Molecular Omics

A pan-cancer analysis of progression mechanisms and drug sensitivity in cancer cell lines

Journal:	*Molecular Omics*
Manuscript ID	MO-RES-07-2019-000119.R1
Article Type:	Research Article
Date Submitted by the Author:	31-Aug-2019
Complete List of Authors:	Lima Fleck, Julia; Pontifical Catholic University of Rio de Janeiro, Industrial Engineering Brandusa Pavel, Ana; Boston University, Bioinformatics Program; Icahn School of Medicine at Mount Sinai Cassandras, Christos; Boston University, Division of Systems Engineering and Center for Information and Systems Engineering

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

A pan-cancer analysis of progression mechanisms and drug sensitivity in cancer cell lines

Julia L. Fleck,*[a‡] Ana B. Pavel,[b,c‡] and Christos G. Cassandras[d]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Biomarker discovery involves identifying genetic abnormalities within a tumor. However, one of the main challenges in defining such therapeutic targets is accounting for the molecular heterogeneity of cancer. By integrating somatic mutation and gene expression data from hundreds of heterogeneous cell lines from the Cancer Cell Line Encyclopedia (CCLE), we identify sequences of genetic events that may help explain common patterns of oncogenesis across 22 tumor types, and evaluate the general effect of late-stage mutations on drug sensitivity and resistance mechanisms. Through gene enrichment analysis, we find several cancer-specific and immune pathways that are significantly enriched in each of our three proposed phases of cancer progression. By further analyzing the drug activity area associated with compounds that target the BRAF oncogene, a known predictor of drug sensitivity for several compounds used in cancer treatment, we verify that the acquisition of new driver mutations interferes with the targeted drug mechanism, meaning that cells without late-stage mutations generally respond better to drugs.

1 Introduction

Currently, the main thrust of cancer research is largely based on the concept that causative mutations are responsible for driving a tumor's biological evolution and, as result, its clinical features and response to treatment [1]. In this context, therapeutic decisions must be guided by a tumor's genomic characteristics and, to that end, a number of recent massive-scale efforts have aimed at collecting, organizing and making publicly available multiple data types derived from genetic analysis of cancer cell lines or human tumor samples [2,3,4,5]. Such data are typically collected at one point in time and may aid in uncovering common cancer progression pathways, as well as in classifying cancer patients into groups that will most likely benefit from a certain treatment approach. When additional data types are available, such as solid tumor volume, mathematical models may be developed to forecast tumor growth, evaluate drug efficacy and design rational scheduling of anti-cancerous drugs [6].

In spite of the fact that not all tumors of the same type of cancer possess identical sets of genetic alterations, there seems to be at least a subset of such abnormalities that are consistently verified across a set of tumors. This indicates that different patterns of somatic mutation and gene expression changes may affect cancer initiation and progression mechanisms in a similar manner. In an effort to characterize biological processes and derive biomarkers that indicate disease states or predict medical outcomes, several studies have been conducted using existing cross-sectional data sets. Single data type analyses have been performed using somatic mutation data for inferring the temporal order of genetic alterations [7] as well as for molecular subtyping [8,9]. Copy number variation data and gene-expression levels have also been used independently for deriving causal models of cancer progression [10] and for patient stratification [11].

Due to the fact that one type of data alone may generate an incomplete view of pathway activity, multivariate cancer subtyping has been extensively performed in an effort to uncover genomic instability patterns that could be exploited to inform treatment strategies [12,13,14,15,16,17,18,19,20]. Although several approaches have emerged that correlate different types of cross-sectional data with cancer prognosis, relatively fewer methods have been proposed to *simultaneously* process multiple datasets for biomarker discovery and infer cancer pathways. Moreover, the temporal order of abnormal genomic events is commonly explored in broad time brackets, such as primary vs. metastatic tissues [21], and most attempts at reconstructing tumor progression at the pathway level have thus far considered known, *a priori* defined, pathways.

Novel approaches for simultaneously inferring cancer pathways and the order of genetic mutation occurrence have recently been

[a] *Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro, Rua Marques de Sao Vicente, 225, Rio de Janeiro, Brazil. Fax: 55 21 3527-2181; Tel: 55 21 3527-2167; E-mail: jfleck@puc-rio.br*
[b] *Bioinformatics Program, Boston University, 24 Cummington Mall, Boston, USA.*
[c] *Icahn School of Medicine at Mount Sinai, 1425 Madison Ave, New York, USA.*
[d] *Division of Systems Engineering and Center for Information and Systems Engineering, Boston University, 15 Saint Mary's Street, Brookline, USA.*
‡ These authors contributed equally to this work.

