Supplementary Methods to Iorio et al.: Dissecting the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich

SLAPenrich details, case study and comparison with other tools

S1. Introduction

Here we describe in detail the mathematics underpinning SLAPenrich, its implementation, a case study, as well as a comparison with PathScore and PathScan, two related tools.

SLAPEnrich is implemented as an R package (available at https://github.com/francescojm/SLAPenrich), documentation and submission to Bioconductor in progress).
It includes different collections of pathway gene sets from multiple public available sources [1], together with all the data objects needed to run the analysis described in our manuscript. However, it can be also used with any user-defined collection of gene-sets. An overview of the exposed functions of the R package is provided in Additional File 6.
The statistical framework implemented by SLAPenrich is detailed in the Methods section of our manuscript.
To visualize enriched pathways SLAPenrich makes use of presence/absence matrices visualised as *binary heatmaps* where columns indicate samples, rows indicate genes harboring at least one somatic mutation in at least one sample of the analyzed dataset, and colors indicate the absence or the presence of somatic mutations (respectively) in a given gene/sample combination. To emphasize mutual exclusivity trends among the row-wise mutation patterns, rows and columns of these heatmaps are sorted with a heuristic method (detailed below) that minimizes the superposition of mutated samples column-wisely, thus the overlaps of the mutation patterns across the rows (an example is provided in **Error! Reference source not found.**A described in the next section). To finally summarize the results, an analysis of the enriched-pathway *core-component genes* can be performed. The aim of this final analysis is to visualize in the same heatmap enriched pathways that share a frequently mutated sub-set of genes (the core-component) that is

supposed to lead the pathway enrichments, together with a membership matrix specifying to which enriched pathway each core-component gene belongs to (an example is provided in **Error! Reference source not found.**B, introduced in the next section). This allows filtering out from the results those pathways that are not directly relevant to the disease under consideration, in a supervised way. A final feature of the package is the identification of pathways that are differentially enriched (thus frequently altered) across two sub-populations of samples of the same input dataset, as detailed in the following sections.

S2. Heuristic mutual exclusivity sorting and pathway visualization

The set of somatic mutations of a cancer genomic dataset can be easily modeled as a binary (or Boolean) matrix, whose entries can assume only two possible values, i.e. 0 or 1. In this case, the columns indicate samples, its rows indicate genes (or vice-versa) and a non-zero entry the presence of a somatic mutations in a given gene/sample combination. In a binary matrix, a run is a sequence of consecutive non-zero entries. Reordering rows and columns in a way that the number of runs on the rows and the column-wise marginal totals are minimized is an effective way to highlight patterns of mutual exclusivity among the runs of different rows, i.e. the genes of the considered sub-set. This is an NP-hard problem [2] here referred as *mutual-exclusivity sorting*. In SLAPenrich a heuristic implementation of the mutual-exclusivity sorting is provided in a dedicated R function used by the internal visualization routines, although this function is also available and usaåble on any user defined binary matrix. Here, for simplicity we will describe an execution of this heuristic applied to a binary matrix summarizing a genomic dataset (with genes on the rows, samples on the columns, and binary entries specifying the status of a gene in a given sample).

In the initial step of the algorithm all the samples and all the genes in the input matrix are declared as *uncovered* and an empty vector is initialized: this is the set of *covered* genes ***G***. Then the algorithm proceeds through a series of iterations until the sets of uncovered genes and uncovered samples are both empty. In each of these iterations a *best in class gene* is identified. This is the uncovered gene with the maximal exclusive coverage, which is defined as the number of uncovered samples in which this gene is mutated minus the number of samples in which at least another uncovered gene is mutated. Finally, the identified best in class gene is removed from the set of the uncovered genes, it is attached to ***G***, and the set of samples in which it is mutated are removed from the set of the uncovered samples.

After these iterations have been executed, an empty vector of samples ***L*** is initialized and all the samples of the dataset are labeled again as uncovered. Then for each of the best in class gene ***g*** (in the same order as they appear in ***G***) and until there are uncovered samples, the uncovered samples in which ***g*** is mutated are sorted according to the exclusive coverage of ***g*** across them (in decreasing ordered), they are labeled as covered samples and attached in the resulting order to ***L***.

To obtain the final mutual-exclusivity sorting of the initial dataset, the corresponding inputted binary matrix is rearranged by permuting the genes/rows in the same order as they appear in ***G*** and the samples/columns in the same order as they appear in ***L***.

