Molecular profile to cancer cell line matchmaking

Brendan Reardon

Dana-Farber Cancer Institute <https://orcid.org/0000-0002-1716-5720>

Eliezer Van Allen (✉ EliezerM_VanAllen@dfci.harvard.edu)

Dana-Farber Cancer Institute <https://orcid.org/0000-0002-0201-4444>

Method Article

Keywords: precision oncology, cancer cell lines, high throughput drug screens

Posted Date: October 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.pex-1539/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Profile-to-cell line matchmaking is a computational protocol to identify cancer cell lines that are genomically similar to a patient's case profile. In doing so, high-throughput drug screens applied to the same cancer cell lines may be used for therapeutic hypothesis generation in research settings and potentially in clinical settings. To evaluate the metrics of the matchmaking, a hold-one-out approach of the considered cancer cell lines is applied, and molecular similarity models are assessed based on their ability to identify cancer cell lines that share therapeutic sensitivity.

Introduction

Preclinical data from high-throughput therapeutic screens of cancer cell lines are routinely used to identify individual molecular features associated with therapeutic response or resistance, with the ultimate goal of translating these findings to impact clinical care^{1,2}. A limitation of utilizing such results for translational hypothesis generation is that cell lines that share therapeutic response may be genomically highly dissimilar and therefore have questionable biological relevance to another molecular profile. Therefore, we were motivated to study similarity models which identified cancer cell lines that shared more extensive similarities while maintaining therapeutic sensitivities. Previous approaches evaluated genomic similarity based on shared mutated genes weighted by their recurrence in TCGA^{3,4}. However, we chose to assess models based on shared therapeutic sensitivity independent of ontology-specific priors in this protocol to emphasize potential clinical relevance.

We note areas that may improve this protocol for translational hypothesis generation. First, not all cancer cell lines are tested with every therapy, preventing us from characterizing shared drug response in a more nuanced manner than boolean status. Second, there is likely an opportunity to improve developed genomic similarity models to align with therapeutic sensitivity. The advent of large, clinically annotated and molecular profile patient cohorts may enable these techniques and patient similarity networks to be evaluated for precision cancer medicine on patient profiles rather than cancer cell lines⁵⁻⁷.

Reagents

Somatic variants and copy number alterations for cancer cell lines catalogued in the Cancer Cell Line Encyclopedia were downloaded from cBioPortal and fusions and therapeutic sensitivity were downloaded from the Sanger Institute's Genomics of Drug Sensitivity in Cancer (GDSC)^{1,2}.

Equipment

Molecular Oncology Almanac version 0.4.1
(<https://github.com/vanallenlab/moalmanac/releases/tag/0.4.1>).

A Python 3.7 environment with the following software packages installed:

ipykernel==5.1.0

jupyter==1.0.0

matplotlib==3.0.2

numpy==1.18.2

pandas==1.0.1

scipy==1.4.1

scikit-learn==0.20.1

snfpy==0.2.2

tabulate==0.8.7

oauth2client==4.1.3

openpyxl==3.0.6

xlrd==1.2.0

xmljson==0.2.0

Procedure

1. Cancer cell lines are standardized by name and filtered by requiring: all four data types being available, being of solid tumor origin, not subject to genetic drift between Broad Institute and Sanger Institute characterizations of the cell line per Ghandi et al. 2019, and not reclassified as fibroblast-like by Weck et al. 2017 and Ghandi et al. 2019^{1,8}.

2. Observed somatic variants, copy number alterations, and fusions are processed by the Molecular Oncology Almanac (MOAlmanac) to identify clinically relevant molecular features and annotate for their presence in Cancer Gene Census (CGC)⁹.

3. GDSC's cell line IC50 z-score thresholds are applied to each therapy and cancer cell line pair to generate boolean valued labels for sensitive (z-score < -2.0) and resistant (z-score > 2.0) relationships². Pairwise comparisons are made between all cancer cell lines, noting the intersection of therapies which both

cancer cell lines are sensitive to as well as the intersection size. Each pair of cancer cell lines is deemed to share therapeutic sensitivity if the intersection size of sensitive therapies is greater than zero. Cancer cell lines are further filtered by requiring that they are sensitive to at least one therapy and that there exists at least one other cell line that shares therapeutic sensitivity.