proposed, first using exclusively somatic mutation data[22], and then by combining somatic mutation and gene expression data from cross-sectional measurements[23]. Although these methods are capable of identifying phases of cancer progression that corroborate known interactions between genes in important cancer pathways, to the best of our knowledge, no analysis has yet leveraged such information in light of drug sensitivity and resistance mechanisms. In this paper, we search for common progression and drug sensitivity patterns across different types of cancer. We perform a pan-cancer analysis of cell line data in order to stratify known oncogenes and tumor suppressors[24,25] into a number of phases of cancer progression, and predict their effect on gene expression. We then investigate how these phases may help explain drug response data from the Cancer Cell Line Encyclopedia (CCLE) by evaluating the effect of late-stage mutations on drug resistance.

2 Results

The problem of partitioning somatic mutation and gene expression data into a temporal sequence of events may be formulated as a Mixed Integer Linear Program (MILP)[23]. The assumptions underling our MILP formulation are the following:

(A1) Exclusivity of driver mutations within each cancer progression phase. It has been shown that a typical tumor contains only about two to eight mutations in genes that promote tumorigenesis, the remaining mutations occurring in genes that confer no selective growth advantage[24]; hence we assume that each sample can only have one mutated gene in each phase.

(A2) Progression of mutation across subsequent phases. The notion that cancer accumulates mutations over time is widely accepted[22,26]; hence, we assume that each sample must have one gene mutated in the previous phase in order to have a mutation in a subsequent phase.

(A3) Dependency relationship between mutated genes and genes with abnormal expression. Abnormal gene expression is due to driver gene mutations; if a sample has no mutated genes in a given phase, no changes in gene expression may occur in that phase.

(A4) The strength of the connection between expression and mutation genes determines the assignment of abnormal expression genes to the corresponding phases. This means that each expression gene is assigned to a certain phase based on the strength of this gene's connection to the mutations genes that belong to that phase.

Of note, our MILP formulation derives from the assignment problem, a well established linear programming formulation. In the assignment problem, there are n persons and n projects, and we wish to assign a different person to each project; this is referred to as the exclusivity constraint. In our model, we wish to assign n genes to K phases, where the exclusivity constraint is enforced across each sample, meaning that each sample can only have one mutated gene in each phase.

Assumption *(A1)* adapts the exclusivity constraint in the context of cancer heterogeneity. This constraint reflects current observations that different individuals may harbor driver mutations in different genes within the same pathway[24]. Because driver mutations target pathways, it has been suggested that the order in which mutations (and, more generally, different types of genetic alteration) occur should be analyzed at the pathway level, not at the gene level. Considering that no information regarding known pathway interactions is provided to our MILP model, and in light of recent observations that driver mutations tend to be mutually exclusive within pathways[27], assumption *(A1)* implies that each phase is driven by a mutation defining a "unit of time" of cancer progression, and should not be interpreted at the gene level, but at the pathway level. More importantly, the phases proposed by our algorithm do not identify with clinical staging, meaning that we are not, in fact, suggesting that only one mutation occurs as a tumor progresses from one clinical stage to the next; our "unit of time" is actually defined by a mutation event, and each phase defines a set of equivalent mutations that may occur at the same time during the cancer progression of different patients, explaining the increased heterogeneity observed in advanced clinical stages across different cancer types and even within each cancer type. We also note that our work focuses on driver mutations occurring on driver genes, and no attempt is made to temporally stratify passenger mutations.

Assumptions *(A2)-(A4)* are not present, in the form of constraints, in the assignment problem, but must be used in our MILP formulation, as explained next. Assumption *(A2)* may be understood as a progression constraint. It enforces a linear progression of somatic mutations at the pathway level and, as shown in[22], both exclusivity and progression constraints are necessary in order for our linear programming formulation to generate correct partitions. Assumptions *(A3)* and *(A4)* connect the occurrence of somatic mutation events and changes in gene expression during cancer progression. More specifically, assumption *(A3)* reflects several underlying issues in cancer biology. For one, it accounts for the issue of molecular heterogeneity by considering that distinct mutational events may result in under/over expression of several genes that affect the cell's state in a similar way. Moreover, it illustrates the fact that cancer progression is a multi faceted process involving both the accumulation of mutations as well as changes in gene expression. Finally, it points to a reinforcement mechanism whereby changes in gene expression may lead to the appearance of new mutations. Nevertheless, this assumption does not imply that driver mutations in driver genes are the only types of genetic abnormalities to influence the occurrence of expression changes in cancer-associated genes, and vice-versa.

