Identification of ovarian cancer driver genes by using module network integration of multi-omics data

Olivier Gevaert¹, Victor Villalobos², Branimir I. Sikic² and Sylvia K. Plevritis¹

1Cancer Center for Systems Biology, Department of Radiology, Stanford University, Lucas Center for Imaging, 1201 Welch Road, Stanford, CA 94305, USA
2Division of Medical Oncology, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA

The increasing availability of multi-omics cancer datasets has created a new opportunity for data integration that promises a more comprehensive understanding of cancer. The challenge is to develop mathematical methods that allow the integration and extraction of knowledge from large datasets such as The Cancer Genome Atlas (TCGA). This has led to the development of a variety of omics profiles that are highly correlated with each other, however it remains unknown which profile is the most meaningful and how to efficiently integrate different omics profiles. We developed AMARETTO, an algorithm to identify cancer drivers by integrating a variety of omics data from cancer and normal tissue. AMARETTO first models the effects of genomic/epigenomic data on disease-specific gene expression. AMARETTO’s second step involves constructing a module network to connect the cancer drivers with their downstream targets. We observed that more gene expression variation can be explained when using disease-specific gene expression data. We applied AMARETTO to the ovarian cancer TCGA data and identified several cancer driver genes of interest, including novel genes in addition to known drivers of cancer. Finally, we showed that certain modules are predictive of good versus poor outcome, and the associated drivers were related to DNA repair pathways.

1. Introduction

The unprecedented wealth of data currently being generated for cancer patients has provided us with the challenge in its interpretation and translation to personalized medicine. Personalized medicine aims to tailor medical care to the individual through the meaningful characterization of biological heterogeneity present in cancer. Technological innovation has enabled the acquisition of multiscale information ranging from genotypes to several phenotypic layers. For example, advances in high throughput sequencing allow quantification of global DNA variation and RNA expression of tissue or blood samples [1–3]. These platforms produce a variety of omics profiles that, while highly correlated to each other, often raise difficulty in discerning meaningful interpretation and integration.

Previous data integration efforts in cancer have focused on integrating a subset of omics profiles. For example, Ciriello et al. [4] used a method based on mutual exclusivity to model copy number and mutation data and identified driver genes in glioblastoma. Similarly, Vandin et al. [5] developed a method to identify driver genes in cancer, but focused on finding pathways with a significant enrichment of approximately mutually exclusive genes. In addition, other groups are focusing on identifying driver genes through network analysis of copy number data to filter potential regulators in Bayesian module network analysis [6].

We selected a module network approach to integrate copy number, DNA methylation and gene expression data. We developed an algorithm called AMARETTO to unravel cancer drivers by using data integration of omics data, conditioned on differential expression between cancer and normal samples. We applied AMARETTO to the ovarian cancer data from The Cancer Genome Atlas.
Figure 1. Workflow of AMARETTO. Step 1 involves generating a list of candidate cancer driver genes by using a linear model capturing the relationship between genomic and transcriptomic data for each gene separately. First, we create disease-specific gene expression profiles (DSE) by taking normal gene expression profiles into account. Secondly, only genes identified by GISTIC and MethylMix are modelled. Step 2 connects the cancer drivers from step 1 with their downstream targets by reconstructing a module network. This module network associates a set of co-expressed genes with cancer driver genes through learning a regulatory programme. The regulatory programme is modelled using linear regression with elastic net regularization. (Online version in colour.)

(TCGA) project [7]. AMARETTO builds further on previous work by adding a number of new concepts [8]. AMARETTO focuses on identifying what is specifically aberrant in cancer tissue compared with normal tissue, although matched cancer and normal samples are not required for each patient; this specific aspect is integrated throughout our algorithm. In addition, AMARETTO focuses on candidate cancer drivers that are recurrent in different samples and are regarded functional. Drivers are regarded as functional if there is a significant relationship between the genomic/epigenomic event and their resulting gene expression. AMARETTO also incorporates DNA methylation data, which to date are often not integrated due to lack of methods to extract cancer-specific DNA methylation aberrations. To address this issue, we developed a method called MethylMix [9] that allows us to integrate DNA methylation in our workflow as a complementary driving genomic force of cancer cells, thus enabling us to expand beyond copy number aberrations.

