Transcriptome and entropy analysis of lung squamous cell carcinoma progression and staging

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

Stage-specific signatures were detected followed by analysis of their respective sub-interactomes. Remarkably, 23, 62 and 169 genes were determined as potential biomarkers, respectively for stages I, II and III. From the stage-specific networks constructed from sub-interactome of biomarkers, shannon entropy was calculated to test whether there is more cancer information on expression of genes in the later stages. Entropy values of 3.669, 3.713 and 3.786 for stages I, II and III shows that perhaps cancer progression is detected at transcriptome level. Stage-wise functional analysis showed that for stage I anontations with greatest number of genes are : (GO) tertiary granule lumen, specific granule lumen, cytoplasmic stress granule; and (KEGG) Cell surface interactions at the vascular wall and Neutrophil degranulation. For Stage II, (GO) mesenchyme development, axonogenesis and regulation of nervous system development. Finally, for stage III, (GO) ribosome and (KEGG) Ribosome. Moreover, Translation and SARS-CoV Infections annotate genes from both stages II and III but not for stage I. Most abundant protein families detected in stage III are PSM (PSMA7, PSMB1, PSMC2, PSMC6, PSMD4, PSMD7) and MRPL (MRPL17, MRPL18, MRPL40, MRPL48, MRPS16, MRPS23).

**Keywords**:

# 1. Introduction

Lung squamous cell carcinoma

biomarkers for Lung squamous cell carcinoma

classification and prediction of Lung squamous cell carcinoma stages

Transcriptome of cancer progression and staging

function annotation and transcriptome of cancer progression and staging

# 3. Results

## 3.1. Stage-specific signatures for squamous cell lung cancer

While 4882 genes are identified to be up-regulated in tumor (RPKM ≥ 4 and paired t-test FDR ≤ 0.05), 1618, 1694 and 1795 were distributed to stages I, II and III, respectively (RPKM ≥ 4 , log2foldchange ≥1 and unpaired t-test FDR ≤ 0.05). From these, the number of stage-specific genes, i.e. without overlap to other stages, is 23, 62, 169 in stages I, II and III, respectively (see Figure VenDiagram).

## 3.2. Cancer progression measured by shannon entropy

Genes distributed respectively to stages I, II and III were used to sample sub-interactomes with 1140/3291, 1198/3564, 1280/4002 vertex/edges. From these, entropy values of 3.669, 3.713 and 3.786 for sub-interactomes of stages I, II and III were calculated.

## 3.3. Stage-wise network and functional analysis

ClusterProfiler was used to annotate stage-specific genes against GO, KEGG and REACTOME databases (see Table annotation). For each of these three layer, top three annotations in number of genes were selected for inspection (see Figure Annotaton). The most abundant terms in number of genes annotated for Stage I are (GO) tertiary granule (2), (GO) tertiary granule lumen (2), (KEGG) Steroid biosynthesis (1), (KEGG) Viral myocarditis (1), (REACTOME) Cell surface interactions at the vascular wall (2) (REACTOME) and Neutrophil degranulation (2) (REACTOME). For stage II, terms more abundant are (GO) mesenchyme development (7), (GO) regulation of nervous system development (7), (KEGG) Endocytosis (5), (KEGG) Axon guidance (6), (REACTOME) Signaling by ROBO receptors (4) and (REACTOME) Translation (4). Finally for stage III, we computed: (GO) mitochondrial protein-containing complex (15), (GO) mitochondrial inner membrane (16), (KEGG) Pathways of neurodegeneration - multiple diseases (11), (KEGG) Alzheimer disease (13), (REACTOME) PIP3 activates AKT signaling (12) and (REACTOME) Translation (12).

Stage I and II overlap in annotation for WWWWW (genes) and PPPP (genes) and Stage II and III for DDDDD (genes) and TTTT (genes). RRRR is annotated consistently for genes of the three stages.

# 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) for stage II, 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.

## Signature detection framework

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.

## 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.

## Functional annotation analysis

ClusterProfiler [7] was used to annotate stage-specific genes against GO (enrichGO, pvalueCutoff = 0.1,qvalueCutoff = 0.15, minGSSize = 3), KEGG (enrichKEGG, pvalueCutoff = 0.15) and REACTOME (enrichPathway, pvalueCutoff = 0.15) databases. For each of these three layer, top anonontations in number of genes were selected for further analysis.

# 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

7. “clusterProfiler: an R package for comparing biological themes among gene clusters.” OMICS: A Journal of Integrative Biology, 16(5), 284-287. doi:10.1089