Here we solved the MILP using CPLEX v12.6 with default parameters for a varying number of phases K such that $K \in \{2,3,4\}$. The most appropriate number of phases, for the CCLE dataset we consider, was then determined by assessing the MILP output based on its biological significance. Figures 1, 2 and 3 illustrate the distribution of mutation and expression genes assigned to each phase for $K = 2$, $K = 3$ and $K = 4$, respectively. Interestingly, although the most appropriate value of K is expected to be dataset-specific, here the most meaningful partition of mutation and expression genes was obtained for $K = 3$, similarly to what was observed in The Cancer Genome Atlas (TCGA) breast cancer dataset[23]. Indeed, when running the MILP with two (Figure 1) or four (Figure 3) phases, most of the expression genes were

grouped in only one phase. Clearly these results do not reflect a gradual progression and hence lack biological significance.

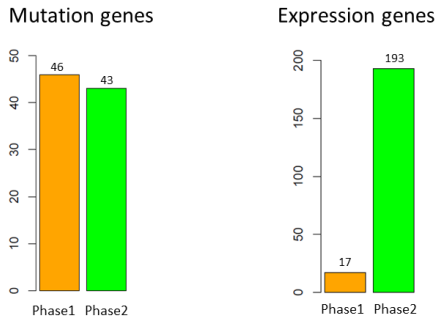

Fig. 1 Distribution of mutation and expression genes for $K = 2$

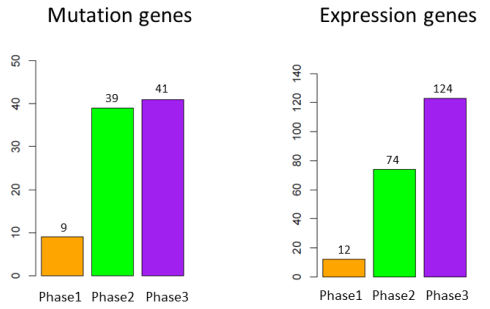

Fig. 2 Distribution of mutation and expression genes for $K = 3$

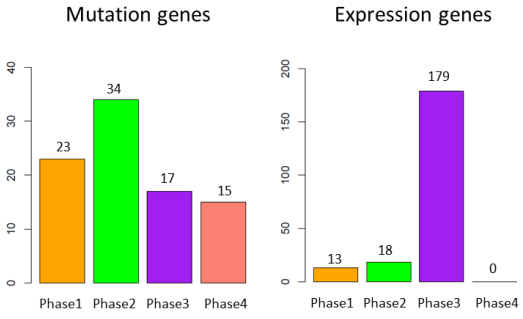

Fig. 3 Distribution of mutation and expression genes for $K = 4$

Subsequent to defining the most meaningful partition of mutation and expression genes, gene enrichment analysis was performed using Enrichr[28]. Our results reveal several cancer-specific and immune pathways that are significantly enriched within each of the three phases of cancer progression. First, although Phase 1 contains the smallest gene set among all phases (21 genes vs. 113 and 165 genes in Phases 2 and 3, respectively), these few genes were found to be significantly enriched ($FDR < 0.05$) in 50 cancer-related pathways, including "Apoptosis" ($FDR < 10^{-8}$), "PI3K-Akt signaling pathway" ($FDR < 10^{-7}$) and "MAPK signaling pathway" ($FDR < 10^{-6}$), "mTOR signaling pathway" ($FDR < 10^{-4}$), "Wnt signaling pathway" ($FDR < 10^{-4}$), "RAS signaling pathway" ($FDR < 10^{-4}$), "TNF signaling pathway" ($FDR = 0.0002$), "Jak-STAT signaling pathway" ($FDR = 0.0005$), "P53 signaling pathway" ($FDR = 0.002$), "Notch signaling pathway" ($FDR = 0.046$), suggesting that alterations in these key signaling pathways begin early on in oncogenesis. Interestingly, these pathways continue to acquire important alterations during cancer progression as they are also significantly enriched in both Phases 2 and 3. Second, additional cancer pathways which are not significant in Phase 1 achieve significance in Phases 2 and 3, such as "VEGF signaling pathway" ($FDR < 10^{-11}$), "Gap junction" ($FDR < 10^{-14}$) and "Inflammatory mediator regulation of TRP channels" ($FDR < 10^{-14}$). Our MILP-based stratification revealed that while the "Apoptosis" pathway is significantly enriched in all phases of cancer progression, different mutation mechanisms are in play in each phase (Figure 4). Finally, mechanisms of "Proliferation" pathway are particularly present in Phase 3, suggesting that abnormal functioning of this process occurs later on in oncogenesis. These results not only support the widely accepted notion that a continuous accumulation of genomic alterations in cancer signaling pathways occurs during cancer progression, but also explain resistance mechanisms to specific inhibitors in later phases of progression.

