## METHOD

# Population-level characterization of pathway alterations with SLAPenrich dissects heterogeneity of cancer hallmark acquisition

Francesco Iorio[1,5†], Luz Garcia-Alonso[1,5], Jonathan Brammeld[2], Iñigo Martincorena[2], David R Wille[3,5], Ultan McDermott[2,5] and Julio Saez-Rodriguez[1,4,5*†]

### Abstract

We present SLAPenrich, a statistical approach implemented in an R package to search for pathways enriched for genomic alterations in large datasets. SLAPenrich mines for sample-population level enrichments, accounting for mutation rates, gene exonic lengths, and mutational mutual exclusivity. We show that SLAPenrich detects in lung adenocarcinoma pathways known to be typically altered, therapeutic targets, and associations with clinicopathological features. Finally, we explore with SLAPenrich the landscape of pathways contributing to the acquisition of the cancer hallmarks in large cohorts of patients across 10 cancer types, highlighting potential novel cancer driver genes and networks.

**Keywords:** cancer-genomics; hallmark-analysis; pathways; populations; mutual-exclusivity; enrichment-analysis

## Background

The swift progression of next-generation sequencing technologies is enabling a fast and affordable production of an extraordinary amount of genome sequences. Cancer research is particularly benefiting from these advances, and comprehensive catalogues of somatic mutations involved in carcinogenesis, tumour progression and response to therapy are becoming increasingly available and ready to be exploited for the identification of new diagnostic, prognostic and therapeutic markers [1, 2, 3, 4]. Exploration of the genomic makeup of multiple cancer types has highlighted that driver somatic mutations typically involve few genes altered at high frequency and a long tail of more genes mutated at very low frequency [5, 6], with a tendency for both sets of genes to code for proteins involved into a limited number of biological processes [7]. As a consequence, a reasonable approach is to consider these alterations by grouping them based on prior knowledge of the cellular mechanisms and biological pathways where the products of the mutated genes operate [8]. This facilitates the identification of the spectrum of possible alterations underpinning a common cancer hallmark

[9, 10], which can include genes that would be otherwise not considered on the basis of their (low) alteration frequency alone. Additionally, a pathway-based description reduces the dimensionality of large genomic datasets involving thousands of altered genes into a sensibly smaller set of altered mechanisms that are more interpretable, and possibly actionable in a pharmacological or experimental way [11].

### Existing methods

Several computational methods have been designed to perform pathway analysis on genomic data, aiming at prioritizing sets of genomically altered genes whose products operate in the same cellular process or functional network. All the approaches proposed so far toward this aim can be classified into two main classes [8].

The first class of approaches aims at identifying pathways whose composing genes are significantly over-represented in the set of altered genes across all the samples of a dataset compared against the background set of all studied genes. Generally, a Fisher's exact test is used to calculate the statistical significance of this over-representation. Many tools exist and are routinely used to perform this analysis [12, 13, 14],

sometimes incorporating additional features, such as accounting for inter-gene dependencies and signal correlations [15]. More sophisticated methods take into account the differences in pathway mutation probabilities expected by random chance [16], variations in gene lengths, and combine p-values from tests on individual samples into multiple-sample p-values [17].

The second class of approaches aims at identifying novel pathways by mapping alteration patterns on large protein networks. The combinatorial properties occurring among the alterations are then analyzed and used to define cost functions, for example based on the tendency of a group of genes to be mutated in a mutual exclusive manner. On the basis of these cost functions, optimal sub-networks are identified and interpreted as novel cancer driver pathways [18, 19, 20]. However, at the moment there is no consensual way to rigorously define a mathematical metric for mutual exclusivity and compute its statistical significance, and a number of interpretations exist [18, 19, 21, 22, 23].

### SLAPenrich

Here we present SLAPenrich (Sample Level Analysis of Pathway alteration Enrichments): a computational method implemented into an R package to perform pathway analyses at the population level. We propose this tool as a mean to characterize in an easily interpretable way sparse somatic mutations detected in heterogeneous cancer sample populations, which share traits of interest and are subjected to strong selective pressure, leading to combinatorial patterns. SLAPenrich does not require somatic mutations in a pathway to be statistically enriched among those detected in each sample nor the merged (or aggregated) set of mutations in the population. Relying on the mutual exclusivity principle [24], SLAPenrich assumes that a single mutation involving a node of a pathway in an individual sample can be sufficient to deregulate the activity of the pathway, providing selective growth advantages. Hence, SLAPenrich belongs roughly to the first category described above, although it shares the mutual exclusivity consideration with the methods in the second.

SLAPenrich includes a visualization/report framework allowing an easy exploration of the outputted enriched pathways across the analyzed samples, in a way that highlights their mutual exclusivity mutation trends, and a module for the identification of core-components genes, shared by related enriched pathways.

We illustrate SLAPenrich in a case-study where we analyzed somatic mutations across lung adenocarcinoma (LUAD) patient samples from a public dataset [25]. In addition, we used SLAPenrich to characterise across 4,415 patients covering 10 cancer types the contributions of the underlying somatic mutations to the acquisition of the canonical cancer hallmarks [9, 10]. This novel 'hallmark heterogeneity analysis' has the potential to provide a data-driven evaluation of each hallmark across disease types, to explore how these hallmarks are achieved via different pathways, and to identify novel putative cancer driver genes and networks.
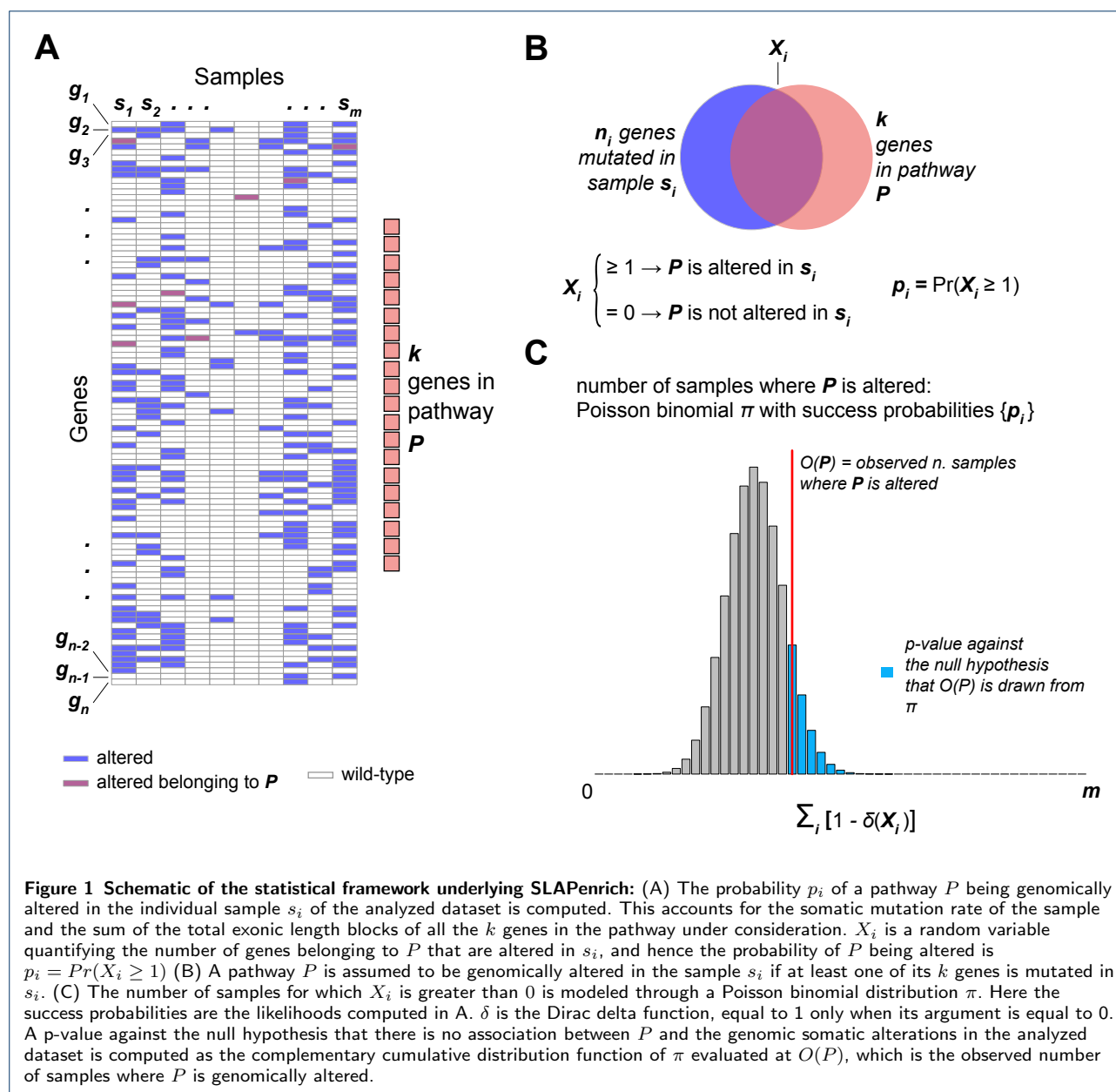
## Results

### Summary of the approach and analytical features

SLAPenrich assumes that an individual point mutation in a gene belonging to a given pathway can be sufficient to deregulate the activity of that pathway in the sample under consideration. More precisely, after modeling the probability of a genomic alteration in at least one member of a given pathway across the individual samples, a collective statistical test is performed against the null hypothesis that the number of samples with at least one alteration in that pathway is close to that expected by random chance, therefore no association exists between the analyzed population and the pathway under consideration. An additional advantage of modeling probabilities of at least an individual mutation in a given pathway, instead of modeling the probability of the actual number of mutated genes, is that this prevent signal saturations due to hypermutator samples.

