Systematic Use of Computational Methods Allows Stratifying Treatment Responders in Glioblastoma Multiforme

Riku Louhimo*, Viljami Aittomäki*,+, Ali Faisal**,+, Marko Laakso*,+, Ping Chen*, Kristian Ovaska*, Erkka Valo*, Leo Lahti**, Vladimir Rogojin*, Samuel Kaski**,***, Sampsa Hautaniemi*

*Computational Systems Biology Laboratory, Genome-scale Biology Research Program, University of Helsinki, Finland; **Aalto University School of Science, Helsinki Institute of Information Technology HIIT, Finland; ***Department of Computer Science, University of Helsinki, Finland; +Equal contribution.

ABSTRACT

Cancers are complex diseases whose comprehensive characterization requires genome-scale molecular data at several levels from genetics to transcriptomics and clinical data. We use our recently published Anduril framework and introduce novel approaches, such as dependency analysis, to identify key variables at miRNA, copy number variation, expression, methylation and pathway level in glioblastoma multiforme (GBM) progression and drug resistance. We also present methods to identify characteristics of clinically relevant subgroups, such as patients treated with temozolomide drug and patients with an EGFRvIII mutation, which is a constitutively active variant of EGFR. Our results identify several novel genomic regions and transcript profiles that may contribute to GBM progression and drug resistance. All results and Anduril scripts are available at http://csbi.ltdk.helsinki.fi/camda/.

1. Introduction

Glioblastoma multiforme (GBM) is the most frequent and aggressive brain tumor type with incidence of 2-3 cases in 100,000 people per year. Over the past 25 years the advances in GBM treatment have been very modest and the median survival of a GBM patient has remained 15 months [7]. To improve diagnosis and treatment of GBM, The Cancer Genome Atlas (TCGA) consortium provides hundreds of GBM primary tumors with high-throughput molecular data at genetics, transcriptomics and epigenetics levels together with clinical data [15]. These data provide a basis for gaining a holistic view on GBM progression and drug resistance.

Integration of such massive amounts of data requires a computational infrastructure that allows systematic data processing and interpretation. We recently introduced a computational platform, Anduril, that facilitates analysis and integration of large-scale data, systematic software development and rapid use of bio-databases [14]. Anduril provides a framework for joining reusable algorithms (components) into executable workflows. A component can be implemented with any programming language, such as R, MATLAB, Java and C++, which allows us to take advantage of the efforts of the bioinformatics community as well as parallelize computationally demanding tasks.

An underlying assumption in using all TCGA GBM samples to identify survival associated genomic regions or transcript profiles is that the samples belong to the same tumor subtype and have been treated similarly. In reality these assumptions are oversimplifications as shown in a study by Verhaak et al. in which four subtypes of GBM were suggested based on gene expression profiles [16]. Furthermore, the GBM patients in TCGA are treated using a wide spectrum of drugs, and while some patients have received only chemotherapy, others have been administered up to 15 different compounds. Here our major objective is to identify genomic regions and transcript profiles that have a significant survival or drug response association for a subset of the GBM samples. Furthermore, to deal with noise and uncertainty in the heterogeneous cancer data we use a Bayesian dependency approach that incorporates a suitable prior, which is well-suited for multi-source analysis.

2. Results

2.1 Automated TCGA data import

The first step in the analysis of TCGA data is to fetch it. As TCGA repository is frequently updated, fetching the data needs to be done automatically and periodically. To this end we

implemented an Anduril component (GetFromTcga) that automatically imports and integrates data from the TCGA data portal into Anduril workflows. GetFromTcga automatically generates file and sample reference tables and reports that establish relations between the data files and the samples that those files contain. The latest versions of data are imported by default, and a user can also specify any earlier version of the data. In this way, GetFromTcga enables fine-grained selection of data to be imported, as well as automated data download and update functionality. Since GetFromTcga is incorporated into our workflow, our analysis always contains the most up-to-date version of the TCGA data. Here all data were accessed on May 6th 2011. The data processing protocols are available at http://csbi.ltdk.helsinki.fi/camda/.