We applied AMARETTO to identify driver genes based on multi-omic profiles from 511 ovarian cancer patients. In step 1 of AMARETTO, we identify the main cancer driver genes based on a linear model of the relationship between copy number data or DNA methylation data and disease-specific gene expression data. We observed that both copy number and DNA methylation are more explanatory of disease-specific gene expression compared with raw gene expression. This suggests that using normal tissue variation as a reference improves the signal and potential integration of copy number or DNA methylation data with gene expression. Next, AMARETTO’s second step connects candidate cancer driver genes with their downstream targets by constructing a module network. We discuss several cancer driver genes identified by our model. We then associate the module network with the TCGA identified molecular subtypes in ovarian cancer to identify the main cancer driver genes. Finally, we correlate the module network with overall survival and therapy response, and found some promising associations.

2. Material and methods

2.1. AMARETTO

We developed a method called AMARETTO to identify cancer drivers by using data integration of omics data that considers the differential expression between cancer and normal samples. AMARETTO accomplishes this using a multi-step algorithm that integrates copy number, DNA methylation and gene expression data to identify cancer driver genes and subsequently associates them with their downstream targets through module network analysis. Figure 1 gives an overview of how each step is linked to each other and which algorithms are used for each step.

— Step 1 attempts to identify cancer driver genes by modelling the relationship between genomic and transcriptomic data on an individual gene basis. In addition, we add two important filters to this model. First, we take gene expression variation in normal tissue into account and focus on disease-specific
We recently developed a method called MethylMix to identify methylated. We combined both GISTIC and MethylMix to identify methylation and identifies which genes are hyper- and hypo-major methylation states, compares each state with normal DNA methylation with this paper. MethylMix uses a mixture model to identify the from aberrations found in cancer [9], a summary of MethylMix methylation states by differentiating normal DNA methylation We recently developed a method called MethylMix to identify statistically model DNA methylation aberrations in cancer. general methods are available to identify recurrent amplifications and number data has been extensively studied in this context, and several filters in the following paragraphs.

**2.1. Step 1: generating a list of candidate cancer driver genes**

To generate a list of candidate cancer driver genes, we used a linear regression model to estimate the effect of copy number and DNA methylation on gene expression. This step focuses on modelling the cis-regulatory effects of genomic data on gene expression. This linear regression model was built for each gene independently. Next, we evaluated whether copy number data had a significant positive effect on gene expression and whether DNA methylation had a significant effect on gene expression. These effects were quantified using the $R^2$ statistic.

**2.2. Data preprocessing**

We used gene expression, copy number and DNA methylation data from TCGA ovarian cancer [7]. The gene expression data was produced using Agilent G4502A microarrays. Preprocessing was done by log-transformation and quantile normalization of the arrays. Next, we used DNA methylation data generated using the Illumina Infinium Human Methylation 27 Bead Chip. We used the level-3 methylation data containing methylation data on 27 578 CpG sites in 14 473 genes. DNA methylation is quantified using $\beta$-values ranging from 0 to 1, with values close to 0 versus 1 indicating low versus high levels of DNA methylation, respectively. We removed CpG sites with more than 30 per cent of missing values in all samples, reducing the number of usable CpG sites to 23 030 for ovarian cancer. We used 15-K nearest neighbour algorithm to estimate the remaining missing values in the dataset [14]. Next, we used copy number data produced by the Agilent Sure Print G3 Human CGH Microarray Kit 1 M x 1 M 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 HGNC (version August 2012). Owing to the size of the Q1 TCGA project, the TCGA samples are analysed in batches and a significant batch effect was observed based on a one-way analysis of variance in most data modes. We applied Combat to adjust for these effects [15]. In total, we used 511 primary tumour and eight normal fallopian tube samples for ovarian cancer [7]. These normal samples were profiled using the same TCGA pipeline and platform. In addition, fallopian tube tissue has been shown as a cell or origin for serous ovarian cancer [16,17]. All TCGA data is accessible at the TCGA data portal (https://tcga-data.ncc.nih.gov/tcga/).

**2.3. Gene set enrichment analysis**

To evaluate the enrichment of modules with gene sets, we used several databases, namely: MSigDB v. 3 [18], GeneSetDB v. 4 [19], CHEA for CHIP-X gene sets v. 2 [20] and manually curated gene sets related to stem cells and immune gene sets. We use a hyper-geometric test to check for enrichment of gene sets in the lists of hyper- and hypo-methylated genes. We corrected for multiple hypothesis testing using the false discovery rate (FDR; [21]).