The input to SLAPenrich is a collection of samples accounting for the mutational status of a set of genes, such as a cohort of human cancer genomes. This is modeled as a dataset where each sample consists of a somatic mutation profile indicating the status (point-mutated or wild-type) of a list of genes (Figure 1A). For a given biological pathway $P$, each sample is considered as an individual Bernoulli trial that is successful when that sample harbors somatic mutations in at least one of the genes belonging to the pathway under consideration (Figure 1B). An additional advantage of modeling probabilities of at least an individual mutation in a given pathway (instead of, for example, the probability of the actual number of mutated genes) is that this prevent signal saturations due to hypermutated samples.

The probability of success of each of these trials is computed by either (i) a general hypergeometric model

**Figure 1 Schematic of the statistical framework underlying SLAPenrich:** (A) The probability $p_i$ of a pathway $P$ being genomically altered in the individual sample $s_i$ of the analyzed dataset is computed. This accounts for the somatic mutation rate of the sample and the sum of the total exonic length blocks of all the $k$ genes in the pathway under consideration. $X_i$ is a random variable quantifying the number of genes belonging to $P$ that are altered in $s_i$, and hence the probability of $P$ being altered is $p_i = Pr(X_i \geq 1)$ (B) A pathway $P$ is assumed to be genomically altered in the sample $s_i$ if at least one of its $k$ genes is mutated in $s_i$. (C) The number of samples for which $X_i$ is greater than 0 is modeled through a Poisson binomial distribution $\pi$. Here the success probabilities are the likelihoods computed in A. $\delta$ is the Dirac delta function, equal to 1 only when its argument is equal to 0. A p-value against the null hypothesis that there is no association between $P$ and the genomic somatic alterations in the analyzed dataset is computed as the complementary cumulative distribution function of $\pi$ evaluated at $O(P)$, which is the observed number of samples where $P$ is genomically altered.

accounting for the mutation burden of the sample under consideration, the size of the gene background population and the number of genes in the pathway under consideration, or (ii) a more refined modeling of the likelihood of observing point mutations in a given pathway, accounting for the total exonic block lengths of the genes in that pathway (Figure 1AB) and the estimated (or actual) mutation rate of the sample under consideration [26]. In addition, more sophisticated methods, accounting for example for gene sequence compositions, trinucleotide rates, and other covariates (such as expression, chromatin state, etc) can be used, through user-defined functions that can be easily inte-

grated in SLAPenrich.

Once these probabilities have been computed, the expected number of samples in the population harboring at least one somatic mutation in $P$ can be estimated, and its probability distribution modeled analytically. Based on this, a pathway alteration score can be computed observing the deviance of the number of samples harboring somatic mutations in $P$ from its expectation, and its statistical significance quantified analytically (Figure 1C). Finally, the resulting statistically enriched pathways can be further filtered by looking at the tendency of their composing genes

to be mutated in a mutually exclusive fashion across all the analyzed samples, as an additional evidence of positive selection [27, 18, 19]. A formal description of the statistical framework underlying SLAPenrich is provided in the Methods.

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 in the Methods) that minimizes the superposition of mutated samples column-wisely, thus the overlaps of the mutation patterns across the rows (an example is provided in Figure 2A and described in the next section). To finally summarize the results, an analysis of the enriched-pathway *core-component genes* is 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 Figure 2B, as described in the next section). This allows to filter 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. SLAPenrich is implemented as an R package (available at https://github.com/francescojm/SLAPenrich, submission to Bioconductor is in progress). It includes different collections of pathway gene sets from multiple public available sources [28], together with all the data objects needed to run the analysis, but can be used with any user-defined collection of gene-sets. An overview of the exposed functions of the R package is provided in additional file 7.

### A first case study

To show the ability of SLAPenrich to recover pathways known to be frequently altered in a given disease, and the usefulness of its differential pathway enrichment analysis, we have re-analysed a published dataset encompassing somatic mutations found in 188 lung adenocarcinoma (LUAD) patients, studied in [25].

A SLAPenrich analysis of this dataset, using the entire collection of pathways downloaded from KEGG [29] (post-processed for redundancy removal as described in the Methods), resulted into 48 significantly enriched pathways, at a FDR < 5% and a mutual exclusive coverage (EC) > 50% (additional file 1). 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%) [30] (Figure 2A), *gap junction* (AS = 0.45, EC = 75%) [31], and several pathways found with other methods [25, 17], 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 not reported in the aforementioned studies but recently proposed as potential targets for lung cancer therapy such as *GNRH signaling pathway* (AS = 0.45, EC = 87%) [32], *WNT signaling pathway* (AS = 0.29, EC = 74%) [33], and *VEGF signaling pathway* (AS = 0.33, EC = 80%) [34].

Considering the clinical information of the samples in the analyzed dataset we stratified the corresponding 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 [28], post-processed for redundancy removal as described in the Methods. Outcomes from the first analysis, comparing never-smoker vs. current-smokers, are reported in the additional file 2 and summarized in Figure 2C. 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 (Figure 2C) highlights, consistently with previously reported findings, in the current-smokers population a prominent enrichment of alterations in the RAS/RAF/MEK signaling cascade [35], telomerase activity [36], NOXA and PUMA signaling [37]. On the other hand, in the never-smoker population we observed prominent enrichments in EGFR signaling and EGFR-dependent endothelin signaling pathways [38].

When contrasting mucinous vs. non-mucinous BAC types (additional file 3 and additional figure 1), we again observed correct associations between the mucinous BAC type and pathway alteration enrichments in the RAS/RAF/MEK signaling cascade [39], signaling by leptin [40], PI3K and MTOR signaling pathways [41], and inflammation related pathways such as
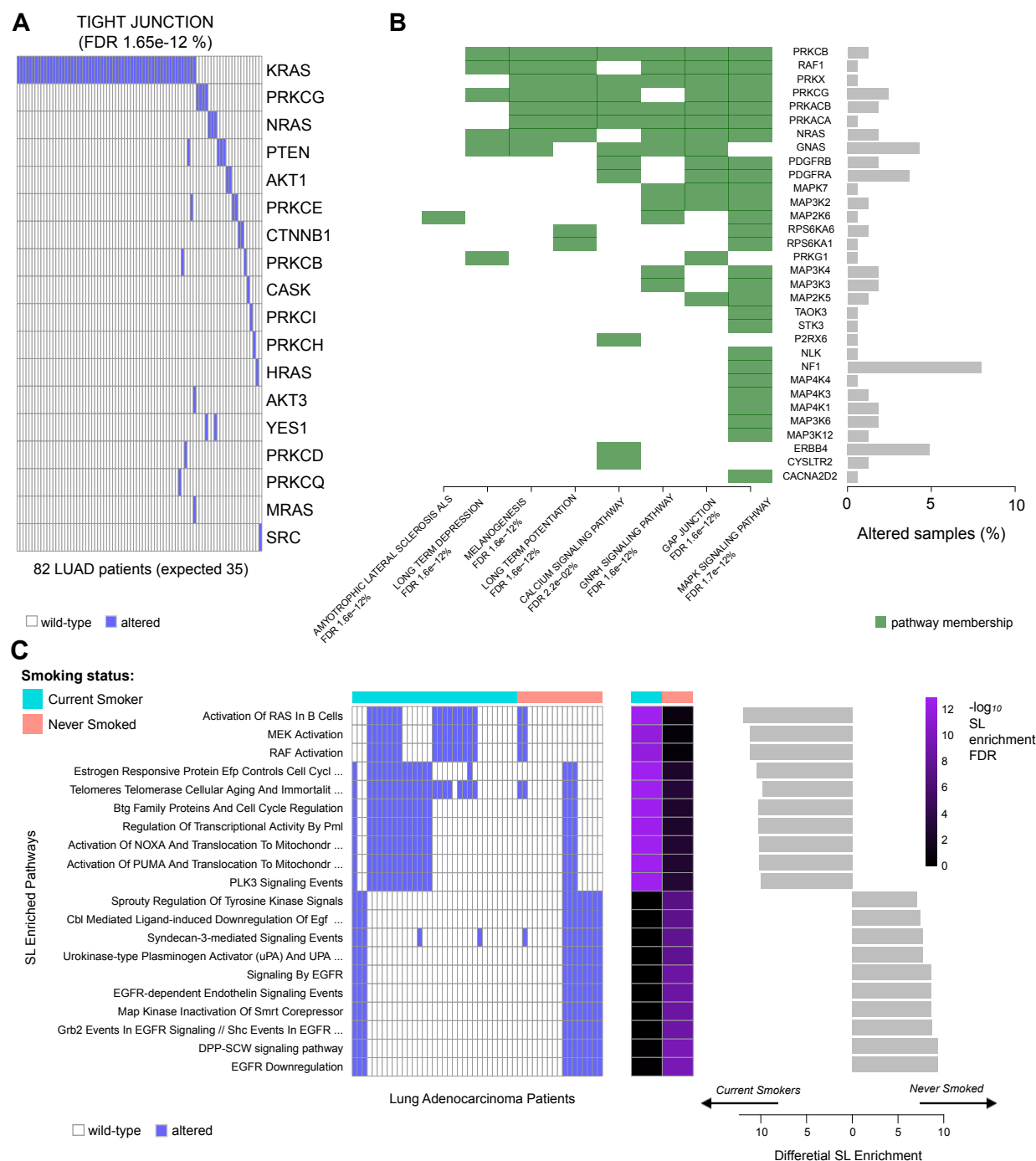
**Figure 2 Visualization of enriched pathways, core-components and differential pathway enrichment analysis:** (A) Heatmap summarizing the status of the genes belonging to a pathway enriched at the population level in the case study of the lung adenocarcinoma dataset. Genes and patient samples (respectively on rows and columns) have been permuted with a dedicated function in order to highlight mutual exclusivity trends in the observed somatic alterations. (B) Heatmap showing a sub-set of genes (on the rows) shared by multiple significantly enriched pathways (on the columns), together with a bar plot diagram (on the right) showing the percentages of patient samples where each gene is altered. These figures are automatically generated by SLAPenrich. (C) Visual output of the differential enrichment analysis function using the case study lung adenocarcinoma dataset in input, and stratifying patients based on their smoking status. The heatmap on the left shows the alteration status of the top/bottom 10 most positively/negatively differentially enriched pathways between the groups of smokers vs non-smokers (on the column); the heatmap in the centre shows enrichment significance of individual pathways in the two sub-populations, and the barplot shows corresponding differential sample level enrichment scores.

