Lung cancer progression and staging with trancriptome bioformatics and entropy analysis

Felipe Leal Valentim1, Carlyle Ribeiro Lima1, Nicolas Carels1,\*

1Laboratory of Biological Systems Modeling, Center of Technological Development in Health (CDTS), Oswaldo Cruz Foundation (Fiocruz), 21040-900, Rio de Janeiro, Brazil.

\*Correspondence to nicolas.carels@fiocruz.br

# Abstract

**Keywords**:

# 1. Introduction

# 2. Materials and Methods

## RNA-seq materials

RNA-seq from TCGA samples of lung squamous cell carcinoma (LUSC) were downloaded from the GDC portal (https://portal.gdc.cancer.gov/) accessed on 2024-04-04. Among the 486 lung samples, 45 were paired samples (45 tumor and 45 non-tumoral samples, each from a same patient), and the remaining (441) were non-paired, which means that no control from healthy lung was available for them. The clinical sheet informed that (I) for stage I, 198 LUSC samples were non-paired while 24 were paired, (ii) I, 130 LUSC samples were non-paired, while 17 were paired, (iii) and for stage III, 68 LUSC samples were non-paired, while 4 were paired.

## RNA-seq processing

RNA-seq counts were normalized according to the reads per kilobase per million mapped reads (RPKM) methodology as described by (Mortazavi et al., 2008) and genes with average RPKM ≤ thresholds were filtered out. Then, we identified tumor genes by comparing expression on 45 tumor samples by reference to the 45 normal samples. Once this list of up-regulated tumor gene has been calculated, we obtained stage-specfic genes.

In both steps, the identification of up-regulated tumor genes and the obtention of stage-specfic genes, log2 fold change is calculated using the same formula: log2(fold change) = log2(expression value in condition A)/log2(expression value in condition B). Moreover, in both cases, a t-test on the RPKM values comparing the two conditions is applied. In the case of up-regulated tumor genes, the average expression in tumor samples and the average expression in normal samples are used. For each gene, a paired t-test is applied and FDR is calculated (method="BH"). About the stage-specfic genes, the average expression among samples in each stage is compared to the average expression among samples in the two other stages. Then, unpaired t-test is applied and FDR (method="BH") is calculated for each stage-specfic gene.

## Sub-interactome networks

To uncover the pathways in which up-regulated genes being specific of a given stage were involved, we crossed them with interactome data. Sub-interactome networks were constructed for each stage by sampling the IntAct interactome (Orchard et al., 2013) with their respective specific up-regulated genes. The IntAct interactome was obtained from the intact-micluster.txt file (version updated December 2017) accessed on January 11, 2018, at ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt. We excluded incomplete and non-human interactions from this file, and the resulting file presented 151,631 interactions among 15,526 human proteins with UniProtKB accessions. To construct sub-interactome networks, pairwise combinations among stage-specific genes are created, and then filtered to keep only those edges that overlap with the interctome.

## Gene regulatory networks

Gene regulatory networks (GRN) are valuable information to uncover genes whose co-expression is correlated. To construct the gene co-expression network, low variation genes are removed based on a threshold for the percentage of genes to be retained (75%). Then a correlation matrix is constructed using Spearman rank correlation, but only the upper diagonal is kept to avoid redundant edges. Finally, a correlation threshold was used to maintain edges with significant co-expression.

## Bipartite network

Up-regulated specific genes may belong to independent clusters portraying independent pathways. A bipartite networks is expected to help connecting these pathways through the information obtained from gene co-regulation (GRN) or/and to diagnosis genes that are under a same regulation process. Bipartite network for stage-specific genes were constructed from the co-expression netwoks using igraph.

## Shannon entropy

Quoting Enneking et al. (1980), the purposes of a staging system are to “(1) incorporate the significant prognostic factors into a system that describes progressive degrees of risk of local recurrence and distant metastases to which a patient is subject, (2) stratify the stages so they have specific implications for surgical management, and (3) provide guidelines for adjunctive therapies”. Staging is correlated with 5-year overall survival (OS) in lung (Jeon et al., 2015). 5-year OS is, therefore, a measure of aggressiveness and it has been shown that it is correlated to Shannon entropy of up-regulated gene sub-networks (Conforte et al., 2019). Thus, we postulated that tumor staging could be correlated to Shannon entropy (Shannon, 1948) implemented in R package according to formula 1.

(1)

where p(k) is the probability of occurrence of a vertex with a rank order k (k edges) in the sub-network considered. Since entropy is an extensive thermodynamic function of states, it should not be normalized for network size. The sub-networks were generated automatically from gene lists found to be up-regulated and specific of a given stage regarding the others.

# 3. Results

2333 genes are identified to be up-regulated in tumor (RPKM ≥ 10, log2foldchange tumor ≥1, and paired t-test FDR ≤ 0.05). Furthermore, we obtained 201 genes specific to stage I, 438 to stage II and 438 genes to stage III (log2foldchange per stage ≥1 and unpaired t-test FDR ≤ 0.05). From these stage-specific genes, co-expression networks were constructed; resulting in networks with 208 edges, 471 edges and 491 edges, for stages II and III, respectivelly. Shannon entropy from the co-expression network showed values of 0.3646, 0.6405 and 0.7498, from stages I to III.

In addition to co-expression networks, sub-interactomes were also generated from the same stage-specific genes of stage I, II and III; resulting in networks with 40, 480, 350 edges, and respective shannon entropy values of 1.2713, 2.8299 and 2.5851.

# 4. Discussion

# 5. Conclusion

# Acknowledgement

Aqui, os detalhes da sua bolsa.

# References

1. Conforte AJ, Tuszynski JA, da Silva FAB, Carels N. Signaling Complexity Measured by Shannon Entropy and Its Application in Personalized Medicine. Front Genet. 2019 Oct 21;10:930. doi: 10.3389/fgene.2019.00930.

2. Enneking WF, Spanier SS, Goodman MA. A system for the surgical staging of musculoskeletal sarcoma. Clin Orthop Relat Res 1980;153:106–120

3. Jeon DS, Kim HC, Kim SH, Kim TJ, Kim HK, Moon MH, Beck KS, Suh YG, Song C, Ahn JS, Lee JE, Lim JU, Jeon JH, Jung KW, Jung CY, Cho JS, Choi YD, Hwang SS, Choi CM; Korean Association for Lung Cancer; Korea Central Cancer Registry. Five-Year Overall Survival and Prognostic Factors in Patients with Lung Cancer: Results from the Korean Association of Lung Cancer Registry (KALC-R) 2015. Cancer Res Treat. 2023 Jan;55(1):103-111. doi: 10.4143/crt.2022.264.

4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5: 621–628. pmid:18516045

5. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014 Jan;42(Database issue):D358-63. doi: 10.1093/nar/gkt1115.

6. Shannon C. E. (1948). A mathematical theory of communication. Bell. Syst. Tech. J. 27 (July 1928), 379–423. 10.1002/j.1538-7305.1948.tb01338.x