4. Somatic variants, copy number alterations, and fusions are coded into matrices indexed by cancer cell line name with each column associated with a different molecular feature to be used to calculate genomic similarity between pairs of cancer cell lines. The coding of features is dependent on the model implemented, as follows in alphabetically:

- Compatibility (compatibility): Similarity measures between case and comparison profiles are generated based on shared observed clinically relevant features.

First, a total possible score is calculated for each profile based on the set of clinically relevant somatic variants, copy number alterations, and fusions observed in the tumor. Specifically, molecular features which match fully characterized MOAImanac entries (by gene, feature type, and alteration details) receive 75 points, those which only match by gene and feature type receive 25 points, and only 10 points for matching by gene. For example, the features *PIK3CA* p.H1047R, *NUP214-ABL1*, and *CDK12* amplification respectively score 75, 25, and 10 points for a total of 110 points as *PIK3CA* p.H1047R is catalogued in MOAImanac, *ABL1* is catalogued with other fusion partners, and *CDK12* somatic variants but not copy number alterations are catalogued in MOAImanac.

Next, pairwise comparisons are performed to score the intersection of observed molecular features relative to each case profile. Consider a second profile (B) with features *BCR-ABL1*, *CDK* p.L21S, and *TP53* deletion being compared relative to the one described above (A). B scores 35 of 110 potential points from A, resulting in a score of 0.318. Likewise, A relative to B is calculated and the mean of the two values is taken as the similarity measure, or compatibility, between two molecular profiles. This approach was inspired by dating algorithms.

- Jaccard of CGC feature types (jaccard-cgc-feature-types). Sort by agreement-based measure (jaccard) by considering variants in a Cancer Gene Census gene and feature type (e.g. *CDKN2A* copy number alterations match but not a *CDKN2A* deletion and *CDKN2A* nonsense somatic variant). Matrix elements are boolean.

- Jaccard of CGC genes (jaccard-cgc-genes). Sort by agreement-based measure (jaccard) by considering any variant in a Cancer Gene Census gene. Matrix elements are boolean.
- Jaccard of MOAlmanac feature types (jaccard-almanac-feature-types). Sort by agreement-based measure (jaccard) by considering both gene and data type for all somatic variants, copy number alterations, and rearrangements catalogued in the Molecular Oncology Almanac (e.g. CDKN2A copy number alterations match but not a CDKN2A deletion and CDKN2A nonsense somatic variant). Matrix elements are boolean.
- Jaccard of MOAlmanac features (jaccard-almanac-features). Sort by agreement-based measure (jaccard) by considering all somatic variant, copy number, and rearrangement molecular features with alteration details as catalogued in the Molecular Oncology Almanac. Matrix elements are boolean.
- Jaccard of MOAlmanac genes (jaccard-almanac-genes). Sort by agreement-based measure (jaccard) by considering any somatic variant, copy number alteration, and rearrangement in any gene catalogued in Molecular Oncology Almanac. Matrix elements are boolean.
- Multi-pass sort: FDA & CGC (multi-pass-sort_fda-cgc). A weakness of agreement-based measure is that there will be tied values. Tied similarity based on Molecular Oncology Almanac features associated with FDA evidence are further sorted by using similarity based on CGC genes.
- Nonsynonymous variant count (nonsynonymous-variant-count). Similarity is evaluated based on the absolute value of the difference of the number of coding somatic variants between the two cancer cell lines. This is a proxy for mutational burden, as the number of somatic bases considered when calling variants to use a denominator is not available from data sources.
- PCA of CGC genes (pca-cgc-genes). Principal Component Analysis is applied to the matrix of CGC genes, with elements populated if a gene is mutated in a sample either as a somatic variant, copy number alteration, or fusion. Matrix elements are boolean. For example, both *TP53* nonsense variants and copy number deletions can populate elements in the column associated with the gene *TP53*.

