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

1. 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...Potential biomarkers for each stage were selected based on t-test (FDR ≤ 0.05).... This work supports the fact that transcritpome can reveal markers for cancer staging and study of cancer progression....

**Keywords**:

# 1. Introduction

# 2. Results

## 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 are distributed to stages I, II and III, respectively (RPKM ≥ 4 , log2foldchange ≥1 and unpaired t-test FDR ≤ 0.05). These genes were used to sample sub-interactomes with 1140/3291, 1198/3564, 1280/4002 vertex/edges for each stage. Then, entropy values of 3.669, 3.713 and 3.786 were obtained for sub-interactomes of stages I, II and III. Finally, 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 1).

**Figure 1 : A) Ven diagram of turmor genes and their stage-specific distribution. B) PCA of tumor genes contrasting tumor and normal paired samples.**

## Stage-wise functional analysis

ClusterProfiler [cite] was used to annotate stage-specific genes against GO , KEGG databases. Additionally, the Human Gene Database [cite] was queried Genecards summary was stored. Diseases associated, related pathways and gene ontology annotation were extrated from summary (see Table 1). For each of these terms, we counted the number of genes. Then, ten most abundant terms in number of genes per stage were inspected (see Table 2). Terms common to all stages are protein homodimerization activity, Asparagine N-linked glycosylation, Cell surface interactions at the vascular wall, Neutrophil degranulation, Infectious disease, Metabolism of proteins, Innate Immune System, SARS-CoV Infections andRNA binding.

Terms most abundant in the later stage III are: Signaling by WNT (13), Pathways of neurodegeneration - multiple diseases (14), mitochondrial matrix (16), mitochondrial protein-containing complex (16), Alzheimer disease (16) and mitochondrial inner membrane (16). The human proteasome gene family (PSM) consists of 49 genes that play a crucial role in cancer proteostasisFrom this gene family we found PSMB1, PSMC6, PSMC2, PSMD4, and PSMD7 annotated with terms Signaling by WNT, Pathways of neurodegeneration - multiple diseases and Alzheimer disease. Morever, mitochondria plays an important role in cancer through macromolecular synthesis and energy production. Genes from MRP (7), TIMM (2) and COX (2) gene families are annotated with terms related to mitochondrial activitymitochondrial matrix, mitochondrial protein-containing complex and mitochondrial inner membrane).

**Table 1 : Annotation interpretation via GeneCards and ClusterProfiler.**

**Table 2 : Ten most abundant terms in number of genes per stage.**

* 1. **Tumor genes**

1. Pubmed was searched for known biomarkers of lung cancer [PMC10982305, PMC2742385, PMC4254270, PMC4630522, PMC5137804, PMC7881207, PMC9537298, PMC9876670] and 67 genes were found. Among these, 24 genes are listed as tumor genes in our dataset (FDR ≤ 0.05).
2. **Table 3 : Statistics of stage-specific gene.**
3. **Table 4 : Published tumor biomarkers.**
   1. **Figure 5 : Expression of known tumor biomarkers.**
   2. **Biomarkers selection for squamous cell lung cancer staging**From the stage-specific genes, statistics were computed t test was computed from paired and tumoral samples (see Table 3). Noteworthy, we used FDR topoint to most prominent makers per stage. For stage I COPB2 [PMC8798258, PMID:29674272, PMC8455385] and DHX36 [PMC8111079], both known for promoting metastasis in lung cancer, are selected. From stage II, we note RNPS1 and NONO with literature associating these genes to lung cancer [PMC10800354, PMID:38493226]. From stage III, we hightlight RPE, GLRX5 and PPP4C as potential biomarkers [PMC3623115, PMC4731000, PMC10290155].
4. **Table 5 : Selected stage-specific biomarkers.**

**Figure 2 : Analysis potential biomarkers for stage I.**

**Figure 3 : Analysis potential biomarkers for stage II.**

**Figure 4 : Analysis potential biomarkers for stage III.**

# 3. 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 441 tumor samples by reference to the 45 normal samples. The average expression in tumor samples and the average expression in normal samples were used to calculate FDR of paired t-test (method="BH"). Once this list of up-regulated tumor gene has been calculated, we obtained stage-specfic genes by comparing expression of genes in samples of each stage by reference to the normal samples, i.e. log2(fold change) = log2(expression value in stage A/expression value in stages normal samples), 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 , KEGG databases. For each of these three layer, top ten anonontations in number of genes were selected and combined for further analysis if selection had more than one gene. The Human Gene Database [cite] was queried for each ENSEMBL identifier and corresponding Genecards summary was stored. Diseases associated, related pathways and gene ontology annotation were extrated from summary.

# 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