Leveraging the Connectivity Map (CMAP) to Identify Candidate Perturbagens Affecting the Differentiation of Non-Small Cell Lung Cancer Cell Lines

Stephanie Martinez, Monica Arniella, Arvind Ravi, Chip Stewart, Gad Getz

Getz Lab, Broad Institute of MIT and Harvard

Abstract

Differentiation therapy is a developing but promising approach to treating cancer with stem cell-like qualities. Due to this, it is valuable to understand how perturbagens (small molecule or genetic) may affect the differentiation of cancer cells. By utilizing CMAP, a library that contains the measured effect on gene expression for a wide range of perturbagens, experimental conditions, and cell lines, we created a list of candidate perturbagens that can be further investigated for their potential utility in differentiation therapy. Specifically, for non-small cell lung cancer cell lines, analyzing the effect that perturbagens have on both lung development and stem cell markers allowed us to define such a list. Quantile thresholds for average z-score and standard deviation, as well as nonparametric testing of the z-score distributions (Wilcoxon Rank-Sum test) were used to define perturbagens of interest. We also used Principal Component Analysis (PCA) on alveolar epithelial differentiation data in order to investigate a hypothesized developmental gene set. Those perturbagens that were consistent between our different set of tests were investigated using the CMAP database. Ultimately, we found interesting hits among the candidate perturbagens, specifically those which function as NFkB inhibitors, cell-cycle inhibitors, and oncogene inhibitors.

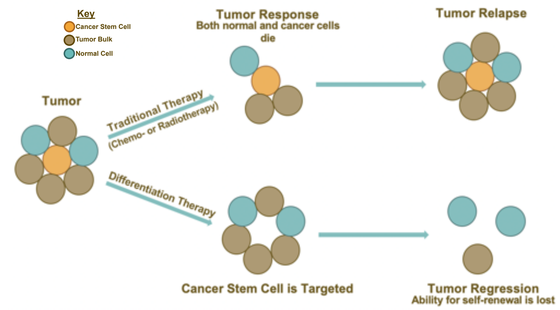
Background

Stem Cells and Cancer

Stem cells are unique for their ability to differentiate into mature cell types while also replicating through self-renewal. Through this cycle, stem cells can regenerate damaged tissue when needed. Although stem cells have therapeutic potential, certain stem cell traits are found in cancer. Cancer can develop when normal cells undergo various mutations which often involve the downregulation of tumor suppressor genes and the upregulation of oncogenes. As a result, cancer cells experience excessive proliferation, metastasize, and cause death. It is essential that cancer with stem cell like traits be further studied to understand how their progression can be stopped.

Differentiation Therapy

Differentiation therapy specifically targets cancer stem cells, which are dedifferentiated mature cells that exhibit a stem cell-like phenotype. Cancer stem cells are then induced with drugs to undergo differentiation and acquire a normal phenotype. Treatments such as chemotherapy are commonly used but kill both cancerous and normal cells, causing detrimental side effects. Furthermore, if the treatment fails to target the cancer stem cell, the cancer can continue to progress.



Literature Markers for Lung Development

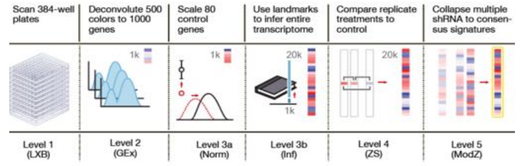
In lung development, there are several stages that span from the embryonic stage to the various mature lineages of the adult lung. Networks of transcription factors involved in development have been shown to be relevant in different cancer types. In lung cancer specifically, the developmental lineage and cell of origin are related to disease characteristics and therapeutic implications.

Data

What is CMAP?

CMAP is a large database that contains gene expression profiles for perturbing agents such as small-molecule compounds and genetic reagents across multiple cell lines. Analyzing CMAP can allow one to draw similarities between cellular signatures which, in effect, elucidates the effect of a perturbagen on its gene target.

Right figure is from the GEO CMAP LINCS User Guide v.2.1.



Acquire CMAP — CMAP was developed in the Broad Institute using the L1000 gene expression assay.

Literature Marker Search — Lung development and stem cell markers were identified.

Filtering for Significant Perturbagens — Average z-score and standard deviation across all genes for each perturbagen were used for quantile thresholds.

