Reverse causal inferencing on molecular data from lung adenocarcinoma patients reveals a stem cell-like molecular subtype associated with pack year history

Renée Deehan, Sergey Korkhov, Alexis Foroozan, Scott Marshall, Nimisha Schneider

QuartzBio, a Precision for Medicine Company, Boston, MA

INTRODUCTION

Advances in high throughput measurement technologies (-omics data) have made it possible, and increasingly affordable, to generate high complexity, high volume data for oncology research. Coordinate efforts in computational modeling and machine learning applications to biological data have yielded increasingly sophisticated methods to deeply characterize the mechanisms of disease pathogenesis and heterogeneity among patients, which are predicates to the rapid development and timely and safe administration of efficacious treatments. Open source repositories that catalog, harmonize and host -omics data collected from clinical or preclinical studies, generously donated by patients and researchers, provide revolutionary access to otherwise siloed data. The combination of these advancements enabled us to characterize the molecular phenotypic heterogeneity that exists within a lung adenocarcinoma cohort from The Cancer Genome Atlas [1]. Using an inferencing method on our proprietary knowledgebase of interconnected, scorable, directed networks, we identified mechanisms driving lung adenocarcinoma pathogenesis using mRNA expression data from tumors compared to adjacent controls, and from heavy smokers (>40 pack years) compared to light smokers (<10 pack years). Lung adenocarcinoma tissue was defined by a molecular phenotype indicative of tumor stem cell-like signatures. This was more pronounced in heavy smokers compared to light smokers, indicating a potential dose-dependent relationship between cigarette smoke exposure and the active mechanisms in the tumor microenvironment, which may impact response to tumor microenvironment-sensitive therapies, such as immune checkpoint inhibitors.

METHODS

Description of Lung Cancer Adenocarcinoma Data Set

Two hundred thirty resected lung adenocarcinomas were characterized by mRNA sequencing and published as part of the The Cancer Genome Atlas [1]. Tumor and normal tissues were evaluated, as well as smoking status of individuals. Approximately 80% of patients in the data set had a previous smoking history.

Differentially Expressed Gene (DEG) Analysis

Significant changes in gene expression were calculated using the limma package from the R Bioconductor suite [2]. DEGs were calculated by comparing read counts from the following contrasts: (1) all tumor tissue to all adjacent tissue; (2) tumor tissue from heavy smokers (>40 pack years) to tumor tissue from light smokers (<10 pack years). 4487 and 2410 significant changes in gene expression were identified, respectively.

Reverse Causal Inferencing

We applied a reverse inferencing approach that systematically interrogates RNAseq measurements from tumor and control biopsies against a graph database of cause and effect interactions (Figure 1A). Within this graph, there exist directed signatures that describe the impact of a perturbation on a set of downstream changes in mRNA expression (Figure 1A, inset). Our library was nucleated through an open sourced knowledge graph [3] and enhanced with updated and relevant knowledge using the Biological Expression Language framework [4]. Significant changes in gene expression from the previously defined contrasts were mapped onto 526 directed graphs contained in the knowledgebase (e.g., Figure 1B). Each network was evaluated for significance of network enrichment of mapped changes and confidence in the inferred directionality of the upstream node in the directed network. Mechanisms that met or were below near a $p \leq 0.1$ for both metrics were considered for further biological interpretation analysis. Inferred mechanisms were connected using known prior knowledge links into a more holistic graph describing the molecular phenotype present in lung cancer tissue.

FIGURE 1.