**2.4. Survival analysis**

We used Cox proportional hazards modelling to investigate univariate relationships between modules and overall survival
observed a tendency towards higher $R^2$ values for cancer-specific genes after correction for normal gene expression data ($p$-value 0.0652 Wilcoxon rank sum test).

Regarding STAT3, for example, we observed an increase of 14 per cent in the $R^2$ value when explaining its disease-specific gene expression (36% versus 22% for the disease-specific model versus the model not adjusted for normal variation) solely based on its copy number. Similarly, TNFRSF1A shows an increase of 11 per cent for the $R^2$ in the disease-specific model versus the model not adjusted for normal. Both genes have functions related to cell growth and apoptosis, which represent major processes that are deregulated in cancer. Next, we looked at examples of a decreasing $R^2$ in the disease-specific gene expression. This corresponds to genes where adjusting for normal variation reduces the correlation between genomic markers and gene expression. We found, for example, LY86 shows a decrease of 13 per cent indicating that normal gene expression variation is confounding the correlation between genomic data and gene expression data, and after correcting for normal gene expression variation, the genomic data explain less of the disease-specific gene expression.

3.2. Recurrent genomic or epigenomic events

In addition to focusing on disease-specific gene expression, we select genes that are recurrently altered genomically or epigenomically. In the case of copy number data, this entails that we look for genes that are recurrently amplified or deleted. Copy number data has been extensively studied and several methods are available to identify recurrent amplifications and deletions [11,12]. We used GISTIC to identify recurrent amplifications and deletions across 481 TCGA ovarian cancer samples. GISTIC identified 670 and 2353 genes that are in recurrently amplified and deleted regions, respectively, consistent with earlier reports that ovarian cancer has a large number of copy number alterations.

In the case of DNA methylation, we focused on genes that are significantly and differentially methylated in a subset of patients. To accomplish this, we developed an algorithm called MethylMix that uses a mixture model to identify subgroups of patients with similar DNA methylation for a
3.3. Cancer driver genes associated with their downstream targets

The candidate driver genes still constitute a large set, even after selecting for significant relationships between genomic and transcriptomic data, and recurrent events. The second step of AMARETTO associates candidate driver genes with their downstream targets. This provides insight into the processes that candidate driver genes are regulating, and also serves to focus only on driver genes that are predictive of downstream gene expression.

We selected the top half of the genes that exhibited the highest variance to build the module network, producing 8907 genes on 560 samples. This variance based filter was also applied to the candidate driver genes further refining the list to 865 candidate driver genes. We built a network consisting of 100 modules with corresponding regulatory programmes (see figure 3 and the electronic supplementary material, figure S1) that are functionally enriched for key processes in cancer such as cell cycle, immune response, RNA regulation and extracellular matrix signalling. Electronic supplementary material, table S1 contains the modules and regulatory programmes of the complete ovarian cancer module network. This network contains 339 selected cancer driver genes including 213 that are copy number driven and 144 that are DNA methylation driven. Interestingly, a higher proportion of genes selected that are DNA methylation driven compared with the initial number of DNA methylation candidate driver genes. The selected cancer driver genes include well-known genes such as CCNE1, CDKN2A, KRAS, PTEN and RB1 but also genes with unknown functions in cancer such as EVI2A, C1orf114 and LCP2.

A number of biological hypotheses can be deduced from this network. For example, the network suggests that CCL5 is a master immune system response regulator. This gene is significantly hypo-methylated in a subset of samples. CCL5 is one of the top cancer drivers in the network and is part of five regulatory programmes. Each of the corresponding modules is highly enriched with gene sets related to immune response or defence response. For example, module 37 is highly enriched with immune response genes and is also significantly increased in the immunoreactive molecular subtype (figure 4, p-value $1.2009 \times 10^{-29}$ Wilcoxon rank sum test; [7]). CCL5 is a chemokine that facilitates disease progression by recruiting and modulating the activity of inflammatory cells, which subsequently remodel the tumour microenvironment. Moreover, CCL5 has been shown to promote metastasis in basal breast cancer cells [24].