In order to further investigate sensitivity vs. resistance mechanisms, we looked into the BRAF oncogene, a known predictor of drug sensitivity for several compounds used in cancer treatment, such as AZD6244, PD-0325901, PLX4720, RAF265[2]. The results reported next represent an initial study focusing on only one predictor of drug sensitivity, and are intended to set the stage for an extended analysis of other known predictors.

Our MILP stratified BRAF mutation in Phase 1, and this fact may explain the heterogeneity of sensitive vs. resistant phenotypes of cells that harbor such mutation[2,29]. To test this hypothesis, we analyzed the drug activity area associated with the aforementioned compounds, whose action mechanism involves targeting BRAF mutant cells. The activity area is a measure of cell growth inhibition relative to drug concentration[2]. In this context, larger values of activity area for a given drug indicate greater potency and efficacy. To verify whether our proposed 3 phase progression pattern may help explain drug response, cell lines were categorized into three groups: G_1 consisting of samples with BRAF mutation in Phase 1 and few mutations (≤ 3, arbitrary threshold) in genes that belong to Phase 2 or 3; G_2 including samples with BRAF mutation in Phase 1, many mutations in Phase 2 (> 3), and few mutations in Phase 3 (≤ 3); G_3 containing samples with BRAF mutation in Phase 1 and many mutations in both Phase 2 and 3 (> 3 mutations in Phase 2 and > 3 mutations in Phase 3). Of note, the threshold value of 3 was arbitrarily selected in light of previous studies suggesting that a typical tumor contains no more than eight driver gene mutations[24]. Moreover, note that exceeding a threshold of 3 implies that (at least) more than $1/3$ of a given sample's driver gene mutations have occurred in a certain phase.

In this context, the activity areas of cell lines were compared across the three categories. As shown in Figures 5, 6 and 7, the mean activity area decreases gradually from G_1 to G_3 for all four

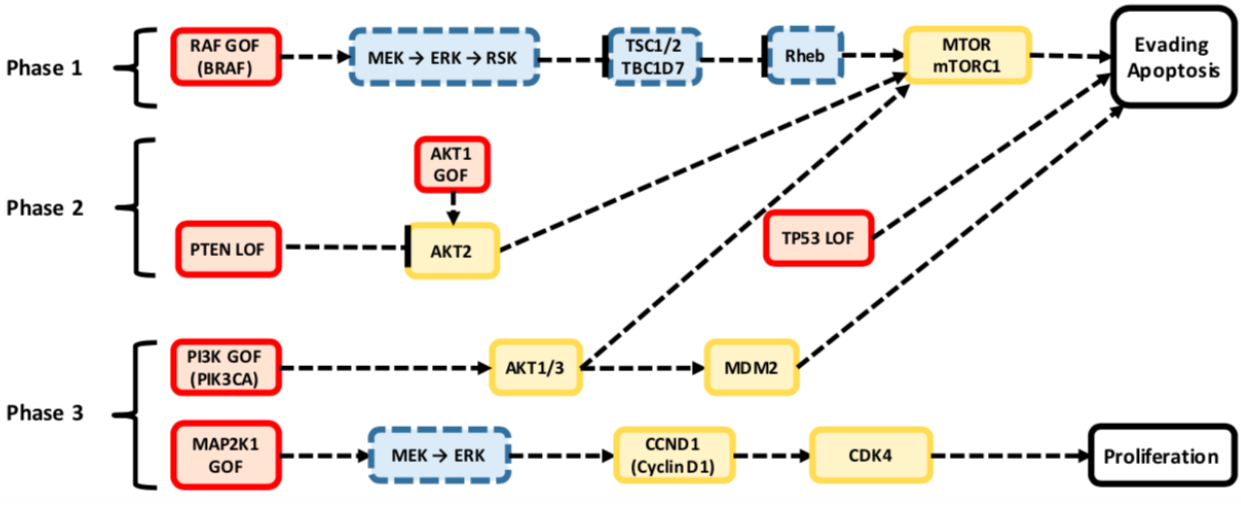

Fig. 4 MILP model identifies temporal relationships in "Evading Apoptosis" and "Proliferation" pathways. GOF: Gain of Function Mutation (in the context of oncogenes); LOF: Loss of Function Mutation (in the context of tumor suppressors). Red indicates GOF/LOF mutation; Yellow indicates expression changes; Blue indicates phosphorylation or other activation processes not included in the MILP model.

tested compounds, reaching significance between groups G_1 and G_3 for PLX4720 ($p-value = 0.01$), AZD6244 ($p-value = 0.02$) and PD-0325901 ($p-value = 0.04$). These results suggest that the acquisition of new cancer driver mutations interferes with the targeted drug mechanism, in our case BRAF mutation. This means that cell lines harboring patterns of genomic alterations similar to those verified in Phases 2 and 3 tend to have a decreased response to targeted therapy.

