE | 8           |
| 19        | 4                   | JAG1 DLG1 DGKZ MYOD1 CDK4 PTK2B YWHAB SETD2 STK10 MAP3K3 RAF1 CTNND1 GSK3B | 9           |
| 20        | 3                   | FLT3LG PRR5L SMARCA4 CEBPZ PTPN11 | 3           |
| 25        | 2                   | PDGFRB PAWR CDC20 TAOK2 FOXO3 | 3           |
| 30        | 3                   | PTPRB NFATC1 NOL3 MSN HIF1A PDLA6 HEY2 DLG4 PPARGC1A TRAF1 | 7           |
| 55        | 9                   | CAMK1 MEN1 RB1 FANCF BID IKBKG PTPRE FUBP1 POGlut1 CFLAR PALB2 UBE2B PPP1CB | 8           |
| 71        | 1                   | CCDC88A XIAP GNA13 GABPA ARID4A RIPK1 USP8 MAST2 MXD1 TIAM1 MAP3K2 | 10          |
| 109       | 3                   | BUB1 PSMC3IP NCAPH BRCA2 POU2F1 CASP8AP2 GRK6 | 6           |
| 143       | 1                   | TAP1 TAPBP | 2           |
| 172       | 1                   | CASP8 | 0           |
| 177       | 1                   | CD74 | 0           |
| 190       | 1                   | CD79B CD19 | 1          |
| 205       | 1                   | BUB1B | 0           |
| TOTAL     | 35                  | | 57          |

Red: Known driver genes which are also parents; Black: Parents and possible candidate genes; Underlined: Parents in a parent-driver gene relationship in the given cancer module.
signaling network, CDK4, YWHAB, and GSK3B. Module 55 has nine driver genes, but only five were parents. RB1 was the only driver gene with the neighboring parent PPP1CB from the signaling network. This process prioritized five parents, which are underlined in Table 2. They are neighbors of existing drivers in the cancer modules, and are parents to the same driver gene in the signaling network. Of these five, CDK4 is already present in the CGC list of known driver genes. The other four genes PRKDC, YWHAB, GSK3B, and PPP1CB are shortlisted.

**Discussion/conclusion**

We used three diverse genomic networks to reveal novel candidate genes from pathways underlying cancer: the physical PPI, the functional co-expression, and the signaling networks. Our method incorporated the simple m-reach topology metric in disease module identification. The reach measure has been shown to be useful in the key player problem\(^1\), and has been used to identify cancer driver genes\(^4\). Recent publications in the literature highlight the four candidate genes that were shortlisted. PRKDC has been shown to be associated with poor clinical outcome in gastric cancer patients\(^5\). Recent studies proposed the YWHA family is able to regulate a vast number of proteins involved in key cellular processes, with implications for tumorigenesis and cancer progression\(^6\). GSK3B is known to regulate epithelial-mesenchymal transition and cancer stem cell properties, and is a novel drug target for triple-negative breast cancer\(^7\). PPP1CB has been identified as a protein safeguarding nuclear integrity; altered nuclear shape is a defining feature of cancer cells\(^8\).

**Data availability**

**Underlying data**

Synapse: Disease Module Identification DREAM Challenge. Synapse ID syn11944943.

**Extended data**

Zenodo: Parent-child signals identify candidate cancer driver genes. [http://doi.org/10.5281/zenodo.3740805](http://doi.org/10.5281/zenodo.3740805).

This project contains the scripts and source code used to perform the analysis in this study.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0).