CXCR3 and GM-CSF mediated signaling. Whereas for the non-mucinous BAC type population prominent enrichments were observed in pathways involving EGFR signaling [42]. The presented analyses and results are fully detailed in the vignette of the SLAPenrich package (see Code Availability).

## SLAPenrich analyses across different cancer types highlights the heterogeneity of cancer hallmark acquisition

SLAPenrich can be used to systematically analyze large cohorts of cancer genomes providing a data-driven exploration of mutated pathways that can be easily compared across cancer types. Additionally, the format of the results allows a wide range of novel investigations at a higher level of abstraction.
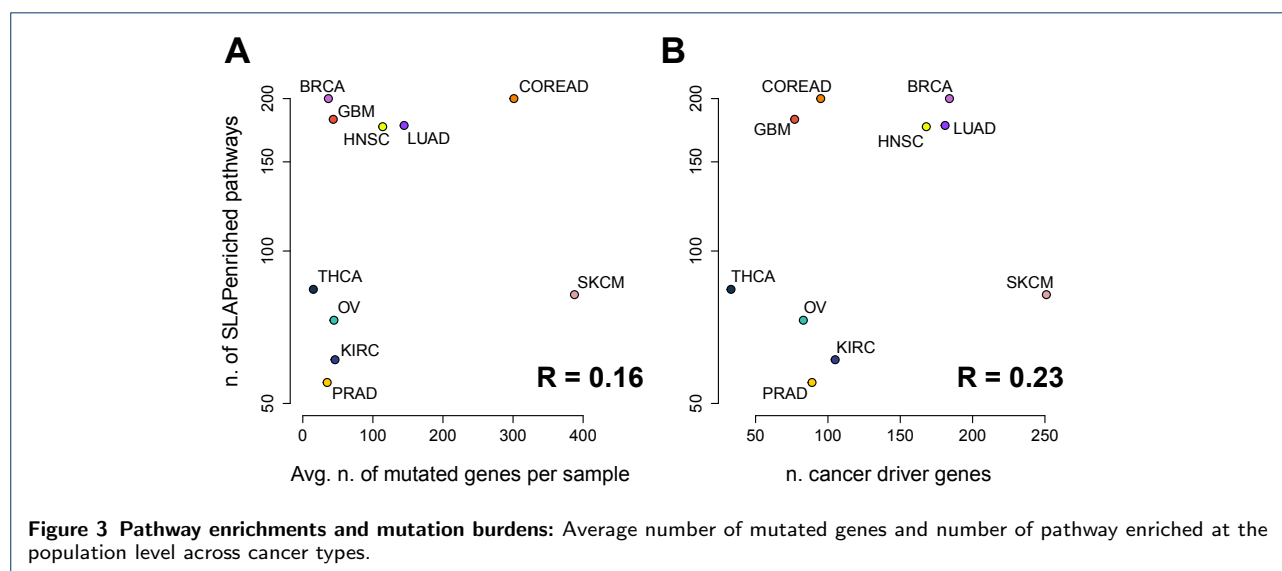
As an illustration of this, we performed individual SLAPenrich analyses on 10 different genomic datasets containing somatic point mutations, preprocessed as described in [43], from 4,415 patients across 10 different cancer types, from public available studies, in particular The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). These samples (see methods) comprise breast invasive carcinoma (BRCA, 1,132 samples), colon and rectum adenocarcinoma (COREAD, 489), glioblastoma multiforme (GBM, 365), head and neck squamous cell carcinoma (HNSC, 375), kidney renal clear cell carcinoma (KIRC, 417), lung adenocarcinoma (LUAD, 388), ovarian serous cystadenocarcinoma (OV, 316), prostate adenocarcinoma (PRAD, 242), skin cutaneous melanoma (SKCM, 369), and thyroid carcinoma (THCA, 322). We observed a weak correlation ($R = 0.53$, $p = 0.11$) between the number of enriched pathways across the different analyses and the number of available samples in the analysed dataset (additional Figure 3A), but a down-sampled analysis showed that our results are not broadly confounded by the sample sizes (see methods and additional Figure 3B). Results from all these individual SLAPenrich analyses are contained in the additional file 4.

We investigated how our pathway enrichments capture known tissue specific cancer driver genes. To this aim, we used a list of high-confidence and tissue-specific cancer driver genes [43, 44] (from now high-confidence Cancer Genes, HCGs, assembled as described in the Methods). We observed that the majority of the HCGs was contained in at least one SLAPenriched pathway, across the 10 different tissues analyses (median percentage = 63.5, range = 88.5%, for BRCA, to 28.7% for SKCM) (additional Figure 3C).

Interestingly, we found that the number of SLAPenriched pathways per cancer type (median = 130, range

= 55 for PRAD, to 200 for BRCA and COREAD) was independent from the average number of mutated genes per sample across cancer types (median = 46, range from 15 for THCA to 388 for SKCM) with a Pearson correlation $R = 0.16$ ($p = 0.65$), Figure 3A, as well as from the number of high confidence cancer driver genes (as predicted in [44], median = 100, range from 33 for THCA to 251 for SKCM, Figure 3B). Particularly, THCA has the lowest average number of mutations per sample (15.03), but there are 4 tissues with a lower number of pathways mutated. In contrast, SKCM has the highest average number of point mutations per sample (387.63), but the number of affected pathways is less than half of those of BRCA and GBM (82 enrichments against an average of 191), which have on average less than 100 mutations per sample (Figure 3). GBM, OV, KIRC, PRAD and BRCA are relatively homogeneous with respect to the average number of somatic mutations per sample (mean = 41.03, from 34.76 for KIRC to 45.95 for PRAD) but when looking at the number of enriched pathways for this set of cancer types we can clearly distinguish two separate groups (Figure 3). The first group includes BRCA and GBM that seem to have a more heterogeneous sets of processes impacted by somatic mutations (average number of SLAPenriched pathways = 191) with respect to the second group (63 SLAPenriched pathways on average). These results suggest that there is a large heterogeneity in the number of processes deregulated in different cancer types that is independent of the mutational burden. This might be also indicative of different subtypes with dependencies on different pathways (and at least for BRCA this is expected) but could be also biased by the composition of the analysed cohorts being representative of a selected subtype only.

Subsequently, we reasoned that since the main role of cancer driver alterations is to enable cells to achieve a series of phenotypic traits called the 'cancer hallmarks' [9, 10], that can be linked to gene mutations [45], it would be informative to group the pathways according to the hallmark they are associated with. Towards this end, through a computer aided manual curation (see Methods and additional file 5) we were able to map 374 gene-sets (from the most recent release of pathway commons [28]) to 10 cancer hallmarks [9, 10] (additional Figure 2AB), for a total number of 3,915 genes (included in at least one gene set associated to at least one hallmark; additional file 6). The vast majority (99%, 369 sets) of the considered pathway gene-sets were mapped on two hallmarks at most, and 298 of them (80%) was mapped onto one single hallmark (additional Figure 2C). Regarding the individual genes contained in at least one pathway

**Figure 3 Pathway enrichments and mutation burdens:** Average number of mutated genes and number of pathway enriched at the population level across cancer types.

gene-set, about half (49%) were associated with a single hallmark, 22% with two, 12% with three, and 7% with four(additional Figure 2D). Finally, as shown in additional Figure 2E, the overlaps between the considered pathway gene-sets was minimal (74% of all the possible pair-wise Jaccard indexes was equal to 0 and 99% < 0.2). In summary, our manual curation produced a non-redundant matching in terms of both pathways- and genes-hallmarks associations.

Mapping pathway enrichments into canonical cancer hallmarks through this curation allowed us to explore how different cancer types might acquire the same hallmark by selectively altering different pathways. examples are provided in Figure 4, and additional Figure 4. Heatmaps in these figures (one per each hallmark) show different level of enrichments of pathways associated to the same hallmark across different tissues.

We investigated at what extent the identified enriched pathways were dominated by somatic mutations in the high-confidence cancer genes (HCGs) [44], across cancer types. To this aim, for each pathway $P$ enriched in a given cancer type $T$, we computed an HCG-dominance score as the ratio between the number of $T$ samples with mutations in HCGs belonging to $P$ and the total number of $T$ samples with mutations in any of the gene belonging to $P$. Results of this analysis are shown in additional figures 5 and 6. We observed a median of 15% of pathway enrichments, across hallmarks, with an HCG-dominance score < 50%, thus not led by somatic mutations in HCGs (range from 9% for *Deregulating Cellular Energetics* to 21% for *Genome Instability and Mutation*). Additionally, a median of 3% of pathway enrichments had a null HCG-

dominance, thus not involved somatic mutations in HCGs (range from 0.25% for *Evading Growth Suppression* to 15% for *Avoiding Immune Destruction*). Across all the hallmarks, the cancer type with the lowest median HCG-dominance was KIRC (33%), whereas that with the highest was THCA (91%).

Patterns and well defined clusters can be clearly distinguished in the heatmaps of Figure 4. Patterns and well defined clusters can be clearly distinguished. As an example, in Figure 4 the heatmap related to the *Genome Instability and mutation* hallmark shows that BRCA, OV, GBM, LUAD and HNSC might achieve this hallmark by selectively altering a group of pathways related to homologous recombination deficiency, whose prevalence in BRCA and OV is established [46]. This deficiency has been therapeutically exploited recently and translated into a clinical success thanks to the introduction of PARP inhibition as a very selective therapeutic option for these two cancer types[47]. Pathways preferentially altered in BRCA, OV, GBM, LUAD and HNSC include *G2/M DNA Damage Checkpoint // Processing Of DNA Double Strand Break Ends*, *TP53 Regulates Transcription Of DNA Repair Genes*, and other signaling networks related to BRCA1/2 and its associated RING Domain 1 (BARD1). Conversely, the *Androgen receptor* pathway, known to regulate the growth of glioblastoma multiforme (GBM) in men [48] is also exclusively and preferentially altered in this cancer type. The acquisition of the *Genome Instability and mutation* hallmark seems to be dominated in COREAD by alterations in the *HDR Through Single Strand Annealing (SSA), Resolution Of D Loop Structures Through Synthesis Dependent Strand Annealing (SDSA), Homolo-*

**Figure 4 Heterogeneity of hallmark acquisition across cancer types:** Heatmaps showing pathways enrichments at the population level across cancer types for individual hallmarks (representative cases). Color intensities correspond to the enrichment significance. Cancer types and pathways are clustered using a correlation metric. See also additional figure 4.