S3. Identification and visualization of enriched pathway core-components

To identify shared core-components across significantly enriched pathways, the set of enriched pathways and their composing genes are modeled as a bipartite network, in which nodes in the first set correspond to enriched pathways and nodes in the second set to genes belonging to at least one of the enriched pathways. Finally a pathway node is connected with an edge to each of its composing gene nodes. The resulting bipartite network is then mined for communities, i.e. groups of densely interconnected nodes, by using a fast community detection algorithm based on a greedy strategy [3]. The resulting communities are finally visualized as independent heatmaps where nodes in the first set (pathways) are on the columns, nodes in the second set (genes) are on the rows and a not-empty cell in position ***i,j*** indicates that the ***i***-th gene belongs to the ***j***-th pathway (an example is provided in **Error! Reference source not found.**B).

S.4 Differential pathway enrichment analysis

Similarly to differential gene expression analysis, the two sub-populations to be contrasted are defined through a contrast matrix. Then individual SLAPenrichment analyses are performed on these two populations, yielding two sets of results. The pathways that are significantly enriched in at least one of the two analyses (according to a user defined false discovery rate (FDR) threshold) are then selected and, for each of them, a differential enrichment score is computed as:

$$\Delta_{A,B}(P) = -\log_{10} FDR_{A(P)} + \log_{10} FDR_{B(P)}$$

where A and B are the two contrasted sub-populations (respectively, positive and negative) and $FDR_{A(P)}$ and $FDR_{B(P)}$ are the two SLAPenrichment FDRs obtained in the two corresponding individual analyses, and P is the pathway under consideration. Graphic routines included in our package allow a pathway level visualization of the inputted alterations across the two contrasted population, on the domain of the differentially enriched pathways as well as heatmaps and barplots of the differential enrichment scores (see an example in **Error! Reference source not found.**C).

S.5 LUAD case study analysis

To test the ability of our method in recovering pathways that are known to be associated to given a disease state and different clinico-pathological features, we have re-analysed, using different reference pathway collections, a published dataset encompassing somatic mutations found in 188 lung adenocarcinoma (LUAD) patients, studied in [4], downloading annotations of somatic variants and associated clinical information from http://genome.wustl.edu/pub/supplemental/tsp_nature_2008/ (files: supplementary_table_2.tsv and supplementary_table_15.tsv, respectively).
The variants annotations were converted into a genomic event matrix (EM) with altered genes on the rows, patient sample identifiers on the columns, and generic *i*,*j* entries specifying the number of observed point mutations hosted by the *i*-th gene in the *j*-th patient.

A first SLAPenrich analysis on the resulting dataset was performed using the `SLAPE.analyse` function with default values for all the parameters (including a Bernoulli model [5] for the individual pathway alteration probabilities across all the samples, and the choice of the set of all the altered genes in the dataset as background population), and a pathway gene sets collection from KEGG [6] (embedded in the package as R data object: `SLAPE.20160211_MSigDB_KEGG_hugoUpdated`).
This analysis yielded 48 significantly enriched pathways, at a FDR < 5% and a mutual exclusive coverage (EC) > 50% (**Error! Reference source not found.**). Among these, we found pathways whose deregulation is known to be involved in lung cancer, such as *tight junction* (alteration score (AS) = 0.37, EC = 89%) [7] (**Error! Reference source not found.**A), *gap junction* (AS = 0.45, EC = 75%) [8], and several pathways found with PathScan [9] and other computational methods [4], such as for example *focal adhesion* (AS = 0.06, EC = 84%), *ERBB signaling pathway* (AS = 0.27, EC = 69%), *dorsoventral axis formation* (AS = 0.42, EC = 55%). Additionally, we found a number of pathways recently proposed as potential targets for lung cancer therapy such as *GNRH signaling pathway* (AS = 0.45, EC = 87%) [10], *WNT signaling pathway* (AS = 0.29, EC = 74%) [11], and VEGF signaling pathway (AS = 0.33, EC = 80%) [12].

To further validate the ability of SLAPenrich in identifying disease relevant pathways and highlight the possible analytical venues allowed by our tool, we considered the clinical information of the samples in the analyzed LUAD dataset. Using this data, we stratified the considered patients based on their smoking status (never-smoker and current-smokers) and their bronchioalveolar carcinoma type (mucinous and non-mucinous), and performed a differential SLAPenrich analysis contrasting the variant profiles of the obtained sub-populations, using the far larger publicly available collection of pathway gene sets from Pathway Commons [1], post-processed for redundancy removal as described in the Methods. Outcomes from the first analysis, comparing never-smoker vs. current-smokers, are reported in Supplementary Table S2 and summarized in **Error! Reference source not found.**C. In total we found 147 differentially enriched pathways (enriched at FDR < 5% in at least one of the two populations). Ranking these pathways according to their differential enrichment score, in decreasing order (**Error! Reference source not found.**C)

highlights, consistently with previously reported findings, in the current-smokers population a prominent enrichment of alterations in the RAS/RAF/MEK signaling cascade [13], telomerase activity [14], NOXA and PUMA signaling[15]. On the other hand, in the never-smoker population we observed prominent enrichments in EGFR signaling and EGFR-dependent endothelin signaling pathways [16].