We also compared our results with random assignments of mutation genes. Using the same number of genes per phase, as determined by our MILP, we shuffled them randomly across the 3 phases. As shown in Supplementary Figure 1, results no longer show a gradual decrease in drug sensitivity with progression. This provides significant evidence that the stratification obtained by our model is meaningful and captures the sequence of molecular changes with cancer progression. Furthermore, we tested a previous approach that applies an Integer Linear Program (ILP) only to mutation data in order to define a temporal sequence of events[22]. Similarly to our approach, the ILP also shows a significant decrease in drug sensitivity with phase for all 3 drugs (Supplementary Figure 2), producing a meaningful stratification of mutations. Our MILP, however, proposes a configuration of both expression changes and mutation events over time, maintaining a meaningful stratification of mutation events and providing additional information about the relationships between gene expression and mutations during cancer progression.

The above results lend themselves to an additional discussion regarding the relationship between mutation and expression genes across the proposed phases of cancer progression. First, it is interesting to note that our MILP stratified expression gene mTOR in Phase 1. The mTOR signaling pathway is known to regulate the cell cycle, including proliferation and cell survival. It is also known that oncogene BRAF is situated upstream of the mTOR pathway (KEGG Pathways in Cancer,[30]). Our MILP results, which place BRAF and mTOR in the mutation and expression gene sets, respectively, of Phase 1, are thus indicative of a dependency relationship between mutations in gene BRAF and

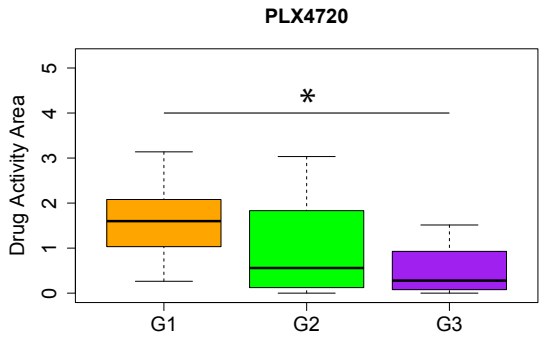

Fig. 5 Drug sensitivity across the 3 progression phases for compound PLX4720. PLX4720 sensitivity significantly decreases in Phase 3 compared to Phase 1 ($p-value = 0.01$).

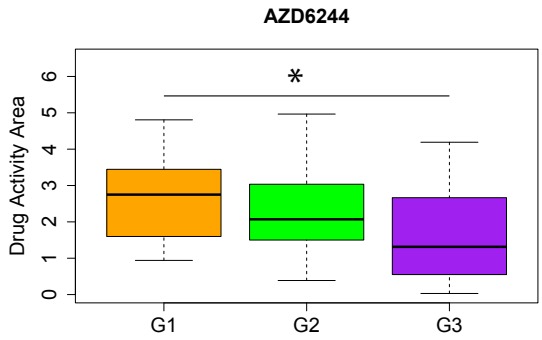

Fig. 6 Drug sensitivity across the 3 progression phases for compound AZD6244. AZD6244 sensitivity significantly decreases in Phase 3 compared to Phase 1 ($p-value = 0.02$).

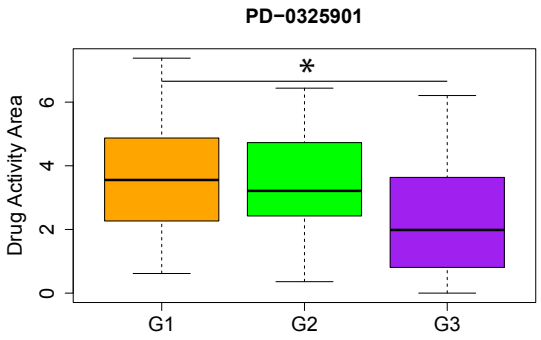

Fig. 7 Drug sensitivity across the 3 progression phases for compound PD-0325901. PD-0325901 sensitivity significantly decreases in Phase 3 compared to Phase 1 ($p-value=0.04$).

expression changes in gene mTOR in early stages of cancer.