*gous DNA Pairing And Strand Exchange* and other pathways more specifically linked to a microsatellite instability led hypermutator phenotype, known to be prevalent in this cancer type [49]. Finally, the heatmap for *Genome Instability and Mutation* shows nearly no enriched pathways for SKCM. This is consistent with the high burden of mutations observed in melanoma originating from cell extrinsic processes such as UV light exposure [50]. The maintenance of genomic integrity is guarded by a network of damage sensors, signal transducers, and mediators, and it is regulated through changes in gene expression. Recent studies show that miRNAs play a crucial role in the response to UV radiation in skin cells [51]. Our analysis strikingly detects *MiRNAs Involved In DNA Damage Response* as the unique pathway associated to *Genome instability and mutation* enriched in SKCM. This suggests that mutations in this pathway, involving ATM (as top frequently mutated gene, and known to induce miRNA biogenesis following DNA damage [52]), impair the ability of melanocytes to properly respond to insult from UV light and may have a significant role in the tumourigenesis of melanoma.

The *Avoiding Immune destruction* heatmap (Figure 4) highlights a large number of pathways selectively enriched in COREAD, wheareas very few pathways associated to this hallmark are enriched in the other analysed cancer types. This could explain why immunotherapies, such as PD-1 inhibition, have a relatively low response rate in COREAD when compared to, for example, non-small cell lung cancer [53], melanoma [54] or renal-cell carcinoma [55]. In fact, response to PD-1 inhibition in COREAD is limited to tumours with mismatch-repair deficiency, perhaps due to their high rate of neoantigen creation [56]. In the context of COREAD, the *Tumor-promoting inflammation* heatmap (Figure 4) also highlights several pathways predominantly and very specificically altered in this cancer type. Chronic inflammation is a proven risk factor for COREAD and studies in animal models have shown a dependency between inflammation, tumor progression and chemotherapy resistance [57]. Indeed, a number of clinical trials evaluating the utility of inflammatory and cytokine-modulatory therapies are currently underway in colorectal cancer [58, 59]. Interestingly, according to our analysis this hallmark is acquired by SKCM by exclusively preferentially altering IRF3 related pathways.

Several other examples would be worthy of mention. For example, the detection of the *Warburg effect* pathway contributing to the acquisition of the *Deregulating cellular energetics* hallmark in GBM only (Figure 4).
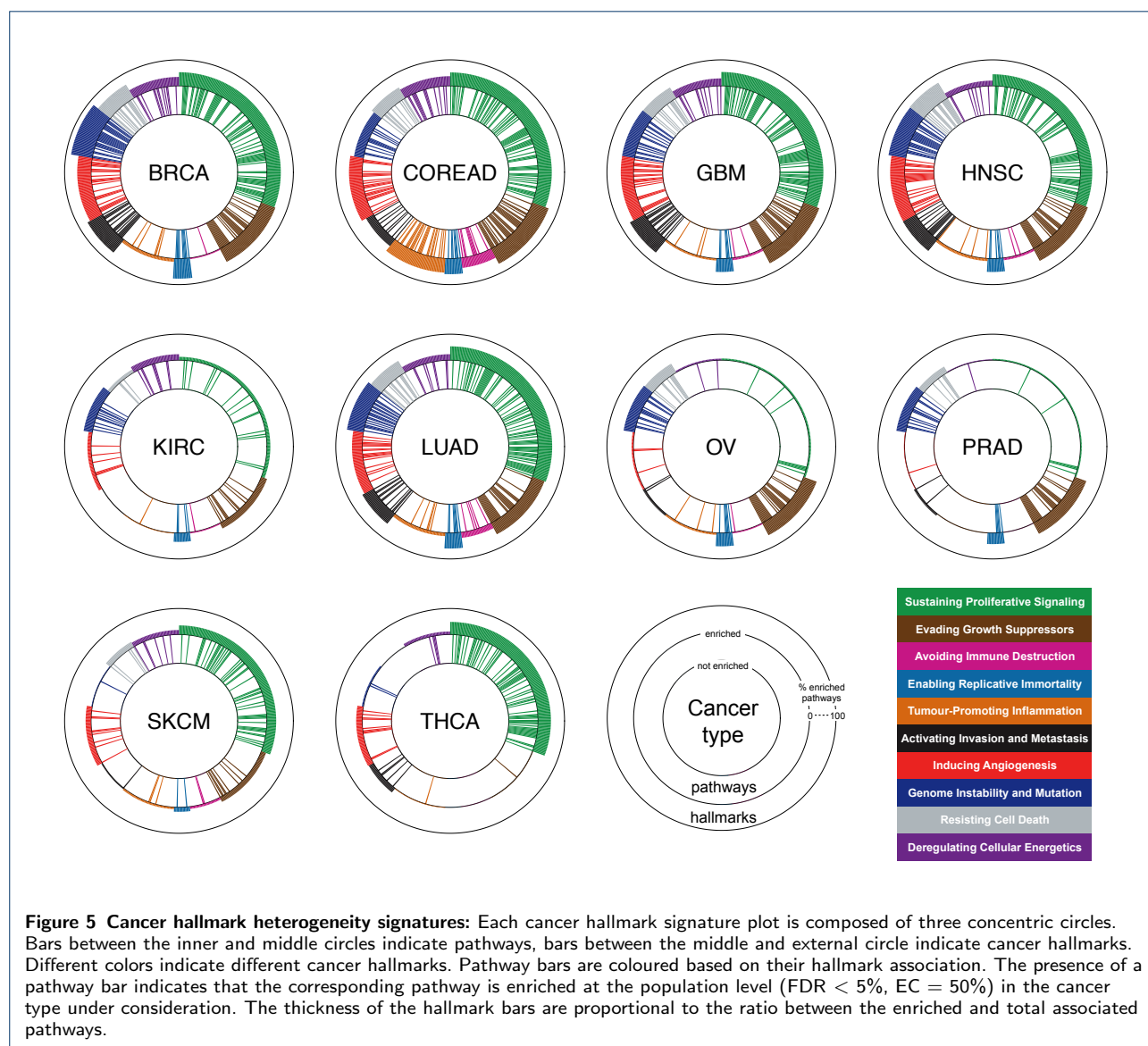
The Warburg effect is a unique bioenergetic state of aerobic glycolysis, whose reversion has been recently proposed as an effective way to decrease GBM cell proliferation [60]. Additionally, the pathway *Formation of senescence associated heterochromatin*, associated to the *Enabling replicative immortality* hallmark is enriched in multiple cancer types. Genomic alterations in this pathway have not been linked to cancer so far. More interestingly the enrichment of this pathway, across cancer types, is not driven by any established cancer gene.

Finally, we quantified the diversity of pathways used to achieve each hallmark in a given cancer type, via a cumulative heterogeneity score (CHS) computed as the proportion of the pathways associated to that hallmark that are enriched. The larger this score the more a given cancer type relies on altering a large number of pathways in order to achieve the considered hallmark. A larger heterogeneity of pathways, in turn, could point to the exploitment of more evolutionary trajectories (reflected by selecting genomic alterations in a large number of associated pathways). As a consequence, the larger this score the higher might be the evolutionary fitness of that hallmark for the cancer type under consideration. Joining the CHSs of all the hallmarks resulting from the analysis of a given cancer type, gives its *hallmark heterogeneity signature*. The hallmark heterogeneity signatures of all the 10 analysed cancer types are reported in Figure 5. Results show consistency with the established predominance of certain hallmarks in determined cancer types, such as for example a high CHS for *Genome instability and mutation* in BRCA and OV [61], for *Tumour-promoting inflammation* and *Avoiding immune-destruction* in COREAD [62]. Lastly, and as expected for *Sustaining proliferative-signaling* and *Enabling replicative immortality*, the key hallmarks in cancer initiation [9], high CHSs are observed across the majority of the analysed cancer types.

Taken together, these results show the potential of SLAPenrich to perform systematic landscape analyses of large cohorts of cancer genomes. In this case this is very effective in highlighting commonalities and differences in the acquisition of the cancer hallmarks across tissue types, confirming several known relations between cancer types, and pinpointing preferentially altered pathways and hallmark acquisitions.

### Hallmark heterogeneity analysis points at potential novel cancer driver genes and networks

To investigate the potential of SLAPenrich in identifying novel cancer driver genes and networks we performed a final analysis (from now the *filtered analysis*) after removing all the variants involving, for each

**Figure 5  Cancer hallmark heterogeneity signatures:** Each cancer hallmark signature plot is composed of three concentric circles. Bars between the inner and middle circles indicate pathways, bars between the middle and external circle indicate cancer hallmarks. Different colors indicate different cancer hallmarks. Pathway bars are coloured based on their hallmark association. The presence of a pathway bar indicates that the corresponding pathway is enriched at the population level (FDR < 5%, EC = 50%) in the cancer type under consideration. The thickness of the hallmark bars are proportional to the ratio between the enriched and total associated pathways.

considered cancer type, the corresponding HCGs. Results of this exercise (Figure 6 and additional figure 7), showed that the majority of the enrichments identified in the original analyses (on the unfiltered genomic datasets) were actually led by alterations in the HCGs (consistent with their condition of high reliable cancer genes). The average ratio of retained enrichments in the filtered analyses across cancer types (true positive rates (TPRs) in Figure 6 and additional figure 7) was 21%, (range from 2.1% for GBM to 56.2% for COREAD). However, several pathway enrichments (some of which did not include any HCGs) were still detected in the filtered analysis and, most importantly, the corresponding hallmark heterogeneity signatures were largely conserved between the filtered and unfiltered analyses for most of the cancer types, with

coincident top fitting hallmarks and significantly high over-all correlations (Figure 6, additional figure 7). If the hallmark signatures from the original unfiltered analyses are faithful representations of the mutational landscape of the analysed cancer types and the filtered analyses still detect this landscape despite removal of known drivers, then the filtered analyses might have uncovered novel cancer drivers in the long tail of infrequently mutated genes. In fact, these new gene modules are typically composed by groups of functionally interconnected and very lowly frequently mutated genes (Figure 7).