Next, we analysed how frequently certain cancer drivers co-occur in a regulatory programme. This analysis showed that NUAK1 and PCOLCE always co-occur and are part of four regulatory programmes. NUAK1 is known to directly phosphorylate TP53 and regulate cell proliferation. PCOLCE is a pro-collagen and is active in the extracellular matrix. All four NUAK and PCOLCE modules are highly correlated with the mesenchymal molecular subtype suggesting that both genes are major drivers of this ovarian cancer subtype. For example, Module 32 is regulated by NUAK1 and PCOLCE and is significantly upregulated in the mesenchymal subtype (figure 4, Wilcoxon rank sum test, p-value $1.3125 \times 10^{-31}$). It is interesting to note that both NUAK1 and PCOLCE are aberrantly methylated genes.
Similarly, CHEK1 and FBXO5 co-regulate two modules: module 55 and module 16. These modules are the top modules enriched for cell proliferation and cell cycle. This suggests that both genes are disease-specific drivers of cell proliferation in ovarian cancer. CHEK1 is a well-known cell cycle gene required for checkpoint-mediated cell cycle arrest in response to DNA damage. FBXO5 is a well-known regulator of the mitotic cell cycle. Both CHEK1 and FBXO5 are recurrently deleted genes.

Finally, EVI2A is among the top regulators in the network. EVI2A is a membrane protein that has unknown function. The predicted targets of EVI2A are highly enriched

---

**Figure 4.** Visualization of (a) module 37 and (b) module 32 regulatory programmes and expression profiles. (A) Visualization of the regulatory programmes. (B) Comparison of the accuracy of the estimated mean expression of the module based on the regulatory programme with the average module expression. (C) Expression of the genes in the module. (D) Indication of membership of samples to the immunoreactive (left panel) and mesenchymal (right panel) molecular subtype. (E) Box-plot comparison of the gene expression in the immunoreactive (left panel) and mesenchymal (right panel) versus the remaining samples.
in the KEGG pathways, ECM receptor interaction and focal adhesion, indicating that this gene may have a role in interactions between cancer cells and the extracellular matrix. EVI2A is a methylation-driven gene expressed in neutrophils and leads us to speculate that we may be measuring the infiltration of neutrophils in ovarian cancer cells which may reflect an immune-mediated driver of oncogenesis.

3.4. Modules that accurately predict molecular subtypes

Next, we focused on how modules can be combined to predict the ovarian cancer molecular subtypes [7]. This allows further identification of the cancer drivers of each of the molecular subtypes. We used linear regression with lasso regularization to build a model for each of the four molecular subtypes based on module expression. The performance was estimated using the area under the ROC curve (AUC) in a 10-fold cross validation loop (figure 5).

All four subtypes can be accurately predicted with AUCs ranging from 0.857 to 0.975. The mesenchymal subtype is most accurately predicted. This model is based on only one module: module 48. Module 48 is one of the four modules regulated by NUAK1 and PCOLCE. Functionally, this module is highly enriched in extracellular matrix genes and focal adhesion (see figure 3 and electronic supplementary material, figure S1). Module 48 is also significantly enriched in genes upregulated in ovarian cancer metastasis versus primary tumours [25]. This is consistent with the mesenchymal subtype. Similarly, the proliferative subtype is also predicted based on only one module: module 34. This module has limited functional enrichment in genes related to protein kinase activity but is otherwise not well characterized. The other two subtypes need a n average 2.9 modules for the differentiated subtype and 2.5 modules for the immunoreactive subtype.

3.5. Correlation of modules with ovarian cancer outcome data

We investigated whether the modules can be used to predict therapy response and survival. We used Cox proportional hazards modelling to correlate the module expression with overall survival. We found that eight modules were significantly correlated with overall survival (table 1, Wald test < 0.05 at FDR < 15%).

We compared patients with good survival versus poor survival to investigate if analysis of the extreme cases results in more significant results. More specifically, we compared patients who had no recurrence or progression for at least 1000 days versus patients who have treatment refractory disease. We found that four modules are significantly correlated with this outcome (Wilcoxon rank sum test < 0.01, FDR < 15%). Module 16 in particular was significantly correlated with this outcome (p-value 8.2757 × 10⁻⁶, FDR 8.2757 × 10⁻⁶). Module 16 is functionally enriched for cell cycle and DNA repair, and by the co-regulators CHEK1 and FBXO5 (figure 6). Expression of module 16 is highly upregulated for good versus poor survival. This is consistent with the copy number data of CHEK1 and FBXO5, both recurrently deleted, corresponding to a tumour suppressor function. Thus patients with intact CHEK1 and FBXO5 expression have better prognosis, consistent with these genes’ function in DNA repair or cell cycle control. In addition, supervised modelling using 10-fold cross validation of this outcome resulted in an accurate model using only module 16 expression (AUC 0.83, figure 6).