Other interesting relationships between mutation and gene expression changes were identified in Phase 2. For example, RAS, PTEN and EGFR are cancer drivers known to interact with the PI3K/AKT signaling pathway (KEGG Pathways in Cancer). Interestingly, AKT2 was stratified in the expression gene set of Phase 2, at the same time as RAS, PTEN and EGFR genes were assigned to the mutation gene set of Phase 2. These findings, which are in agreement with existing knowledge, also suggest that mutations in RAS, PTEN and EGFR genes cause abnormal expression of gene AKT2.

3 Methods

3.1 Mixed Integer Linear Program

The problem of partitioning somatic mutation and gene expression data into a temporal sequence of events is formulated as a Mixed Integer Linear Program (MILP). Details of the problem formulation are given elsewhere[23], but a condensed description of the MILP modeling framework is included here so as to make this paper as self-contained as possible. A script implementing the MILP formulation is available at https://github.com/anabrandusa/MILP, but we note that running it requires access to CPLEX, a commercial mathematical programming solver. Alternatively, the code may be adapted so that other solvers are used to process the MILP model. A simplifying assumption is made by considering that mutations at different genomic loci within a given gene set have a similar effect on tumor biology, similarly to previous works[23,20].

A *phase* of cancer progression is defined in terms of two sets of genes: a set of mutation genes and a set of expression genes. We also define a *sample* as the cancer patient (or cell line) from which genetic data are collected. In this paper, we deal with two types of genetic data: somatic mutation data and gene expression data. The former is input to the MILP as an $m \times n$ binary mutation matrix M, while the latter is input as an $m \times r$ expression matrix E, where m is the number of samples in our database, n is the number of mutation genes considered in our study, and r is the number of expression genes considered in our study.

We further define the *connectivity* between mutation gene j and expression gene h to be the product between the mutation status of gene j and the expression level of gene h, compounded across all samples. Hence, we construct an $r \times n$ real-valued connectivity matrix $C \equiv E^T \cdot M$. The value of entry C_{hj} of the connectivity matrix can be interpreted as follows: values closer to zero indicate that most samples exhibit small absolute values of expression levels for gene h and/or have no mutation in gene j; conversely, the further away the value of C_{hj} is from zero, the stronger is the connectivity between expression gene h and mutation gene j across the data set.

We thus formulate the problem of inferring a model of cancer progression as the search for a partition of the n columns of matrix M into K mutation phases and of the r columns of matrix E into K expression phases. The value of K is externally selected, depending on the desired number of phases, and reflects the level of abstraction of the model. We note that, for the problem we consider here, it is not reasonable to assume that a microscopic model (large value of K) is necessarily superior to a macroscopic model. Moreover, the most appropriate value of K will most likely be dataset-specific. As a result, the number of phases is chosen so as to yield the most biologically meaningful results.

3.2 Data

Our MILP formulation was applied to cell line data made available through the Cancer Cell Line Encyclopedia (CCLE), a collection of gene expression, chromosomal copy number, and massively parallel sequencing data from hundreds of human cancer cell lines. The mutational status of >1,600 genes was assessed by targeted massively parallel sequencing, followed by removal of variants likely to constitute germline events[2]. For our analysis, we considered 629 cell lines that were profiled for both mutation and gene expression data, including 22 tumor types.

Given that many somatic mutations are passengers (i.e., they do not impact cancer progression), it is reasonable to narrow down the >1,600 mutation gene set by selecting those genes that are most likely drivers. The mutation gene subset we consider in this paper consists of 89 known cancer driver genes previously classified as oncogenes or tumor suppressor genes based on the frequency of their mutations[24].

Recall that two sets of genes are input to our MILP: genes that present driver mutations and genes implicated in cancer development. The former, whose selection was detailed above, constitute matrix M. The latter, which make up matrix E, were chosen by overlapping the KEGG Pathways in Cancer set from the Kyoto Encyclopedia of Genes and Genomes database (KEGG)[31,32] with our dataset. Based on this criterion, 210 genes were selected.

Although this paper focuses on cell line data, we remind the reader that our approach is general and can be used to infer the sequence of events from cell line as well as human cancer datasets. In fact, we have previously applied it to The Cancer Genome Atlas (TCGA) data and successfully identified phases of cancer progression that corroborate known interactions between genes in important breast cancer pathways[23].