An example is represented by the pathway *Activation Of Matrix Metalloproteinases* associated with the *Invasion and metastasis* hallmark and highly en-
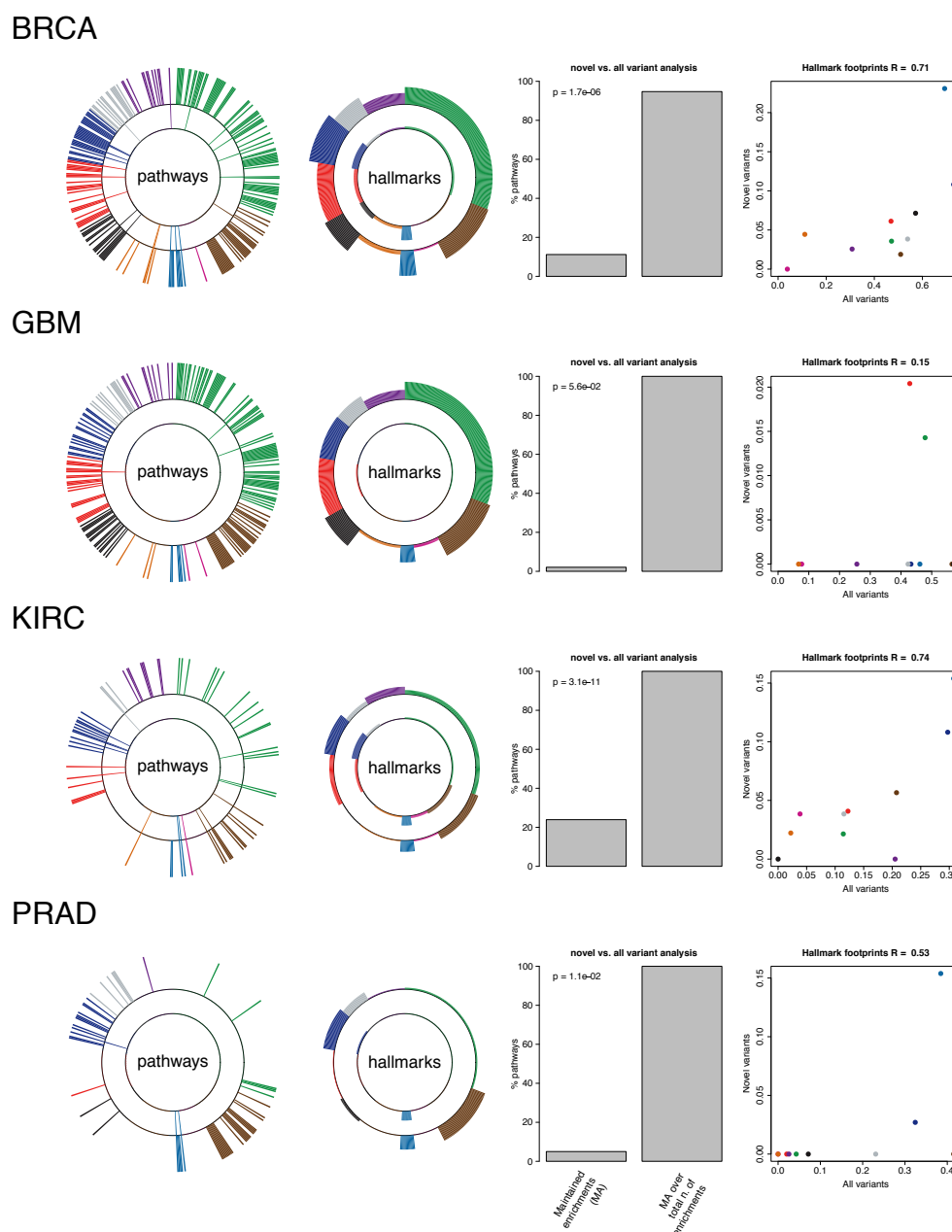
**Figure 6 Hallmark heterogeneity signature analysis with and without known drivers:** In each row, the first circle plot show pathway enrichments at the population level when considering all the somatic variants (bars on the external circle) and when considering only variants not involving known high-confidence cancer driver genes (internal circle); the second circle plot compares the hallmark signatures resulting from SLAPenrich analysis including (bars on the external circle) or excluding (bars on the internal circle) the variants involving known high-confidence cancer genes. The bar plot shows a comparison, in terms of true-positive-rate (TPR) and positive-predicted-value (PPV), of the SLAPenriched pathways recovered in the filtered analysis vs. the complete analysis., The scatter plots on the right show a comparison between the resulting hallmark signatures.

riched in the filtered analyses of COREAD (FDR = 0.002%), SKCM (0.09%) (Figure 7A), LUAD (0.93%), and HNSC (3.1%). The activation of the matrix metalloproteases is an essential event to enable the migration of malignant cells and metastasis in solid tumors [63]. Although this is a hallmark acquired late in the evolution of cancer, according to our analysis this pathway is still detectable as significantly enriched. As a consequence, looking at the somatic mutations of its composing genes (of which only Matrix Metallopep-

tidase 2 - MMP2 - has been reported as harboring cancer driving alterations in LUAD [44]) might reveal novel key components of this pathway leading to metastatic transitions. Interestingly, among these, one of the top frequently mutated genes (across all the 4 mentioned cancer types) is Plasminogen (PLG), whose role in the evolution of migratory and invasive cell phenotype is established [64]. Furthermore, blockade of PLG with monoclonal antibodies, DNA-based vaccination or silencing through small interfering RNAs has been recently proposed to counteract cancer invasion and metastasis [65]. The remaining altered component of this pathway is mostly made of a network of very lowly frequently mutated (and in a high mutual exclusive manner) other metalloproteinases.

Another similar example is given by the *IL 6 Type Cytokine Receptor Ligand Interactions* pathway significantly enriched in the filtered analysis of SKCM (FDR = 4.6%) and associated with the *Tumour-promoting inflammation hallmark* (Figure 7B). IL-6-type cytokines have been observed to modulate cell growth of several cell types, including melanoma [66]. Increased IL-6 blood levels in melanoma patients correlate with disease progression and lower response to chemotherapy [67]. Importantly, studies proposed OSMR, a IL-6-type of cytokine receptor, to play a role in the prevention of melanoma progression [68], and as a novel potential target in other cancer types [69]. Consistently with these findings, OSMR is the member of this pathway with the largest number of mutations in the SKCM cohort (Figure 7B), complemented by a large number of other lowly frequently mutated genes (most of which are interleukins).

In the context of melanoma, we observed other two pathways highly enriched in the filtered analysis: *PDGF receptor signaling network* (FDR = 2.7%) (Figure 7C) and *Neurophilin Interactions with VEGF And VEGFR* (0.21%)(Figure 7D), both associated with the *Inducing angiogenesis* hallmark. Mutations in all the components of these two pathways are not common in SKCM and have not been highlighted in any genomic study so far. The first of these two pathway enrichments is characterised by patterns of highly mutual exclusive somatic mutations in Platelet-derived growth factor (PDGF) genes, and corresponding receptors: a network that has been recently proposed as an autocrine endogenous mechanism involved in melanoma proliferation control [70].

A final example is given by the enriched pathway *Regulating the activity of RAC1* (associated with the *Activating Invasion and Metastasis* hallmark)

in COREAD (Figure 7E). The Ras-Related C3 Botulinum Toxin Substrate 1 (RAC1) gene is a member of the Rho family of GTPases, whose activity is key for cell motility [71]. Previous *in vitro* and *in vivo* studies in prostate cancer demonstrated a marked increase in RAC1 activity in cell migration and invasion, and that RAC1 inhibition immediately stopped these processes [72, 73]. However, although the role of RAC1 in enabling metastasis has already been suggested, the mechanisms underlying such aberrant behaviour are poorly understood, and our findings could be used as a starting point for further investigations [74].

Another interesting case is the high level of mutual exclusivity observed in the mutation patterns involving members of the *TP53 network*, highly enriched in the filtered analysis of SKCM, encompassing TP63, TP73, TNSF10, MYC and SUMD1 (Figure 7F). Whereas alterations in some nodes of this network are known to be an alternative to p53 repression, conferring chemoresistance and poor prognosis [75], dissecting the functional relations between them is still widely considered a formidable challenge [76]. Our results point out alternative players worthy to be looked at in this network (particularly, among the top frequently altered, TNSF10).

Taken together, these results suggest that SLAPenrich can help to identify potential novel cancer driver genes and cancer driver networks composed by lowly frequently mutated genes.

## Discussion

In this paper we present SLAPenrich, a statistical framework implemented in an open source R package to identify pathway enrichment in genomic datasets at the population level. SLAPenrich does not seek for pathways whose alterations are enriched at the individual sample level nor at the global level, i.e. considering the union of all the genes altered in at least one sample. Instead, it assumes that an individual mutation involving a given pathway in a given sample might be sufficient to deregulate the activity of that pathway in that sample and it allows enriched pathways to be mutated in a mutual exclusive manner across samples.