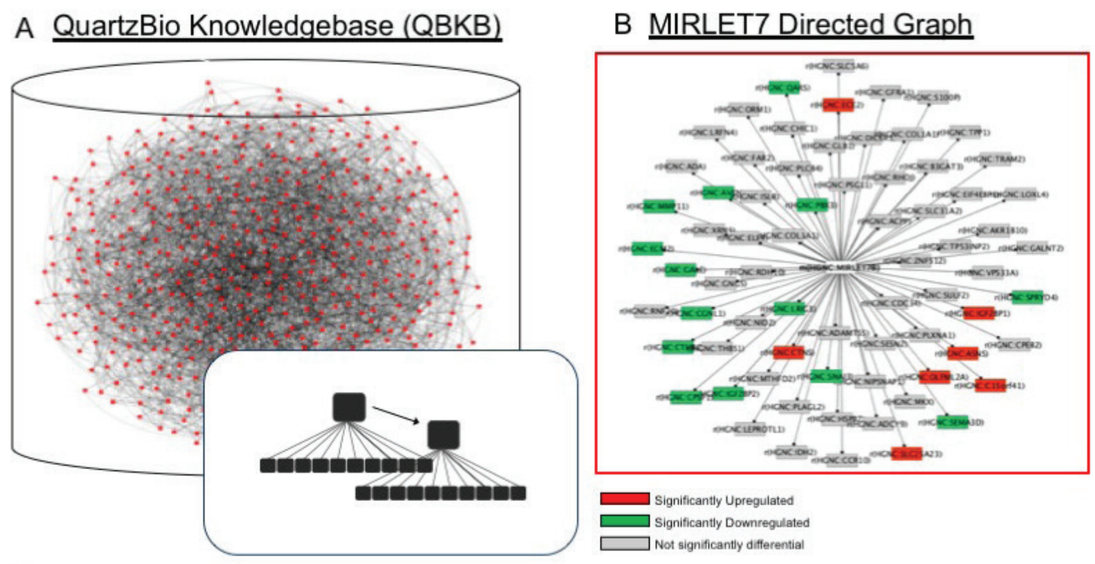
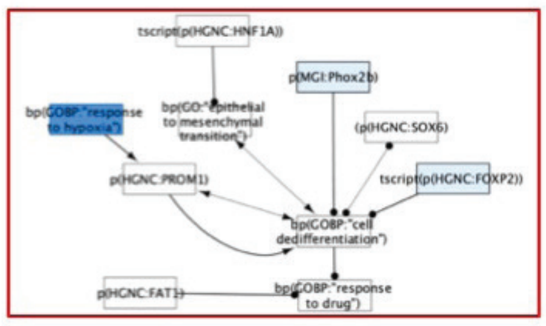


Figure 1. QuartzBio Knowledgebase is made up of directed graphs representing biological mechanisms upstream of gene expression. A, Schematic of QBKB showing network of molecule-interaction-molecule triples; directed graphs can be extracted from the overall network (inset). B, Example of a directed graph representing Mirlet7b activity and its downstream gene targets. Mapping gene expression data allows evaluation of the activity of Mirlet7b in a specific dataset.

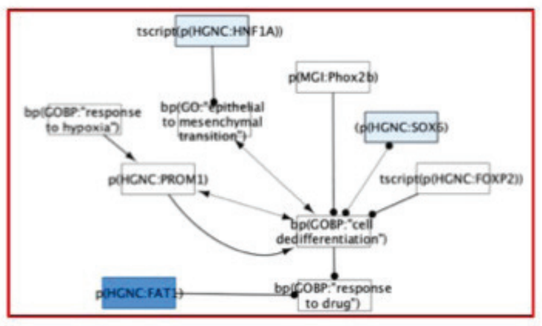
RESULTS

Figure 2.

A



B



C

COMPARISON	NETWORK	ENRICHMENT	DIRECTIONALITY CONFIDENCE
All tumor vs. all normal	Decreased m(HGNC:MIRLET7B)	0.12	0.095
	Increased bp(GOBP:"response to hypoxia")	0.18	1.1E-08
	Decreased p(MGI:Phox2b)	0.12	0.11
	Decreased tscript(p(HGNC:FOXP2))	0.13	0.06
	Increased tscript(p(HGNC:TP63))	1.8E-02	3.2E-03
Heavy tumor vs. light tumor	(not significant) m(HGNC:MIRLET7B)	0.79	1.7E-02
	Decreased (p(HGNC:SOX6)	0.14	3.2E-03
	Increased m(MGI:Mir140)	2.5E-02	0.11
	Decreased tscript(p(HGNC:HNF1A))	7.0E-02	7.6E-03
	Increased p(HGNC:FAT1)	6.3E-02	0.11

Figure 2. Mapping changes in gene expression to directed gene networks elucidates upstream biological activity by highlighting active gene networks that represent molecular mechanisms. A, Mechanistic inference results from lung adenocarcinoma tumor tissue gene expression contrasted with tumor adjacent tissue indicate a dedifferentiation phenotype. B, Inference results from tumor tissue in heavy smokers contrasted with tumor tissue in light smokers indicates additional dedifferentiation mechanisms to be active in a patient subtype defined by heavy smoking. Purple indicates increased mechanism activity; blue indicates decreased mechanism activity.