When contrasting mucinous vs. non-mucinous BAC types (Supplementary Figure S2 and Supplementary Table S3), we again observed correct associations between the mucinous BAC type and pathway alteration enrichments in the RAS/RAF/MEK signaling cascade [17], signaling by leptin[18], PI3K and MTOR signaling pathways [19], and inflammation related pathways such as CXCR3 and GM-CSF mediated signaling. Whereas for the non-mucinous BAC type population prominent enrichments were observed in pathways involving EGFR signaling[20]. The presented analyses and results are fully detailed in the vignette of the SLAPenrich package (see Code Availability).

S.6 Comparison with other methods

To our knowledge there are only two public available tools performing analyses of pathway alterations enrichments in large genomic datasets at the sample population level, and implementing a statistical framework similar to that of SLAPenrich: PathScan [9] and PathScore [21]. While these tools, in particular PathScore, share aspects with SLAPEnrich, a number of features of these two tools make them unsuitable for the analyses described in our manuscript.

Pathscan, even if, like SLAPEnrich, computes aggregated p-values at the sample population level, these are still obtained by merging together enrichment p-values computed at the individual sample level. Additionally, PathScan does not take into account of possible mutual exclusivity trends between patterns of mutations of genes in the same pathway. Finally, in more practical terms, it requires raw sequencing data (BAM files) in input: this is quite uncomfortable for our case making use of public available processed genomic datasets represented through binary presence/absence matrices.

PathScore uses the same mathematical model as SLAPEnrich, but the individual pathway mutation probabilities are computed with a fixed model, using published estimated mutation rates that cannot be changed. In contrast, SLAPEnrich uses a Bernoulli model with customisable mutation rates (which can be also estimated looking at the analysed cohort of patients itself). Furthermore, PathScore is not implemented as a stand-alone tool but as web-application only, and there are not APIs available yet to integrate it in other computational pipelines, or to customize its execution parameters.

Furthermore, both PathScan and PathScore make use of pathway collections from public available repositories (KEGG [6] for PathScan, MsigDB [22] for PathScore). We use a larger pathway collection (2,794 pathways, covering 15,281 genes, against 186 pathways and 5,224 genes for Pathscan, and 1,329 pathways and 8,904 genes for PathScore) from Pathway Commons [1]. Additionally we post-processed this collection for redundancy reduction: pathways with large overlaps are merged together instead of being tested individually. This is a unique feature of our tool, it avoids similar gene-sets to be tested multiple

times and produces a non-redundant mapping between pathways and cancer hallmarks (as detailed in Figure 3).

We report below results for a comparison of results obtained with our tool with respect to PathScan and PathScore both.

Collectively, we found a significant agreement between our results and those obtained with PathScan (and reported in the Supplementary Table 1 of [9]). After applying the same result curation of [9], i.e. removal of known cancer pathways whose mutation lists are invariably collectively dominated by mutations in *TP53*, *KRAS* and *EGFR*, and considering the 129 remaining KEGG pathway, we found 26 enriched pathways (FDR < 5% for both SLAPenrich and PathScan), out of 36 pathways enriched for SLAPenrich and 31 enriched for PathScan (at the same FDR threshold), Fisher exact test p-value = 2.10×10^{-14} (**Error! Reference source not found.**). Additionally, we observed a significant correlation (R = 0.66, p = 0.0002) between the significance levels of the 26 commonly enriched pathways across the two methods (Supplementary Figure S3).