The SLAPenrich package includes (i) fully tunable functions where statistical significance criteria and alternative models, can be defined by the user; (ii) a visualization and reporting framework, and (iii) accessory functions for data management and gene identifier curation and cross-matching. Worthy of note is that many different tools provide the possibility of visualizing a mutual-exclusivity sorted sets of somatic
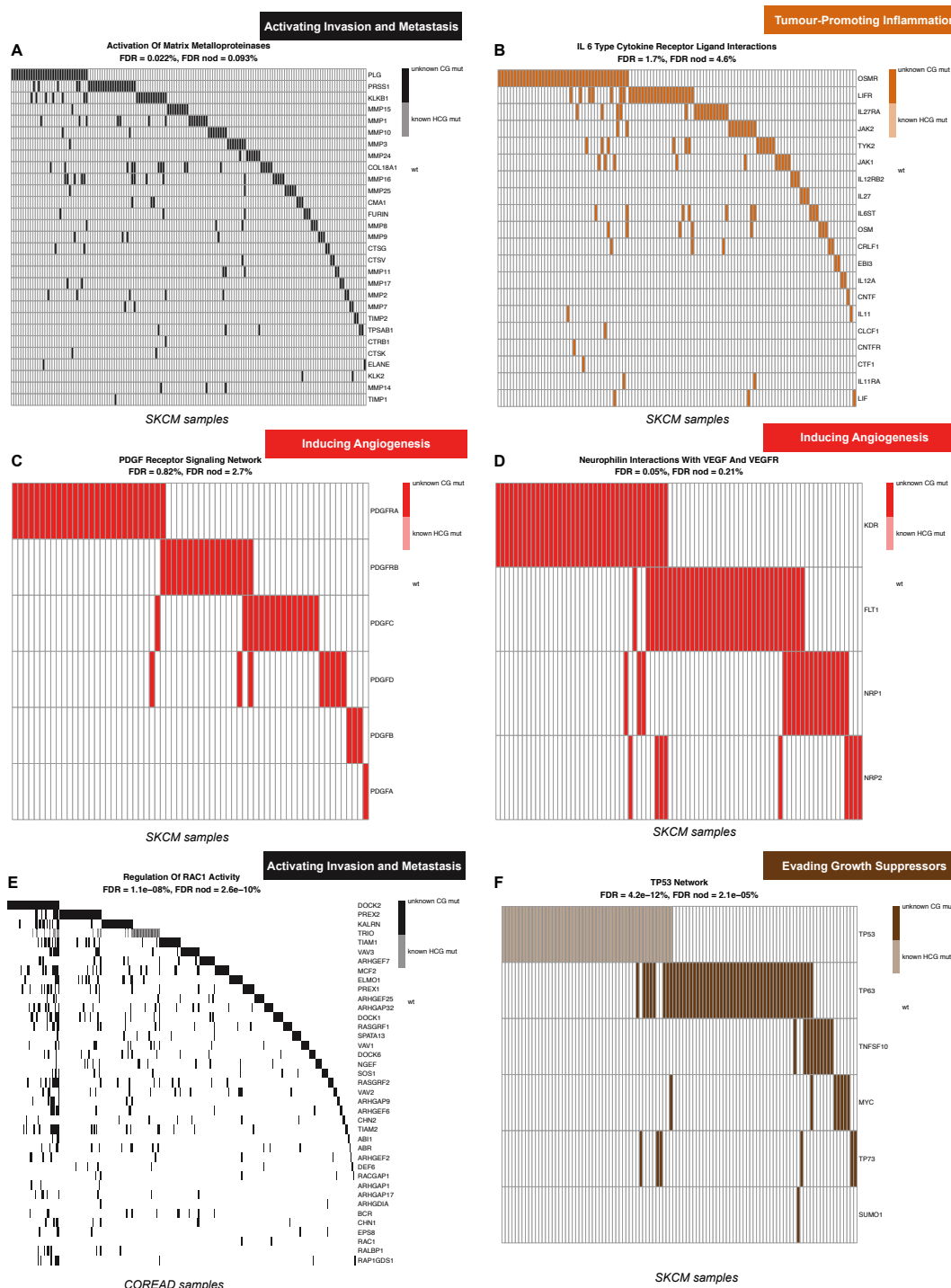
**Figure 7 Example of putative novel cancer genes and networks identified by SLAPenrich** Picked examples of novel putative cancer driver genes and networks. The first FDR value refers to the unfiltered analysis, whereas the second FDR refers to the filtered one.

mutations and other genomic alterations from publicly available or user defined datasets via a browser accessible software suite (e.g. GiTools [77] and cBioPortal [78]) or as a result of combinatorial pattern analysis (such as MEMo [79] and Dendrix [19]). However, none of these tools offer this feature as a mean to visualise

an arbitrarily defined data matrix and, to our knowledge, there is no publicly available R implementation for this.

Beyond extraction of pathway enrichments, as illustrated with our case study on a lung cancer data set, SLAPenrich can also be used to perform large-scale comparative analysis of cancer mutational landscapes. We used this feature to perform a characterization of the mutational status of cancer hallmarks. The obtained results provided a first data-driven landscape of hallmark acquisitions through the preferential alteration of heterogenous sets of pathways across cancer types, confirming established hallmark predominancies, and detecting peculiar pattern of altered pathways for certain cancer types. Finally, by using the identified hallmark signatures as the gold-standard signal, we re-performed SLAPenrich analyses after the removal of variants in established cancer genes and highlighted genes that contributed to those hallmarks as potential novel cancer driver genes and networks.

A number of possible limitations could hamper deriving definitive conclusions from this paper, such as the use of only mutations, the possibility that some of the analysed cohorts of patients are representative only of well-defined disease subtypes, or the limitation of our knowledge of pathways. Nevertheless, we provide the community with a useful tool for the analysis of large genomic datasets, whose produced results (as in our hallmark analysis presented here) could open a wide range of novel in-silico investigations.

## Conclusions

SLAPenrich should be of wide usability for the functional characterization of sparse genomic data from heterogeneous populations sharing common traits and subjected to strong selective pressure. As an example of its applicability we have studied the large cohorts of publicly available cancer genomes patient data that is publicly available in the TCGA. However, SLAPenrich could be of great utility in other scenarios such as the characterizion of genomic data generated upon chemical mutagenesis to identify somatic mutations involved in acquired drug resistance [Brammeld et al, under revision http://dx.doi.org/10.1101/066555]. More generally, SLAPenrich can be used to characterize at the pathway level any type of biological dataset that can be modeled as a presence/absence matrix, where genes are on the rows and samples are on the columns.

## Methods

### Formal description of the SLAPenrich statistical framework

Let us consider the list of all the genes $G = \{g_1, g_2, \ldots, g_n\}$, whose somatic mutational status has been determined across a population of samples $S = \{s_1, s_2, \ldots, s_m\}$, and a function $f : G \times S \rightarrow \{0, 1\}$ defined as

$$f(g, s) = \{1 \text{ if } g \text{ is mutated in } s, \text{ and } 0 \text{ otherwise}\}. \tag{1}$$

Given the set of all the genes whose products belong to the same pathway $P$, we aim at assessing if there is a statistically significant tendency for the samples in $S$ to carry mutations in $P$. Importantly, we do not require the genes in $P$ to be significantly enriched in those that are altered in any individual sample nor in the sub-set of $G$ composed by all the genes harbouring at least one somatic mutation in at least one sample. In what follows $P$ will be used to indicate the pathway under consideration as well as the corresponding set of genes, interchangeably. We assume that $P$ is altered in sample $s_i$ if $\exists g \in G$ such that $g \in P$ and $f(g, s_i) = 1$, i.e. at least one gene in the pathway $P$ is altered in the $i$-th sample (Figure 1B). To quantify how likely it is to observe at least one gene belonging to $P$ altered in sample $s_i$, we introduce the variable $X_i = |\{g \in G : g \in P \text{ and } f(g, s_i) = 1\}|$, accounting for the number of genes in $P$ altered in sample $s_i$. Under the assumption of both a gene-wise and sample-wise statistical independence, the probability of $X_i$ assuming a value greater or equal than 1 is given by:

$$p_i = \Pr(X_i \geq 1) = \sum_{x=1}^{k} H(x, N, k, n_i), \tag{2}$$

where $N$ is the size of the gene background-population, $k$ is the number of genes in $P$, $n_i$ is the total number of genes $g$ such that $f(g, s_i) = 1$, i.e. the total number of genes harbouring an alteration in sample $s_i$, and $H$ is the probability mass function of a hypergeometric distribution:

$$H(x, N, k, n_i) = \frac{\binom{k}{x}\binom{N-k}{n_i-x}}{\binom{N}{n_i}}. \tag{3}$$

To take into account the impact of the exonic lengths $\lambda(g)$ of the genes $(g)$ on the estimation of the alteration probability of the pathway they are part of $P$, it is possible to redefine the $p_i$ probabilities (of observing

at least one genes in the pathway $P$ altered in sample $s_i$) as follows:

$$p_i = \ \Pr(X_i \geq 1) = \sum_{x=1}^{k} H(x, N', k', n_i'), \qquad (4)$$

where $N' = \sum_{g \in G} \lambda(g)$, with $G$ the gene background-population, i.e. the sum of all the exonic content block lengths of all the genes; $k' = \sum_{g \in P} \lambda(g)$ is the sum of the exonic block length of all the genes in the pathway $P$; $n'$ is the total number of individual point mutations involving genes belonging to $P$ in sample $s_i$, and $H$ is defined as in equation 3, but with parameters $x, N', k'$, and $n_i'$. Similarly, the $p_i$ probabilities can be modeled accounting for the total exonic block lengths of all the genes belonging to $P$ and the expected/observed background mutation rate [26], as follows:

$$p_i = \ \Pr(X_i \geq 1) = 1 - \exp(-\rho k'), \qquad (5)$$

where $k'$ is defined as for equation 4 and $\rho$ is the background mutation rate, which can be estimated from the input dataset directly or set to established estimated values (such as $10^{-6}$/nucleotide)[26].
If considering the event "the pathway $P$ is altered in sample $s_i$" as the outcome of a single test in a set of Bernoulli trials $\{i\}$ (with $i = 1, \ldots, M$) (one for each sample in $S$), then each $p_i$ can be interpreted as the success probability of the $i - th$ trial. By definition, summing these probabilities across all the elements of $S$ (all the trials) gives the expected number of successes $E(P)$, i.e. the expected number of samples harbouring a mutation in at least one gene belonging to $P$:

$$E(P) = \sum_{i=1}^{M} p_i. \qquad (6)$$

On the other hand, if we consider a function $\phi$ on the domain of the $X$ variables, defined as $\phi(X) = 1 - \delta(X)$, where $\delta(X)$ is the Dirac delta function (assuming null value for every $X \neq 0$), i.e. $\phi(X) = \{1$ if $X > 0$, and 0 otherwise$\}$, then summing the $\phi(X_i)$ across all the samples in $S$, gives the observed number of samples harbouring a mutation in at least one gene belonging to $P$:

$$O(P) = \sum_{i=1}^{M} \phi(X_i). \qquad (7)$$

A pathway alteration index, quantifying the deviance of $O(P)$ from its expectation, and thus how surprising is to find so many samples with alterations in the pathway $P$, can be then quantified as:

$$\Delta(P) = \log_{10} \frac{O(P)}{E(P)}. \qquad (8)$$

To assess the significance of such deviance, let us note that the probability of the event $O(P) = y$, with $y \leq M$, i.e. the probability of observing exactly $y$ samples harbouring alterations in the pathway $P$, distributes as a Poisson binomial $B$ (a discrete probability distribution modeling the sum of a set of $\{i\}$ independent Bernoulli trials where the success probabilities $p_i$ are not identical (with $i = 1, \ldots, M$). In our case, the $i$-th Bernoulli trial accounts for the event "the pathway $P$ is altered in the sample $s_i$" and its success probability of success is given by the $\{p_i\}$ introduced above (and computed with one amongst 2, 4, or 5). The parameters of such $B$ distribution are then the probabilities $\pi = \{p_i\}$, and its mean is given by Equation 6. The probability of the event $O(P) = y$ can be then written as

$$\Pr(O(P) = y) = B(\pi, y) = \sum_{A \in F_y} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_i), \qquad (9)$$

where $F_y$ is the set of all the possible subsets of $y$ elements that can be selected from the trial $1, 2, \ldots, M$ (for example, if $M = 3$, then $F_2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$, and $A^c$ is the complement of $A$, i.e. $\{1, 2, \ldots, M\} \backslash A$. Therefore a p-value can be computed against the null hypothesis that $O(P)$ is drawn from a Poisson binomial distribution parametrised through the vector of probabilities $\pi$. Such p-value can be derived for an observation $O(P) = z$, with $z \leq M$, as (Figure 1C):

$$\Pr(O(P) \geq z) = \sum_{i=z}^{M} \Pr(O(P) = i) = \sum_{i=z}^{M} B(\pi, i) \quad (10)$$

Finally, p-values resulting from testing all the pathways in the considered collection are corrected for multiple hypothesis testing with a user-selected method among (in decreasing order of stringency) Bonferroni, Benjamini-Hochberg, and Storey-Tibshirani [80].

## Pathway gene sets collection and pre-processing

To highlight the versatility of SLAPenrich and guarantee results' comparability with respect to previously published studies, we have conducted the analyses described in the Results section using different collections of pathway gene sets, all included (as R objects) in our software package.

For the case study analysis on the LUAD dataset we downloaded the whole collection of KEGG [29] pathway gene sets from MsigDB [81], encompassing 189 gene sets for a total number of 5,224 genes included in at least one set.

The following differential enrichment analyses and the hallmark signature analyses were performed on a larger collection of pathway gene sets from the Pathway Commons data portal (v8, 2016/04) [28] (http://www.pathwaycommons.org/archives/PC2/v4-201311/). This contained an initial catalogue of 2,794 gene sets (one for each pathway) that were assembled from multiple public available resources, such as Reactome [82], Panther [83], HumanCyc [84], pid [85], smpdb [86], KEGG [29], ctd [87], inoh [88], wikipathways [89], netpath [90], and mirtarbase [91], and covering 15,281 unique genes.

From this pathway collection, those gene sets containing less than 4 or more than 1,000 genes, were discarded. Additionally, in order to remove redundancies, those gene sets (*i*) corresponding to the same pathway across different resources or (*ii*) with a large overlap (Jaccard index $(J) > 0.8$, as detailed below) were merged together by intersecting them. The gene sets resulting from this compression were then added to the collection (with a joint pathway label) and those participating in at least one of these merging were discarded. Finally, gene names were updated to their most recent HGCN [92] approved symbols (this updating procedure is also executed by a dedicate function in of the SLAPenrich package, by default on each genomic datasets prior the analysis). The whole process yielded a final collection of 1,911 pathway gene sets, for a total number of 1,138 genes assigned to at least one gene set.

Given two gene sets $P_1$ and $P_2$ the corresponding $J(P_1, P_2)$ is defined as:

$$J(P_1, P_2) = \frac{|P_1 \cap P_2|}{|P_1 \cup P_2|}. \qquad (11)$$

## Curation of a pathway/hallmark map

We implemented a simple routine (included in the SLAPenrich R package) that assigns to each of the 10 canonical cancer hallmarks a subset of the pathways in a given collection. To this aim this routine searches for determined keywords (typically processes or cellular components) known to be associated to each hallmark in the name of the pathway (such as for example: 'DNA repair' or 'DNA damage' for the *Genome instability and mutations* hallmark) or for key nodes in the set of included genes or key word in their name prefix (such as for example 'TGF', 'SMAD', and 'IFN' for *Tumour-promoting inflammation*. The full list of keywords used in this analysis are reported in the additional file 5. Results of this data curation are reported in the additional file 6.

## Mutual exclusivity filter

After correcting the p-values yielded by testing all the pathways in a given collection, the enriched pathways can be additionally filtered based on a mutual exclusivity criterion, as a further evidence of positive selection. To this aim, for a given enriched pathway $P$, an exclusive coverage score $C(P)$ is computed as

$$C(P) = 100\frac{O'(P)}{O(P)} \qquad (12)$$

where $O(P)$ is the number of samples in which at least one gene belonging to the pathway $P$ is mutated, and $O'(P)$ is the number of samples in which exactly one gene belonging to the pathway gene-set $P$ is mutated. All the pathways $P$ such that $C(P)$ is at least equal to a chosen value pass this final filter.

## 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. $\{0, 1\}$. 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 [93] here referred as the '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 usable 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, all the samples of the dataset are labeled again as uncovered, an empty vector of samples $S$ is initialized. 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 $L$.

### 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 [94]. 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 (Figure 2 B).

### 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:

$$\Delta_{A,B}(P) = -\log_{10} FDR_A(P) + \log_{10} FDR_B(P) \quad (13)$$

where $A$ and $B$ are the two contrasted sub-populations (respectively, positive and negative) and $FDR_A$ and $FDR_B$ are the two SLAPenrichment FDRs obtained in the two corresponding individual analyses. 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 Figure 2C).

### LUAD case study analysis

Annotations of somatic variants identified in a cohort of 188 lung adenocarcinoma patients, and associated clinical information, were downloaded 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 SLAPenrich analysis on the resulting dataset was performed using the SLAPE.analyse function with default values for all the parameters (including a Bernoulli model [26] 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 [29] (embedded in the package as R data object:
SLAPE.20160211_MSigDB_KEGG_hugoUpdated).

### Hallmark heterogeneity signature analysis: genomic datasets and high-confidence cancer genes

Tissue specific catalogues of genomic variants for 10 different cancer types (breast invasive carcinoma, colon and rectum adenocarcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, kidney renal clear cell carcinoma, lung adenocarcinoma,

ovarian serous cystadenocarcinoma, prostate adenocarcinoma, skin cutaneous melanoma, and thyroid carcinoma) were downloaded from the GDSC1000 data portal described in [43] (`http://www.cancerrxgene.org/gdsc1000/`). This resource (available at `http://www.cancerrxgene.org/gdsc1000/`) encompasses variants from sequencing of 6,815 tumor normal sample pairs derived from 48 different sequencing studies [44] and reannotated using a pipeline consistent with the COSMIC database [95] (Vagrent: `https://zenodo.org/record/16732#.VbeVY2RViko`). Lists of tissue specific high-confidence cancer genes [44] were downloaded from the same data portal (`http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources//Data/suppData/TableS2A.xlsx`) These were identified by combining complementary signals of positive selection detected through different state of the art methods [96, 97] and further filtered as described in [43] (`http://www.cell.com/cms/attachment/2062367827/2064170160/mmc1.pdf`)

### Hallmark heterogeneity signature analysis: Individual SLAPenrich analysis parameters

All the individual SLAPenrich analyses were performed using the `SLAPE.analyse` function of the SLAPenrich R package (`https://github.com/francescojm/SLAPenrich`) using a Bernoulli model for the individual pathway alteration probabilities across all the samples, the set of all the genes in the dataset under consideration as background population, selecting pathways with at least one gene point mutated in at least 5% of the samples and at least 2 different genes with at least one point mutation across the whole dataset, and and a pathway gene sets collection downloaded from pathway commons[28], post-processed for redundancy reduction as explained in the previous sections, and embedded in the SLAPE package as R data object:

`PATHCOM_HUMAN_nonredundant_intersection_hugoUpdated.RData`.

A pathway in this collection was considered significantly enriched, and used in the follow-up computation of the hallmark cumulative heterogeneity score, if the SLAPenrichment false discovery rate (FDR) was less than 5% and its mutual exclusive coverage (EC) was greater than 50%.

### Down-sampling analyses

To investigate how differences in sample size might bias the SLAPenrichment results due to a potential tendency for larger datasets to produce larger number of SLAPenriched pathways, down-sampled SLAPenrich analyses were conducted for the 5 datasets with more than 350 samples (for BRCA, COREAD, GBM, HNSC, LUAD). Particularly, for $n \in \{800, 400, 250\}$

for BRCA and $n = 250$ for the other cancer types, 50 different SLAPenrich analyses were performed on $n$ samples randomly selected from the genomic dataset of the cancer type under consideration, with the parameter specifications described in the previous section. The average number of enriched pathways (FDR $< 5\%$ and EC $> 50\%$) across the 50 analysis was observed.