In LUAD tumor cells, we detected a pattern of gene signatures which indicated a tumor stem cell-like phenotype characterized by:

- predicted decreases in the activity of pro-differentiation factors FOXP2 and PHOX2B
- an increased response to hypoxia, a feature of the tumor microenviroment with implications for dedifferentiation of tumor cells [5].

Analysis of patients with heavy (>40) versus light (<10) pack-year burden suggested an augmented dedifferentiation profile in heavy smokers, including

- decreased SOX6 [6],
- decreased HNF1A [7], and
- increased FAT1 signaling [8].

FAT1 signaling has also been recently implicated in resistance to immune checkpoint inhibitors in NSCLC patients [9]. Expression of PD-L1 has been associated with cumulative inhaled smoke exposure in lung cancer patients, and with mechanisms inferred from our analysis (e.g., MIR140 signaling) [10,11]. However, MIR140 on PD-L1 is reported to suppress PD-L1, while increased smoke exposure is reported to associate with higher PD-L1 expression; the behavior of PD-L1 in this cohort requires further investigation.

CONCLUSIONS

Our in silico analysis of lung cancer patient biopsies generated hypotheses implicating stem cell signaling in tumors and a further stratification of this signal based on patient pack year burden. Given the implication of tumor stem cells in resistance to drug treatments in cancers, there is increased focus on immunotherapies as a mechanism to circumvent this resistance, and the role of PD-L1 in this milieu requires further examination. Understanding the molecular makeup of these subtypes is critical when considering clinical research treatment strategies in immuno-oncology. Our results indicate computationally derived molecular mechanisms with potential involvement in tumor cell stemness and drug resistance with the potential for experimental follow-up.

REFERENCES

1. The Cancer Genome Atlas. https://portal.gdc.cancer.gov/.
2. Ritchie M, et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
3. Catlett N, et al. Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data. *BMC Bioinformatics*. 2013;14-340.
4. Open BEL. https://openbel.org.
5. Yun Z and Lin Q. Hypoxia and regulation of cancer cell stemness. *Adv Exp Med Biol*. 2014;772:41-53.
6. Qin YR, et al. Characterization of tumor-suppressive function of SOX6 in human esophageal squamous cell carcinoma. *Clin Cancer Res*. 2011;17(1):46-55.
7. Pelletier L, et al. HNF1α inhibition triggers epithelial-mesenchymal transition in human liver cancer cell lines. *BMC Cancer*. 2001;11:427.
8. Srivastava C, et al. FAT1 modulates EMT and stemness genes expression in hypoxic glioblastoma. *Int J Cancer*. 2018;142(4):805-812.
9. Fang W, et al. Comprehensive genomic profiling identifies novel genetic predictors of response to anti-PD-(L)1 therapies in non-small cell lung cancer. *Clin Cancer Res*. 2019 [epub ahead of print].
10. Villanueva N, Bazhenova L. New strategies in immunotherapy for lung cancer: beyond PD-1/PD- L1. *Ther Adv Respir Dis*. 2018;12:1753466618794133. doi:10.1177/1753466618794133.
11. Xie W, et al. MiR-140 expression regulates cell proliferation and targets PD-L1 in NSCLC. *Cell Physiol Biochem*. 2018;46(2):654-663.

ACKNOWLEDGMENTS

We would like to extend our appreciation to the patients that participated in the TCGA study, and to the QuartzBio's team for giving the authors room to explore using these data. We'd also like to acknowledge Mehrab Bari for initial exploration into the data, and Natalie Catlett and William Hayes for answering numerous BEL and knowledge-related questions.