Similarly, we performed a comparison between the output obtained with SLAPenrich and PATHscore [21] when analysing the LUAD dataset described above. To this aim, in order to obtain comparable results we downloaded the whole collection of 1,392 *canonical pathway signatures* from the Molecular Signature Database (MsigDB) [22], as this is the reference collection used by the PATHscore online tool. We performed a SLAPenrich analysis of the LUAD dataset (coded as an EM, as detailed above) using this reference collection of pathway gene sets as input, and a PATHscore analysis using the online tool available at http://pathscore.publichealth.yale.edu/, and the list of variants of the LUAD dataset coded as required by this PATHscore (included in **Error! Reference source not found.**). As with PAthScan, we observed a high and significance overlap (181 enriched pathways, Fisher exact test p-value = 1.76×10^{-70}) between the 198 enriched pathways outputted by SLAPenrich (at an FDR < 5%) and the 479 outputted by PATHscore (adjusted p-value < 0.05), Supplementary Table S5. As most of the significantly enriched pathways outputted by PATHscore have a null p-value it was not possible to check the correlation between the patterns of enrichment significance across the two methods. However when looking at the top enriched pathways across the two analyses (SLAPenrich FDR = 1.76×10^{-12} and PATHscore adjusted p-value = 0) the results' concordance was even more pronounced (100 overlapping pathways out of the 117 outputted by SLAPenrich and the 176 outputted by PATHscore, Fisher exact test p-value = 8.63×10^{-83}), Supplementary Table S5.

Supplemental References

1. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2011;39:D685–90.

2. Johnson D, Krishnan S, Chhugani J, Kumar S. Compressing large boolean matrices using reordering techniques. 2004.

3. Newman MEJ. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69:066133.

4. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008;455:1069–75.

5. Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics. 2011;27:175–81.

6. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.

7. Soini Y. Tight junctions in lung cancer and lung metastasis: a review. Int J Clin Exp Pathol. 2012;5:126–36.

8. Guy S, Geletu M, Arulanandam R, Raptis L. Stat3 and gap junctions in normal and lung cancer cells. Cancers (Basel). 2014;6:646–62.

9. Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, et al. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics. 2011;27:1595–602.

10. Baudot AIS, Torre V de L, Valencia A. Mutated genes, pathways and processes in tumours. EMBO reports. 2010;11:805.

11. Yang J, Chen J, He J, Li J, Shi J, Cho WC, et al. Wnt signaling as potential therapeutic target in lung cancer. Expert Opin. Ther. Targets. 2016;:1–17.

12. Aita M, Fasola G, Defferrari C, Brianti A, Bello MGD, Follador A, et al. Targeting the VEGF pathway: antiangiogenic strategies in the treatment of non-small cell lung cancer. Crit. Rev. Oncol. Hematol. 2008;68:183–96.

13. Larsen JE, Minna JD. Molecular biology of lung cancer: clinical implications. Clin. Chest Med. 2011;32:703–40.

14. Yim HW, Slebos RJC, Randell SH, Umbach DM, Parsons AM, Rivera MP, et al. Smoking is associated with increased telomerase activity in short-term cultures of human bronchial epithelial cells. Cancer Lett. 2007;246:24–33.

15. Sakakibara-Konishi J, Oizumi S, Kikuchi J, Kikuchi E, Mizugaki H, Kinoshita I, et al. Expression of Bim, Noxa, and Puma in non-small cell lung cancer. BMC Cancer. 2012;12:286.

16. Pirie K, Peto R, Green J, Reeves GK, Beral V, Million Women Study Collaborators. Lung cancer in never-smokers. Int. J. Cancer. 2016.

17. Finberg KE, Sequist LV, Joshi VA, Muzikansky A, Miller JM, Han M, et al. Mucinous differentiation correlates with absence of EGFR mutation and presence of KRAS mutation in lung adenocarcinomas with bronchioloalveolar features. J Mol Diagn. 2007;9:320–6.

18. Woo H-J, Yoo WJ, Bae CH, Song S-Y, Kim Y-W, Park S-Y, et al. Leptin up-regulates MUC5B expression in human airway epithelial cells via mitogen-activated protein kinase pathway. Exp. Lung Res. 2010;36:262–9.

19. Raina D, Kosugi M, Ahmad R, Panchamoorthy G, Rajabi H, Alam M, et al. Dependence on the MUC1-C oncoprotein in non-small cell lung cancer cells. Molecular Cancer Therapeutics. 2011;10:806–16.

20. Sakuma Y, Matsukuma S, Yoshihara M, Nakamura Y, Noda K, Nakayama H, et al. Distinctive evaluation of nonmucinous and mucinous subtypes of bronchioloalveolar carcinomas in EGFR and K-ras gene-mutation analyses for Japanese lung adenocarcinomas: confirmation of the correlations with histologic subtypes and gene mutations. Am. J. Clin. Pathol. 2007;128:100–8.

21. Gaffney SG, Townsend JP. PathScore: a web tool for identifying altered pathways in cancer data. Bioinformatics. 2016.

22. Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of

America. 2005;102:15545.

23. Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell. 2015;27:382–96.