- PCA of MOAImanac genes (pca-almanac-genes). Principal Component Analysis is applied to the matrix of MOAImanac genes, with elements populated if a gene is mutated in a sample either as a somatic variant, copy number alteration, or fusion. Matrix elements are boolean. For example, both *TP53* nonsense variants and copy number deletions can populate elements in the column associated with the gene *TP53*.

- Random (random_mean). Cell lines are shuffled against one another randomly across 100,000 seeds. The seed associated with the average mean average precision was chosen.

- SNF: CGC (snf_cgc). The python implementation of Similarity Network Fusion by Ross Markello (<https://github.com/rmarkello/snfpy>) is used to combine similarity across multiple data types¹⁰. Matrices that contain boolean values describing variants in CGC genes altered by (1) somatic variants, (2) copy number alterations, and (3) rearrangements are processed by the tool.

- SNF: FDA & CGC (snf_fda-cgc). The python implementation of Similarity Network Fusion by Ross Markello (<https://github.com/rmarkello/snfpy>) is used to combine similarity across multiple data types¹⁰. Matrices that contain boolean values describing variants (1) in CGC genes that contain a somatic variants, (2) in CGC genes that contain a copy number alterations, (3) in CGC genes that contain a fusion, and (4) associated with an FDA approval, as identified by MOAImanac, are processed by the tool.

- SNF: FDA & CGC genes (snf_fda-cgc-genes). The python implementation of Similarity Network Fusion by Ross Markello (<https://github.com/rmarkello/snfpy>) is used to combine similarity across multiple data types¹⁰. Matrices that contain boolean values describing variants (1) in CGC genes if mutated either as a somatic variant, copy number alteration, or fusion and (2) associated with an FDA approval, as identified by MOAImanac, are processed by the tool.

- SNF: MOAImanac (snf_almanac). The python implementation of Similarity Network Fusion by Ross Markello (<https://github.com/rmarkello/snfpy>) is used to combine similarity across multiple data types¹⁰. Matrices that contain boolean values describing variants in MOAImanac genes altered by (1) somatic variants, (2) copy number alterations, and (3) rearrangements are processed by the tool.

- Somatic tree (somatic-tree). This approach is inspired by CELLector, by Najgebauer et al.⁴. CELLector has a prioritized list of alterations based on cancer type and will report similar cell lines based on mutant / wild type status of each alteration. Likewise, we utilize MOAImanac's prioritization of somatic variants, copy number alterations, and rearrangements observed in a given profile to rank comparison cell lines based on the mutant or wild type status of each molecular feature's feature type and gene altered. The prioritization of somatic events is as appeared in the somatic.scored.txt output of MOAImanac. As an illustrative example, consider a profile with prioritized somatic events, in order: BRAF somatic variant, COL1A1 fusion, and CDKN2A copy number alteration. Cell lines would be sorted into the order of: (1) BRAF, COL1A1, and CDKN2A mutant, (2) BRAF and COL1A1 mutant and CDKN2A wild type, (3) BRAF and CDKN2A mutant and COL1A1 wild type, (4) BRAF mutant and COL1A1 and CDKN2A wild type, and (5) BRAF wild type, etc.

5. Similarity models are evaluated by their ability to sort cancer cell line pairs which share therapeutic sensitivity as more similar than those that do not using evaluation metrics from ranked retrieval¹¹. The metrics precision at rank (k), recall at rank (k), and average precision are used to evaluate a model's ability to sort cell lines relative to one cell line. These metrics are defined as follows, considering a ranked list containing relevant and not relevant documents after a querying many documents:

- Precision at rank k (precision @ k). The number of relevant documents that are in the top k ranked documents divided by k .
- Recall at rank k (recall @ k). The fraction of total relevant documents returned when considering k documents.
- Average precision (AP). The average of precision values at the positions of relevant documents.