2.2 Biomarker candidate search for temozolomide treated GBM patients

GBM patients with a methylated promoter of the *MGMT* DNA-repair gene treated with temozolomide adjuvant therapy and radiotherapy have a longer median survival of 15 to 21 months [5]. Though the treatment is not curative and mutations in mismatch repair genes have been shown to override *MGMT* repression in rendering tumors resistant to alkylating agents [13], temozolomide in GBM is a prime example of the power of "personalized medicine", *i.e.,* choosing the therapeutic strategy using a molecular biomarker status of the patient. Here our objective was to establish a workflow that allows for rapid search for candidate biomarkers for a given treatment strategy. We used 76 GBM patients treated with adjuvant temozolomide and checked whether gene or alternative spliced variant expressions, copy number alterations, single nucleotide polymorphisms (SNPs), methylation patterns or miRNA have survival effect with Kaplan-Meier analysis in these patients.

This analysis suggests several interesting genes for further analysis. For instance, chromosome X open reading frame 1 (*CXorf1*) is significantly down-expressed in GBM and associated with survival in both gene and exon platforms (p<0.0004). As another example, Werner syndrome, RecQ helicase-like (*WRN*) has two SNPs that are significantly associated with survival (p<0.0001) as shown in Figure 1. Both of these effects are absent if all patients are included in the analysis. While the number of temozolomide-treated tumors is small, our efforts provide a comprehensive workflow to identify variables that play a key role in temozolomide sensitivity, and provide biomarker candidates for deciding when to use temozolomide.

2.3 Dependency analysis of differentially expressed genes within copy number alteration and methylated regions

Genomic instability is a hallmark of cancer and high-throughput measurements of copy number variation data have become commonplace in cancers. Given that copy number measurements are noisy, one of the most successful approaches in increasing the reliability of putative driver genes involved in tumor progression and drug resistance is integration of copy number data to transcriptomics data. Our objective is to first identify chromosomal regions that have high dependencies between gene expression and copy number changes, and then form patient groups from each of the identified region and a survival analysis to check whether the identified genomic aberrations have survival associations in GBM.

We used our recently developed "similarity constrained" canonical correlation analysis approach (simCCA) as we have observed that the Bayesian formulation with suitable constraints/prior performs better than other learning methods [10]. The method is based on Bayesian formulation of classical canonical correlation analysis [8]. It detects linear dependencies between two data sources by searching for their maximally correlated low-dimensional representation. Briefly, the model defines a chromosomal region via a window that is centered at a gene and spans across ten neighboring genes within the chromosomal arm. The window is slid across all chromosomal arms and a dependency score and each patient's contribution towards the score for each region are calculated. A high score reveals a correlating expression and corresponding chromosomal change; high-scoring regions with $q < 0.05$ were selected for further analysis. For each identified region, patient-wise contribution scores were ordered and three groups were formed based on the 10th percentile, the 90th percentile and the rest.

Our analysis identified three significant chromosomal regions with a stringent cut-off ($q <$ 0.05): 10p13, 10q22.1, 10q26.13. Many corresponding genes had expression profiles that correlated with copy number aberrations such as *HK1*, *HKDC1*, *MCM10*, *DDX21*, and *SLC29A3*.