3.3 Gene enrichment analysis

Gene enrichment analysis is a procedure for inferring which pathways are most prominently active within a given set of genes. The main objective of such method is to map omic measurements to gene sets that represent logical groupings of genes, and its main output is a ranked list indicating which genes are most significantly dysregulated between two conditions. Enrichr[28] is tool for gene enrichment analysis that computes the significance of overlap between a given input list of genes and the gene sets in existing gene libraries. For our analysis, three input gene lists were generated, each list consisting of the genes (from both the mutation and expression gene sets) that had been assigned to a given phase of cancer progression. As such, our first input list included the 21 genes assigned to Phase 1 (9 mutation genes and 12 expression genes, as shown in Figure 2), while our second and third lists included 113 genes and 165 genes, respectively. Comprehensive cancer gene libraries such as KEGG typically map information on which genes are known to interact within important cellular pathways. Hence, by detecting overlaps between our input gene lists and those in KEGG, it is possible to infer the order in which key pathways are altered during oncogenesis.

In this context, the most significantly e nriched p athways in each of our proposed phases of cancer progression were analyzed. In all cases, significance w as a ssessed t hrough Fisher's exact test[28], which tests the null hypothesis that no significant overlap exists between our input gene sets and those in KEGG. Multiple hypothesis testing typically leads to high false positive rates, and to address this we used the corrected the p-values by the False Discovery Rate (FDR) approach, reducing the number of false discoveries in tests that lead to significant r esults. A threshold of 0.05 was chosen for the adjusted p-value.

4 Conclusion

We use a Mixed Integer Linear Program (MILP) to identify sequences of genetic events that may help explain common patterns of oncogenesis across 22 cancer types. We then analyze such temporal sequences in light of drug sensitivity data from the Cancer Cell Line Encyclopedia (CCLE). Our methodology detects which genetic abnormalities occur, and more importantly, *the order* in which they take place as cancer progresses. Our premise that only one driver mutation may occur per phase of cancer progression in a patient suggests that each phase is driven by a mutation defining the "unit of time" of cancer progression. However, we stress that this "unit of time" should be analyzed at the pathway level and does not directly map to the clinical stage of cancer. For late-stage tumors, the model reconstructs events that have already taken place, but for early stage patients, the model is predictive in the sense that it suggests which alterations will likely occur as the disease progresses to more advanced stages. We also verify that the efficacy of a targeted therapy may vary depending on the phase of cancer progression with which a given tumor is associated. Taken together, our results indicate that cells without late-stage mutations generally respond better to drugs. This, in turn, implies that tumors in advanced stages may need to undergo a different drug regimen in order to respond adequately to treatment. These regimen would involve not only a varying number of drugs, but also different, and most likely, case-specific dosages. Such findings may be incorporated into methodologies that evaluate the effect of combining different medications or timing therapy periods on the overall effectiveness of the treatment[33,34].

Our approach advances insights into a number of general mechanisms of drug resistance, but limitations exist in our study. For one, curated data on CCLE has been shown to provide representative genetic proxies for primary tumors in many, but not all, cancer types[2]. Our study analyzed all cell lines from CCLE, including the ones exhibiting weaker genomic similarities with primary tumors. Hence, further analysis of molecular correlates of pharmacologic sensitivity *in vivo* would be useful in ascertaining to what extent our results directly translate anticancer drug response mechanisms of human tumor samples. Additionally, in this study we assume that mutations at different genomic loci lead to similar effects on tumor biology and do not consider the effect of amino acid alterations and protein mutational rates. By extending our model to account for such features, as well as for additional data types, such as copy number alterations and noncoding RNAs, a more robust tool would be generated for analyzing the relationship of cancer initiation and progression, and drug resistance mechanisms. Finally, our ongoing work includes performing experiments to test *in vitro* some of the temporal relationships we have found, thus providing lab validation for our computational findings.

5 Conflicts of Interest

There are no conflicts of interest to declare.

6 Acknowledgements

This work was supported in part by the National Council for Scientific a nd T echnological D evelopment (CNPq) u nder grant 428907/2018-0; the Coordination for the Improvement of Higher Education Personnel (CAPES) under Finance Code 001; the Pontifical C atholic U niversity o f R io d e J aneiro; N SF u nder grant DMS-1664644. The authors acknowledge the Broad Institute for generating and making publicly available the data used in this paper.

References

1 M. Gerstung, E. Papaemmanuil, I. Martincorena, L. Bullinger, V. Gaidzik, P. Pashka, M. Heuser, F. Thol, N. Bolli, P. Ganly, A. Ganser, U. McDermott, K. Dohner, R. Schlenk, H. Dohner and P. Campbell, *Nature Genetics*, 2017, **49**, 332–340.

2 J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin *et al.*, *Nature*, 2012, **483**, 603–607.

3 K. Tomczak, P. Czerwinska and M. Winznerorwicz, *Contemporary Oncology*, 2015, **19**, A68–A77.

4 S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal *et al.*, *Nucleic Acids Research*, 2017, **45**, D777–D783.

5 I. C. G. Consortium, *Nature*, 2010, **464**, 993–998.

6 S. Benzekry, C. Lamont, A. Beheshti, A. Tracz, J. Ebos, L. Hlatky and P. Hahnfeldt, *PLOS Computational Biology*, 2014, **10**, e1003800.