These metrics are illustrated by example by considering a ranked list of four items that are relevant, not relevant, relevant, and not relevant (Figure 1). Considering the first item ($k = 1$) yields a precision of 1.0 (1 relevant item / 1 considered item) but also considering the second yields a precision of 0.5 (1 relevant item / 2 considered items). Recall is calculated to be 0.5 at $k = 1$ (1 relevant item / 2 total relevant item), 1.0 at $k = 3$ (2 relevant items / 2 total relevant items), and 1.0 at $k = 4$. Relevant items exist at $k = 1$ and $k = 3$ with associated precision values of 1.0 and 0.66 so the average precision is calculated to be 0.83.

The mean average precision (mAP) is calculated to evaluate each model's performance across all cancer cell lines. The metric is defined as,

- Mean average precision (mAP). The mean of average precision values from multiple queries.

Given three queries which returned six documents with AP values of 0.66, 0.565, and 0.25, the mAP is calculated to be 0.492 (Figure 2).

The AP for each cancer cell line and mAP across all cancer cell lines is calculated for each similarity model.

6. Models are further compared pairwise using randomized testing (Figure 3). The difference in mAP (delta mAP) values between two models is chosen as a test statistic and recorded. For 10,000 iterations using seeds 0 to 9,999, AP values for all cancer cell lines generated by both models are shuffled and both mAP values and delta mAP values associated with the seed are recorded. A p-value describing the significance in difference between the two models is generated by comparing the test statistic to the distribution of 10,000 delta mAP values.

Troubleshooting

All code associated with an implementation of this procedure may be found on Github (<https://github.com/vanallenlab/moalmanac-paper>).

Time Taken

Anticipated Results

It is anticipated that 377 cancer cell lines will be used for evaluation of similarity models after applying filtering criteria.

The model SNF: CGC & FDA is expected to observe the highest average precision at rank $k = 1$, with a value of 0.191. This means that the most similar cancer cell line identified shares a therapeutic sensitivity for 19.1% of evaluated cancer cell lines. This model is anticipated to be within the noise range of two other models attempted: Multi-pass sort: FDA & CGC and Somatic tree. The random model is anticipated to result in an average precision at rank $k = 1$ of 0.095.

References

1. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
2. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–61 (2013).
3. Sinha, R., Schultz, N. & Sander, C. Comparing cancer cell lines and tumor samples by genomic profiles. *bioRxiv* 028159 (2015) doi:10.1101/028159.
4. Najgebauer, H. *et al.* CELLector: Genomics-Guided Selection of Cancer In Vitro Models. *Cell Syst* **10**, 424–432.e6 (2020).
5. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **7**, 818–831 (2017).
6. Pai, S. & Bader, G. D. Patient Similarity Networks for Precision Medicine. *J. Mol. Biol.* **430**, 2924–2938 (2018).
7. Zitnik, M. *et al.* Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Inf. Fusion* **50**, 71–91 (2019).
8. de Weck, A., Bitter, H. & Kauffmann, A. Fibroblasts cell lines misclassified as cancer cell lines. *bioRxiv* 166199 (2017) doi:10.1101/166199.
9. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
10. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337 (2014).
11. Smucker, M. D., Allan, J. & Carterette, B. A comparison of statistical significance tests for information retrieval evaluation. in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* 623–632 (Association for Computing Machinery, 2007).

Acknowledgements

This work was supported by NIH U01 CA233100, NIH R01 CA227388, NIH R37 CA222574, NIH U2C CA252974, Prostate Cancer Foundation (PCF) PCF-Movember Challenge Award, Mark Foundation Emerging Leader Award, and the ASPIRE Award of the Mark Foundation for Cancer Research.

Figures

Figure 1

Evaluation of models applied to one cancer cell line. Similarity models sort cancer cell lines into an ordered list relative to a one molecular profile based on shared genomic features. Their ability to sort cancer cell lines that share drug sensitivity (relevant items) as more similar than those that do not (not relevant) is evaluated.

Figure 2

Evaluation of models applied to many cancer cell lines. Taking the mean of average precision values of a model applied to many cancer cell lines is used to evaluate a model generally.

Figure 3

Comparing models with a randomized test. The difference in mean average precision (delta mAP) between two values is taken as the test statistic. Average precision values from several applications of the two models are shuffled across many seeds to generate a distribution of delta mAP values. The test statistic is compared to the distribution to identify if two models perform significantly different from one another.