**Author contributions**

ER worked with the DREAM Challenge, downloading data, running analysis, preparing and submitting results. ER also wrote the manuscript, which the other authors reviewed, corrected and approved.
1. Futreal PA, Coin L, Marshall M, et al.: A census of human cancer genes. Nat Rev Cancer. 2004; 4(3): 177–83.
2. Fortunato S: Community detection in graphs. Phys Rep. 2010; 486(3–5): 75–174.
3. Bishop JM: The molecular genetics of cancer. Science. (New York, NY). 1987; 236(486): 305–11.
4. Newman ME: The structure and function of complex networks. SIAM review. 2003; 45(2): 167–256.
5. Wang J, Zuo Y, Man YG, et al.: Pathway and network approaches for identification of cancer signature markers from omics data. J Cancer. 2015; 6(1): 54–65.
6. Mitra K, Carvunis AR, Ramesh SK, et al.: Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013; 14(10): 719–32.
7. Wilkinson DM, Huberman BA: A method for finding communities of related genes. Proc Natl Acad Sci U S A. 2004; 101(supp1): 5241–8.
8. Lamparter D, Marbach D, Ruedi R, et al.: Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput Biol. 2016; 12(1): e1004714.
9. Ramsahai E, Walkins K, Tripathi V, et al.: The use of gene interaction networks to improve the identification of cancer driver genes. PeerJ. 2017; 5: e2568.
10. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995; 57(1): 289–300.
11. Ramsahai E, Tripathi V, John M: Parent-child signals identify candidate cancer driver genes. 2020. http://www.doi.org/10.5281/zenodo.3740805
12. Choobdar S, Ahsen ME, Crawford J, et al.: Assessment of network module identification across complex diseases. Nat methods. 2019; 16(9): 843–852.
13. Borgatti SP: The Key Player Problem. Dynamic social network modeling and analysis: Workshop summary and papers. National Academies Press. 2003.
14. Ramsahai E, Tripathi V, John M: Cancer driver genes: a guilty by resemblance doctrine. PeerJ. 2019; 7: e6979.
15. Sotgia F, Lisanti MP: Mitochondrial biomarkers predict tumor progression and poor overall survival in gastric cancers: Companion diagnostics for personalized medicine. Oncotarget. 2017; 8(40): 67117–67128.
16. Moore S, Järvelin AI, Davis I, et al.: Expanding horizons: new roles for non-canonical RNA-binding proteins in cancer. Curr Opin Genet Dev. 2018; 48: 112–120.
17. Raja GV: GSK3B regulates epithelial-mesenchymal transition and cancer stem cell properties and is a novel drug target for triple-negative breast cancer. 2017. Reference Source
18. Takaki T, Montagner M, Serres MP, et al.: Actomyosin drives cancer cell nuclear dysmorphia and threatens genome stability. Nat commun. 2017; 8: 16013.

References
Open Peer Review

Current Peer Review Status:  

Version 1

Reviewer Report 27 May 2021

https://doi.org/10.5256/f1000research.24706.r83221

© 2021 Kazan H. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Hilal Kazan
Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey

The study extends a method that is originally proposed for finding network modules to cancer driver gene identification task. Cancer driver identification is a well-studied problem in bioinformatics. However, the manuscript does not have enough discussion on the existing literature. Methods such as Hotnet2, MEMCover, DriverNet can be cited.

Also, a graph theoretical measurement, betweenness centrality, has been recently used to find cancer driver genes (https://doi.org/10.1186/s12859-021-03989-w). Since it is related to 2-reach concept, it'd be useful if it is included in the literature discussion.

In the current version of the study, it is very difficult to assess the ability of the proposed method for cancer driver identification task. Several existing methods that predict cancer driver genes using PPI and heat diffusion might be performing better than the current model.

Also there are several problems regarding the content of the manuscript:

- The authors refer to DREAM challenge throughout the paper, however many DREAM challenges have been organized to date and it is not immediately clear which one they refer to. Referring to it as “Disease Module Identification DREAM challenge” would be better.

- Also, the aim and details of this challenge should be better explained.

- It is not clear what the authors mean by anonymity in the following sentence: “These anonymized networks prevented any bias in the module identification process.”

- What’s the intuition behind selecting two genes simultaneously in The 2-reach center cluster formation?

- Figure 1A can be removed, Figure 3’s should be converted to histograms as module ids are not critical.
Figure 4 can be removed. The performance of the cluster identification method in the DREAM challenge is not relevant in the current manuscript.

References
1. Erten C, Houdjedj A, Kazan H: Ranking cancer drivers via betweenness-based outlier detection and random walks. *BMC Bioinformatics*. 2021; 22 (1): 62 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** cancer genomics, statistical genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com