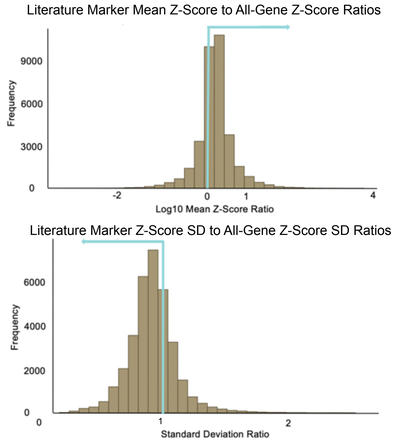
Principal Component Analysis (PCA) of Alveolar Differentiation Data — Used to verify which landmark genes have have a role in the alveolar developmental trajectory.

Filter by Perturbagens with Positive Z-Scores — Perturbagens with negative average z-scores were removed as they were not supported by PCA analysis.

Final Round of Hypothesis Tests — Determining what perturbagens from the high mean z-score set overlap with the Wilcoxon significant set (p.adj < 0.05).

Identifying Perturbagens of Interest Using Gene Set Expression

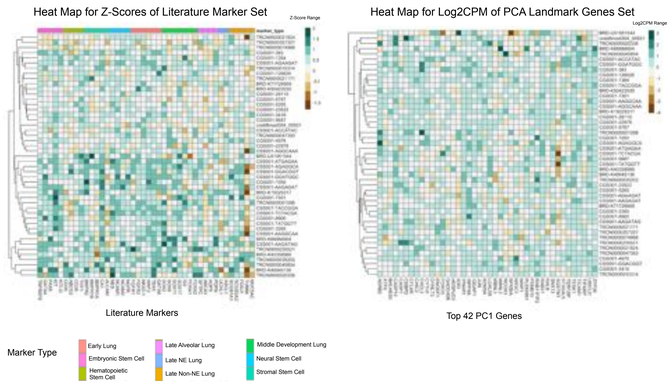
Comparing Distributions of all Genes v. Canonical Stem Cell and Lung Development Markers



As most of the distribution of standard deviations is to the left of the line (at ratio = 1), this indicates that there is greater z-score variability (higher SD) by perturbagen across all genes than within our set of markers. As a greater area of the mean ratio distribution is to the right of the line (at log10(ratio) = 0, or, ratio = 1), this indicates that the majority of the mean z-scores are larger for the marker gene set than across all genes.

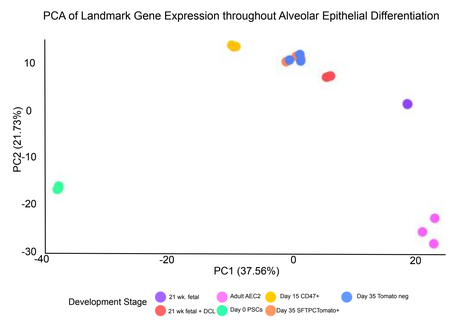
Visualizing Z-Scores of Gene Expression for both PCA-derived and Literature Marker Sets for Top Candidate Perturbagens

We compared the distributions for average z-scores of gene expression for both the PCA gene set (genes derived from PCA on alveolar differentiation data set) and our literature marker set. As the genes contributing to differentiation according to PCA were most similar to the literature gene set for a subset of perturbagens, only those perturbagens (positive average z-scores) were selected for further analysis.

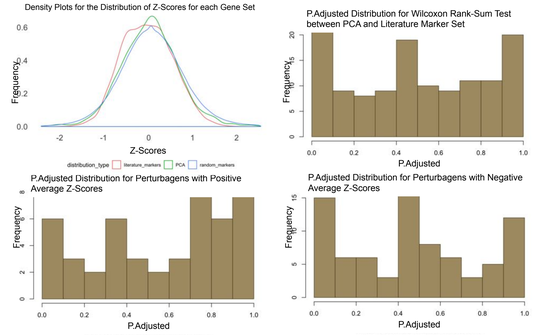


Assessing Concordance of Unsupervised and Canonical Sets

Comparing the Distributions of All Genes v. Canonical Stem-Cell and Lung Development Markers



Determining Factors for P.Adjusted Distributions for Wilcoxon Rank-Sum Test



Wilcoxon Rank-Sum Analysis

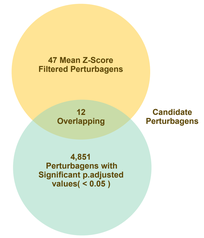
Wilcoxon Rank-Sum Test for PCA and Literature Gene Sets

We compared the distributions of the PCA-derived markers and the literature gene set using the Wilcoxon Rank-Sum test. The distribution for the multiple-hypothesis corrected values showed that perturbagens with negative z-scores had a greater proportion of significant p-values (indicating differences in z-score distribution) than the perturbagens with positive z-scores. To have a consistent perturbagen set, we excluded perturbagens that had overall negative z-scores.