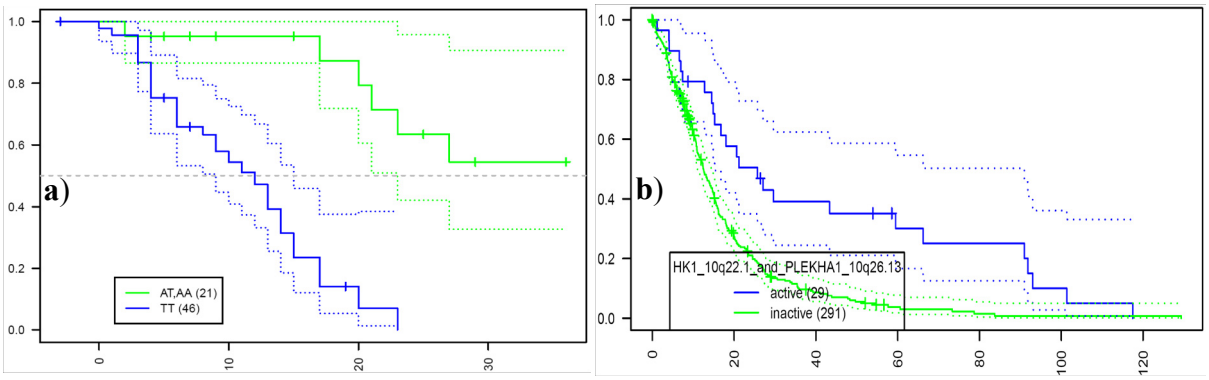

Figure 1. a) Kaplan-Meier analysis of rs11574311 in *WRN* for 67 temozolomide treated patients that were also genotyped. Heterozygous or rare homozygote patients (AA/AT) have significantly better survival than patients with the wildtype homozygote (TT). *WRN* has been suggested to be involved in DNA damage repair, which makes it an intereting candidate for furher studies. b) Kaplan-Meier analysis of a combination of two regions centered at the genes *HK1* and *PLEKHA1*. Patients with a high dependency between copy number alteration and gene expression in all these regions ("active") have better survival association than patients having low dependency ("inactive"). X-axis: months; y-axis: percentage of GBM patients alive. Dotted lines: 95% confidence intervals.

We further maximize survival associations by merging the found regions. The combination of regions centered at *HK1* and *PLEKHA1* resulted in the highest significance ($q < 0.0008$).

In addition to copy number alterations, gene expression levels are affected by methylation patterns. Thus, we integrated also methylation data using the simCCA method and this integrative analysis revealed two statistically significant regions: 10p13 and 12q14.1. The survival associations of all significant regions were checked by stratifying to all individual clinical subgroups and temozolomide treated patients. Interestingly, the survival analyses revealed that *OPTN* and *MCM10* are associated with significant survival increase in white GBM patients less than 30 years old.

2.4 Identifying microRNAs and their target genes with survival association

MicroRNAs (miRNAs) are short non-coding RNAs that typically negatively regulate gene expression and alterations in miRNA expressions are frequently associated with human cancers [2]. In GBM miRNAs play a key role in many hallmarks of glioblastoma, including cell proliferation, invasion, glioma stem cell behavior, and angiogenesis [11].

Here, we integrated 437 GBM patients' gene expression data from exon arrays and 309 patients' miRNA expression data to investigate all cancer related regulatory miRNAs and their potential targets. Initial targets of 63 differentially expressed miRNAs were collected from various sources, such as miRBase [4], microRNAorg [1], miRNAMap2 [6] and TargetScan [12], which contain both validated and predicted targets. These targets were further filtered by Pearson correlation and Kaplan-Meier survival analysis.

In total, we obtained 10 candidate miRNAs along with their 56 negatively regulated targets with that were associated with survival ($p<0.05$). Our integration pipeline for miRNA and gene expression is also able to predict potential miRNA regulated genes.

2.5 Identification of EGFRvIII patients

Epidermal growth factor receptor (*EGFR*) is the most frequently amplified and overexpressed gene in GBM and its amplification is a prognostic marker. EGFRvIII is a variant of *EGFR* with genomic deletion of exons 2-7, corresponding to amino acids 6-273 in the extracellular domain [3]. This variant is not capable of binding ligands but is constitutively active and contributes to tumor progression. The EGFRvIII mutation is found in 20-30% of glioblastoma patients, often combined with EGFR amplification [16].

We developed a method to systematically find EGFRvIII mutations in TCGA data based on Affymetrix exon expression arrays. Exon arrays have the advantage of having high probe resolution in exons. They also enable simultaneous profiling of expression and targeted deletions. Probe sets targeting EGFR are divided into two groups based on whether they match exons deleted in EGFRvIII mutations. Probe sets with low median signal are considered outliers and are removed from statistical analysis. After filtering, there are six probe sets in the deletion