Finally, we investigated whether any of the modules are correlated with drug response. More specifically, we looked at sensitivity (time to failure greater than 365 days) or resistance (time to failure less than 365 days) to platinum-based chemotherapy. We found that three modules were significantly correlated with this outcome (p-value < 0.01, FDR < 15%).

Figure 5. ROC curves of supervised classification of the ovarian cancer molecular subtypes using module expression for each of the four subtypes. AUC, area under the ROC curve; s.e., standard error. (Online version in colour.)
Module 86 was highly correlated with platinum sensitivity ($p$-value 0.00046015, FDR 4.6%). Module 86 is functionally enriched with the Notch signalling pathway, which has been tied to platinum-based chemotherapy response [26].

4. Discussion

AMARETTO is an analytical approach that aims to address the challenges associated with integrating and interpreting multi-omics cancer datasets, for example, through the TCGA project. TCGA now has over 20 cancers that are being studied extensively with multiple omics technologies. This clearly creates a need for methods such as AMARETTO to extract knowledge that leverages all the data.

AMARETTO identifies cancer driver genes by considering that genes that are recurrently altered at the genome or epigenome level with functional consequences, as measured by their gene expression, are the most likely candidates. In addition, AMARETTO takes into account only disease-specific expression variations. This eliminates genes that are naturally expressed in normal tissue and are most likely not cancer drivers. Finally, AMARETTO only focuses on cancer drivers that explain downstream gene expression in the form of modules.

AMARETTO is being continuously developed and improved. Future plans involve integrating microRNA and DNA sequencing data into our models. We investigated the integration of microRNAs in the regulatory programmes. However, our results suggest that microRNAs do not explain additional expression variation at the module level compared with cancer driver genes based on our observation that they are not selected in the regulatory programmes. This is most likely caused by the observation that many microRNAs are located in introns of so-called host genes, confounding gene expression. Accordingly, microRNAs do not seem to explain more expression variation compared with candidate cancer drivers. We also investigated DNA mutation data from sequencing technology, but two issues emerged. First, a mutation does not necessarily have to affect gene expression. For example, mutations that constitutively activate protein function may not result in increased expression, but in increased function. Secondly, mutation data is notoriously sparse thereby limiting our ability to find statistically significant relationships between DNA mutations and gene expression. Further work is necessary on the most optimal way of integrating microRNAs, mutations and other omics data in AMARETTO.

The main limitation of methods such as AMARETTO is the difficulty to evaluate the resulting network models. The most accepted route is experimental validation, a laborious and time-consuming process. In silico validation by comparison with other modelling strategies or networks is a more efficient validation strategy. However, the lack of standardized databases that store computational network models of cancer further complicate this level validation. Most models are now hidden in supplementary files in non-standardized formats. A community effort in this area could create a comprehensive resource of computational models that can serve as a resource for experimental biologists.

In summary, we developed AMARETTO as a biocomputational approach for integrating multi-dimensional cancer data in a manner that enables the identification and analysis of genomic and epigenomic features that influence disease-specific gene expression. Using this method, we identified several novel oncogenic drivers in ovarian cancer and...
developed several new biological and clinical hypotheses. We identified potential drivers of the mesenchymal and proliferative ovarian cancer subtypes. Finally, we identified modules predictive of good versus poor outcome which implicate DNA repair pathways as a marker of ovarian cancer outcome as well as response to platinum therapy.

Figure 6. Visualization of module 16, significantly correlated with very good versus poor outcome in ovarian cancer. (a) Regulatory programme of module 16, (b) Estimated expression of module 16 using the regulatory programme versus actual mean expression, (c) Module 16 gene expression (d) patients with very good outcome are indicated in black versus poor outcome patients in white. (e) Box plot of the module 16 expression in very good versus poor outcome patients. (f) ROC curve of the supervised model based on module 16 to predict the very good versus poor outcome.
References