7 M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein and N. Beerenwinkel, *PLoS ONE*, 2011, **6**, e27136.

8 L. Yang, S. Wang, M. Zhou, X. Chen, W. Jiang, Y. Zou and Y. Lv, *Scientific Reports*, 2017, **7**, 738.

9 D. Amar, S. Izraeli and R. Shamir, *Oncogene*, 2017, **36**, 3375–3383.

10 L. Loohuis, G. Caravagna, A. Graudenzi, D. Ramazzotti, G. Mauri, M. Antoniotti and B. Mishra, *PLoS ONE*, 2014, **9**, e115570.

11 S. Piccolo, I. Andrulis, A. Cohen, T. Conner, P. Moos, A. Spira, S. Buys, W. Johnson and A. Bild, *BMC Medical Genomics*, 2015, **8**, year.

12 S. Piccolo and L. Frey, *Int J Data Min Bioinform*, 2013, **7**, 245–265.

13 B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nature Methods*, 2014, **11**, 333–337.

14 S. MacNeil, W. Johnson, D. Li, S. Piccolo and A. Bild, *Genome Medicine*, 2015, **7**, year.

15 S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann and J. Cox, *Nature Methods*, 2016, **13**, 731–740.

16 S. M. Hill, L. M. Heiser, T. Cokelaer, M. Unger, N. K. Nesser *et al.*, *Nature Methods*, 2016, **13**, 310–317.

17 Z. Wang, B. Li, S. Piccolo, X. Zhang, J. Li, H. Zhou, J. Yang and L. Qu, *Oncotarget*, 2016, **7**, 35044–35055.

18 A. B. Pavel, J. D. Campbell, G. Liu, D. A. Elashoff, S. M. Dubinett, K. Smith, D. Whitney, M. E. Lenburg and A. E. Spira, *Cancer Prevention Research*, 2017.

19 J. J. Bower, L. D. Vance, M. Psioda, S. L. Smith-Roe, D. A. Simpson, J. G. Ibrahim, K. A. Hoadley, C. M. Perou and W. K. Kaufmann, *npj Breast Cancer*, 2017, **3**, 9.

20 J. Dayton and S. Piccolo, *BMC Medical Genomics*, 2017, **10**, year.

21 G. Liu, X. Zhan, C. Dong and L. Liu, *Scientific Reports*, 2017, **7**, year.

22 B. J. Raphael and F. Vandin, *Journal of Computational Biology*, 2015, **22**, 510–527.

23 J. L. Fleck, A. B. Pavel and C. G. Cassandras, *BMC Systems Biology*, 2016, **10**, 12.

24 B. Vogelstein, N. Papadopoulos, V. Velculescu, S. Zhou, L. D. Jr. and K. Kinzler, *Science*, 2013, **339**, 1546–1558.

25 A. B. Pavel and C. I. Vasile, *Journal of Bioinformatics and Computational Biology*, 2016, **14**, 1650031.

26 I. Bozica, T. Antala, H. Ohtsukid, H. Cartere, D. Kime, S. Chenf, R. Karchine, K. Kinzlerg, B. Vogelstein and M. A. Nowaka, *PNAS*, 2010, **107**, 18545–18550.

27 C. Yeang, F. McCormick and A. Levine, *FASEB Journal*, 2008, **22**, 2605–2622.

28 E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark and A. Ma'ayan, *BMC Bioinformatics*, 2013, **14**, 128–128.

29 A. B. Pavel, D. Sonkin and A. Reddy, *BMC Systems Biology*, 2016, **10**, 16.

30 A. S. Prabowo, A. M. Iyer, T. J. Veersema, J. J. Anink, A. Y. N. Schouten-van Meeteren, W. G. M. Spliet, P. C. van Rijen, C. H. Ferrier, D. Capper, M. Thom and E. Aronica, *Brain Pathology*, 2014, **24**, 52–66.

31 M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe, *Nucleic Acids Res.*, 2014, **42**, D199–D205.

32 M. Kanehisa and S. Goto, *Nucleic Acids Res.*, 2000, **28**, 27–30.

33 J. L. Fleck and C. G. Cassandras, *Nonlinear Analysis: Hybrid Systems*, 2017, **25**, 246–262.

34 J. Fleck and C. Cassandras, IEEE Conference on Decision and Control, 2016, pp. 5041–5046.