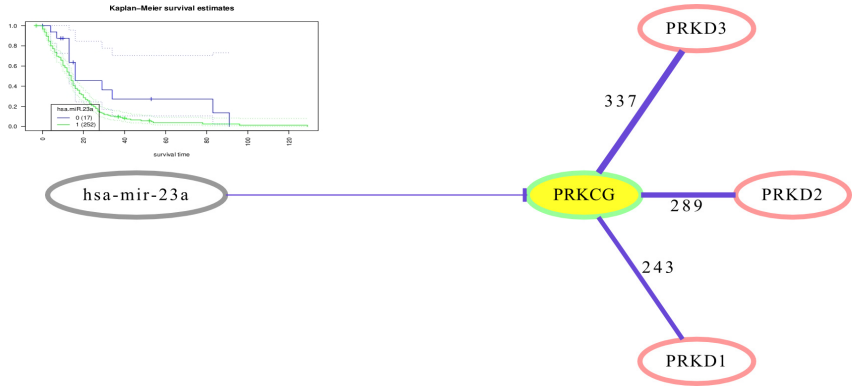

Figure 2. **An example of enzyme-pair analysis.** Upregulated and downregulated genes are in red and green, respectively. The sample frequencies for the enzyme pairs are displayed on edges. Nodes filled with orange or yellow are validated or predicted miRNA targets in databases. hsa-mir-23a has a survival association and *PRKCG* belongs to the KEGG glioma pathway. Here we found three upregulated enzymes from the same protein kinase C group (EC 2.7.11.13).

group and 32 probe sets in the non-deletion group. In EGFRvIII-mutant samples, deletion-group probe sets should have lower signal than other probe sets. For each tumor sample, a one-tailed t-test is applied between these two groups. Samples with a low p-value are considered likely candidates for an EGFRvIII mutation.

Using a p-value threshold of 0.1, 103 out of 397 samples (26%) were classified as potentially having EGFRvIII mutations. In accordance with varied literature on the survival effect of EGFRvIII [16], we did not see a survival difference between candidate EGFRvIII patients and non-EGFRvIII patients (p<0.41).

To find potential drug correlations with EGFRvIII phenotype, we selected the subset of patients (n=196) who have received temozolomide treatment. This drug was selected due to the large number of patients receiving the drug; similar analysis can be conducted for any drug having a suitable number of receiving patients. Temozolomide-treated patients have better survival compared to other patients both in the whole patient set ($p<1.4\times10^{-5}$) as well as in EGFRvIII patients (p<0.051). We compared the survival of EGFRvIII versus non-EGFRvIII patients in the temozolomide-treated patient group. No statistical survival association was found (p<0.17), suggesting that the EGFRvIII mutation does not affect survival in temozolomide-treated patients.

2.6 From context to reactions: Enzyme substitutions in GBM

Enzyme activities are often aberrated in cancer cells and they play a key role in drug resistance. In order to link our efforts to identify genomic regions and transcript profiles having survival effect in GBM to enzymatic reactions, we searched for pairs of enzymes with the same enzyme commission (EC) number, *i.e.*, they catalyze the same reaction, where the same sample has the first enzyme upregulated and the second one downregulated. Our hypothesis is that these enzymes may represent activations of alternative pathways in tumors that mediate tumor progression and drug resistance. Biological properties, such as regulation, binding, efficacy and specificity, typically differ between the original and the substituting enzyme although they share the same EC number.

Enzyme substitutions found from the combined expression profiles were followed up at the sample level. Each gene expression sample was compared against the control population and a fold change limit of two was used to call downregulation and upregulation. The intersection of the common substitution pairs and the DEGs was used to seed *de novo* pathway construction. Pathways were constructed using the Moksiskaan database, which describes known interactions between the genes, proteins, and drugs [9]. The *de novo* pathways were constructed by selecting the genes, drugs, molecular functions, and biological processes connected to the enzyme substitution pairs found from the expression sample profile. Small molecule mediated dependencies were excluded as our primary interested in this study is in signaling cascades. The results reveal a number of interesting substitutions where known GBM related genes have a central role, such as 295 samples with a PYGM -> PYGL phosphorylase transition affecting

the glycogen metabolism and 129 samples with a UPP2 -> UPP1 uridine phosphorylase transitions possibly related to fluorouracil metabolism.

In addition to a global search, we also integrated the identified miRNA-gene pairs (Section 2.4) to the substituted enzymes that were also targets of the survival associated miRNAs. We found several down-regulated enzymes in common enzyme substitutions that are likely to be regulated by miRNAs. For example, hsa-mir-23a has a survival effect and its target *PRKCG* participates a reaction that is associated with GBM progression as shown in Figure 2.

3. Discussion

Multi-dimensional data, such as that provided by TCGA, requires a computational platform, such as Anduril, that allows for inclusion of a large spectrum of computational methods and facilitates collaboration of a team of bioinformaticians. We have presented here several tools to identify key variables defining context for GBM progression and drug resistance, such as probabilistic dependency analysis that provides a flexible and robust approach towards multi-view data integration in functional genomics.