Wilcoxon Rank-Sum Test for Literature and Random Gene Sets

To further filter candidate perturbagens, we set a criteria for perturbagens that are both in the the last round of filtering and contain p.adjusted values below 0.05 to be selected. The test revealed that about 4,800 out of all perturbagens had a p.adjusted < 0.05. However, out of the 4,800 perturbagens, 12 overlapped with the last round of filtering.

Perturbagen ID	Description
BRD-K15025317	NFkB pathway inhibitor (NFkB pathway related to immune response to infection); non-steroidal anti-inflammatory drug
BRD-K40358966	Member of Piperidines
BRD-U51951544	c-Jun N-terminal kinases (JNK) inhibitor (JNK signaling is important for inflammatory response). • Has high connectivity scores with cell-cycle inhibitors
CGS001-1050	NFkB pathway inhibitor; knockdown of CEBPA (transcription factor involved in blood cell differentiation) • Has a high connectivity score with hexamethylenebisacetamide (a differentiation inducer and NFKB inhibitor)
CGS001-3265	Knockdown of HRAS (proto-oncogene, involved in regulating cell division)
CSS001-AAGAGAT	The CMAP database and metadata did not contain annotations for these perturbagens.
CSS001-AAGATAG	
CSS001-AGAAGAT	
CSS001-ATGAGAA	
CSS001-GGATGGC	
TRCN0000010314	



Conclusion

We used different types of filtering to finalize a list of 12 candidate perturbagens that:

- Were above the 90th percentile for positive average z-score across a literature marker set for stem cell and developmental markers
- Below the 35th percentile for standard deviation across literature marker set
- Had a significant p.adjusted value below 0.05 for the Wilcoxon Rank-Sum test between the literature marker set and a random gene set for all perturbagens

However, this list needs to be further validated through different analyses as well as additional, more recent data from CMAP. In particular, testing distributions using randomizations could be made more robust through a greater number of trials. Including perturbagens with negative average z-scores would also be valuable. Future research should also consider that these results were derived from experiments done on cell lines only. Other aspects to note include subjectivity in our filtering methods using average z-score and standard deviation, as there exists the possibility of missing potential significant perturbagens. For future work, we would like to minimize subjectivity in choosing a representative literature marker set and utilize filtering methods that can more accurately select candidate perturbagens. Additionally, we should also consider the results of parametric statistical tests, as the final Wilcoxon Rank-Sum Test is a non-parametric test. Finally, refining our literature marker set is also a potential point of improvement that could aid in our analyses.

Acknowledgements

Many thanks to Rachel Gesserman and Michael Mavros, who provided me with the outline, materials, and invaluable feedback necessary to do this project. Also, many thanks to Arvind Ravi and Chip Stewart for their help developing this project. My wise mentor, Monica Arniella, also guided me every step of the way and taught me the concepts needed to conduct the different analyses. Lastly, I would like to express my gratitude to the Broad's Office of Diversity, Education, and Outreach for all of their support and for giving me the opportunity to conduct research at the Broad Institute this summer.

References:

Thé, Hugues de. "Differentiation Therapy Revisited." *Nature News*, Nature Publishing Group, 1 Dec. 2017, www.nature.com/articles/nrc.2017.103.

Maria Serra, Konstantinos-Dionysios Alysandratos, et al. "Pluripotent Stem Cell Differentiation Reveals Distinct Developmental Pathways Regulating Lung- versus Thyroid-Lineage Specification." *Development*, Oxford University Press for The Company of Biologists Limited, 1 Nov. 2017, dev.biologists.org/content/144/21/3879.

Herriges, Michael, and Edward E Morrisey. "Lung Development: Orchestrating the Generation and Regeneration of a Complex Organ." *Development (Cambridge, England)*, Company of Biologists, Feb. 2014, www.ncbi.nlm.nih.gov/pmc/articles/PMC3899811/.

Subramanian, Aravind, et al. "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles." *Cell*, U.S. National Library of Medicine, 30 Nov. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5990023/.

Cheung, W K C, and D X Nguyen. "Lineage Factors and Differentiation States in Lung Cancer Progression." *Oncogene*, U.S. National Library of Medicine, 19 Nov. 2015, www.ncbi.nlm.nih.gov/pubmed/25823023.

de Castro Barbosa, Maria Letícia, et al. "NF-KB Signaling Pathway Inhibitors as Anticancer Drug Candidates." *Anti-Cancer Agents in Medicinal Chemistry*, U.S. National Library of Medicine, 2017, www.ncbi.nlm.nih.gov/pubmed/27481554.