Our results highlight the need for a large number of samples; when the data set is divided into clinically or therapeutically interesting subgroups, the number of samples decreases rapidly, which poses challenges for genome-scale analysis. For example, we introduced a novel approach to identify EGFRvIII variants using exon array data. However, no EGFRvIII patients were treated with EGFR kinase inhibitors, which prohibited characterization of possible biomarkers related to poor efficacy of the EGFR inhibitors in GBM. Accordingly, we showed a proof-of-principle analysis with temozolomide treated EGFRvIII patients, which resulted in no association between constitutive EGFR signaling to temozolomide efficacy.

Our results also highlight the need for advanced algorithms to define context at several levels in order to identify genomic regions or transcript profiles that play a key role in cancer progression and drug resistance. Here we have shown two novel approaches to identify context at miRNA, gene expression, methylation and copy number alteration levels and association to survival of patients treated with temozolomide. We further linked the miRNA results to cell-network level using a novel concept of finding enzyme substitution pairs. For instance, our results suggest that hsa-mir-23a has a survival association (p<0.029) and its expression correlates (r=-0.55) with the *PRKCG* gene (Figure 2). Enzyme substitutes are feasible drug targets as their activity is higher in cancer cells than healthy cells. The inhibition of these gene products may reduce the malignancy of the cells if they have become addicted to reactions mediated by these genes.

In summary, we have demonstrated the benefits of using a systematic computational framework to include algorithms that enable identification of context and clinically important patient subgroups. Our results provide several genes and genomic regions that have survival effect in GBM or a clinically defined subset, such as temozolomide-treated patients, and thus facilitate translation of large-scale biomedical data to knowledge and further to medical benefits.

4. References

[1] Betel et al. The microRNA.org resource: targets and expression. Nucleic Acids Res. 2008 Jan;36(Database issue):D149-53.

[2] Calin GA, Croce CM. MicroRNA signatures in human cancers. Nat Rev Cancer. 2006 Nov;6(11):857-66.

[3] Gan et al. The EGFRvIII variant in glioblastoma multiforme. J Clin Neurosci. 2009 Jun;16(6):748-54.

[4] Griffiths-Jones S et al. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008 Jan;36(Database issue):D154-8

[5] Hegi ME et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med. 2005 Mar 10;352(10):997-1003.

[6] Hsu et al. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. Nucleic Acids Res. 2008 Jan;36(Database issue):D165-9.

[7] Huse JT, Holland EC. Targeting brain cancer: advances in the molecular pathology of malignant glioma and medulloblastoma. Nat Rev Cancer. 2010 May;10(5):319-31.

[8] Klami A, Kaski S. Local Dependent Components. In Zoubin Ghahramani (Ed.), Proceedings of the 24th International Conference on Machine Learning (ICML 2007), pp. 425-433. Omni Press, 2007

[9] Laakso M, Hautaniemi S. Integrative platform to translate gene sets to networks. Bioinformatics. 2010 Jul 15;26(14):1802-3

[10] Lahti L, Myllykangas S, Knuutila S, Kaski S. Dependency detection with similarity constraints. In Proc. MLSP 2009, IEEE International Workshop on Machine Learning for Signal Processing, pages 89–94. IEEE, 2009

[11] Lawler S, Chiocca EA. Emerging functions of microRNAs in glioblastoma. J Neurooncol. 2009 May;92(3):297-306.

[12] Lewis et al. Prediction of mammalian microRNA targets. Cell. 2003 Dec 26;115(7):787-98.

[13] Liu et al. Mismatch repair mutations override alkyltransferase in conferring resistance to temozolomide but not to 1,3-bis(2-chloroethyl)nitrosourea. Cancer Res. 1996 Dec 1;56(23):5375-9.

[14] Ovaska K, Laakso M, Haapa-Paananen S, Louhimo R, Chen P, Aittomäki V, Valo E, Núñez-Fontarnau J, Rantanen V, Karinen S, Nousiainen K, Lahesmaa-Korpinen AM, Miettinen M, Saarinen L, Kohonen P, Wu J, Westermarck J, Hautaniemi S. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. Genome Med. 2010 Sep 7;2(9):65.

[15] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008 Oct 23;455(7216):1061-8.

[16] Verhaak et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010 Jan 19;17(1):98-110.