### Hallmark signature analysis: signature quantification

For a given cancer type $C$ and a given hallmark $H$ a cumulative heterogeneity score (CHS) was quantified as the ratio of the pathways associated to $H$ in the SLAPenrich analysis of the $C$ variants.

The CDS scores for all the 10 hallmark composed the hallmark signature of $C$.

**Author details**
[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Cambridge, UK. [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, CB10 1SD Cambridge, UK. [3]GlaxoSmithKline, Gunnels Wood Rd, SG1 2NY Stevenage Herts, UK. [4]Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Faculty of Medicine, MTI2 Wendlingweg 2, 52074 Aachen, Germany. [5]OpenTargets, Wellcome Genome Campus, CB10 1SD Cambridge, UK.

## References

1. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)

2. Consortium, T.I.C.G., committee, E., committee, E., policy, group, T., clinical annotation working, group, T.w., group, B.a.w., group, D.c., management working, Data release, d.t., publications working group, centre, D.c., committee, I.d.a., Australia, C.g.p.P.c.d.a., ovarian cancer serous adenocarcinoma, Canada, P.c.d.a., China, G.c.i.-, diffuse-type, European Union/France, R.c.r.c.c.f.o.b.n.l.t.c.c.s., European Union/United Kingdom, B.c.s.d.b.a.a.o.E.H.d.-t., France, B.c.s.d.b.a.a.o.t.H.g., France, L.c.h.c.s.t.a., adiposity, Paediatric brain tumours medulloblastoma, p.p.a.G., India, O.c.g., pancreatic tumours enteropancreatic endocrine tumours, R., rare pancreatic exocrine tumours; intraductal papillary mucinous neoplasms, s.p.t.m.c.n., other rarer tumours Italy, Japan, L.c.h.c.v.a., Spain, C.l.l.w.m., unmutated IgVH, United Kingdom, B.c.t.n., United States, T.C.G.A., committee, I.s.p.: PERSPECTIVES. Nature **464**(7291), 993–998 (2010)

3. Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R.J., Bignell, G.R., Tam, A.T., Davies, H., Stevenson, J.A., Barthorpe, S., Lutz, S.R., Kogera, F., Lawrence, K., McLaren-Douglas, A., Mitropoulos, X., Mironenko, T., Thi, H., Richardson, L., Zhou, W., Jewitt, F., Zhang, T., O'Brien, P., Boisvert, J.L., Price, S., Hur, W., Yang, W., Deng, X., Butler, A., Choi, H.G., Chang, J.W., Baselga, J., Stamenkovic, I., Engelman, J.A., Sharma, S.V., Delattre, O., Saez-Rodriguez, J., Gray, N.S., Settleman, J., Futreal, P.A., Haber, D.A., Stratton, M.R., Ramaswamy, S., McDermott, U., Benes, C.H.: Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature **483**(7391), 570–575 (2012)

4. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M.F., Monahan, J.E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F.A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I.H., Cheng, J., Yu, G.K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M.D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R.C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J.P., Gabriel, S.B., Getz, G., Ardlie, K., Chan, V., Myer, V.E., Weber, B.L., Porter, J., Warmuth, M., Finan, P., Harris, J.L., Meyerson, M., Golub, T.R., Morrissey, M.P., Sellers, W.R., Schlegel, R., Garraway, L.A.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature **483**(7391), 603–607 (2012)

5. Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., Getz, G.: Discovery and saturation analysis of cancer genes across 21 tumour types. Nature **505**(7484), 495–501 (2014)

6. Garraway, L.A., Lander, E.S.: Lessons from the Cancer Genome. Cell **153**(1), 17–37 (2013)

7. Stratton, M.R., Campbell, P.J., Futreal, P.A.: The cancer genome. Nature **458**(7239), 719–724 (2009)

8. Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., Raphael, B.J., Marks, D.S., Ouellette, B.F.F., Valencia, A., Bader, G.D., Boutros, P.C., Stuart, J.M., Linding, R., Lopez-Bigas, N., Stein, L.D.: Pathway and network analysis of cancer genomes. Nature Methods **12**(7), 615–621 (2015)

9. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell **144**(5), 646–674 (2011)

10. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Kinzler, K.W.: Cancer Genome Landscapes. Science (New York, NY) **339**(6127), 1546–1558 (2013)

11. Pe'er, D., Hacohen, N.: Principles and strategies for developing network models in cancer. Cell **144**(6), 864–873 (2011)

12. Reimand, J., Arak, T., Vilo, J.: g:Profiler–a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Research **39**(suppl), 307–315 (2011)

13. Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z.: GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics **10**(1), 48 (2009)

14. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols **4**(1), 44–57 (2008)

15. Wu, D., Smyth, G.K.: Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research **40**(17), 133–133 (2012)

16. Boca, S.M., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., Parmigiani, G.: Patient-oriented gene set analysis for cancer mutation data. Genome Biology **11**(11), 112 (2010)

17. Wendl, M.C., Wallis, J.W., Lin, L., Kandoth, C., Mardis, E.R., Wilson, R.K., Ding, L.: PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics **27**(12), 1595–1602 (2011)

18. Ciriello, G., Cerami, E., Sander, C., Schultz, N.: Mutual exclusivity analysis identifies oncogenic network modules. Genome Research **22**(2), 398–406 (2012)

19. Vandin, F., Upfal, E., Raphael, B.J.: De novo discovery of mutated driver pathways in cancer. Genome Research **22**(2), 375–385 (2012)

20. Schubert, M., Iorio, F.: Exploiting combinatorial patterns in cancer genomic data for personalized therapy and new target discovery. Pharmacogenomics **15**(16), 1943–1946 (2014)

21. Li, H.T., Zhang, J., Xia, J., Zheng, C.H.: Identification of driver pathways in cancer based on combinatorial patterns of somatic gene mutations. Neoplasma **63**(1), 57–63 (2016)

22. Lu, S., Lu, K.N., Cheng, S.-Y., Hu, B., Ma, X., Nystrom, N., Lu, X.: Identifying Driver Genomic Alterations in Cancers by Searching Minimum-Weight, Mutually Exclusive Sets. PLoS computational biology **11**(8), 1004257 (2015)

23. Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., Beerenwinkel, N.: TiMEx: a waiting time model for mutually exclusive cancer alterations. Bioinformatics (2015)

24. Yeang, C.H., McCormick, F., Levine, A.: Combinatorial patterns of somatic gene mutations in cancer. The FASEB Journal **22**(8), 2605–2622 (2008)

25. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., Fulton, L., Fulton, R.S., Zhang, Q., Wendl, M.C., Lawrence, M.S., Larson, D.E., Chen, K., Dooling, D.J., Sabo, A., Hawes, A.C., Shen, H., Jhangiani, S.N., Lewis, L.R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S.E., Clerc, K., Metcalf, G.A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M.L., Osborne, J.R., Meyer, R., Shi, X., Tang, Y., Koboldt, D.C., Lin, L., Abbott, R., Miner, T.L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B.H., Kraja, A., Crosby, S.D., Sawyer, C.S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L.R., Dutt, A., Fennell, T., Hanna, M., Johnson, B.E., Onofrio, R.C., Thomas, R.K., Tonon, G., Weir, B.A., Zhao, X., Ziaugra, L., Zody, M.C., Giordano, T., Orringer, M.B., Roth, J.A., Spitz, M.R., Wistuba, I.I., Ozenberger, B., Good, P.J., Chang, A.C., Beer, D.G., Watson, M.A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W.D., Pao, W., Province, M.A., Weinstock, G.M., Varmus, H.E., Gabriel, S.B., Lander, E.S., Gibbs, R.A., Meyerson, M., Wilson, R.K.: Somatic mutations affect key pathways in lung adenocarcinoma. Nature **455**(7216), 1069–1075 (2008)

26. Youn, A., Simon, R.: Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics **27**(2), 175–181 (2011)

27. Thomas, R.K., Baker, A.C., DeBiasi, R.M., Winckler, W., LaFramboise, T., Lin, W.M., Wang, M., Feng, W., Zander, T., MacConnaill, L.E., Lee, J.C., Nicoletti, R., Hatton, C., Goyette, M., Girard, L., Majmudar, K., Ziaugra, L., Wong, K.-K., Gabriel, S., Beroukhim, R., Peyton, M., Barretina, J., Dutt, A., Emery, C., Greulich, H., Shah, K., Sasaki, H., Gazdar, A., Minna, J., Armstrong, S.A., Mellinghoff, I.K., Hodi, F.S., Dranoff, G., Mischel, P.S., Cloughesy, T.F., Nelson, S.F., Liau, L.M., Mertz, K., Rubin, M.A., Moch, H., Loda, M., Catalona, W., Fletcher, J., Signoretti, S., Kaye, F., Anderson, K.C., Demetri, G.D., Dummer, R., Wagner, S., Herlyn, M., Sellers, W.R., Meyerson, M., Garraway, L.A.: High-throughput oncogene mutation profiling in human cancer. Nature genetics **39**(3), 347–351 (2007)

28. Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research **39**(Database issue), 685–90 (2011)

29. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M.: KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research **44**(D1), 457–62 (2016)

30. Soini, Y.: Tight junctions in lung cancer and lung metastasis: a review. International journal of clinical and experimental pathology **5**(2), 126–136 (2012)

31. Guy, S., Geletu, M., Arulanandam, R., Raptis, L.: Stat3 and gap junctions in normal and lung cancer cells. Cancers **6**(2), 646–662 (2014)

32. Baudot, A.i.s., de la Torre, V., Valencia, A.: Mutated genes, pathways and processes in tumours. EMBO reports **11**(10), 805–810 (2010)

33. Yang, J., Chen, J., He, J., Li, J., Shi, J., Cho, W.C., Liu, X.: Wnt signaling as potential therapeutic target in lung cancer. Expert opinion on therapeutic targets, 1–17 (2016)

34. Aita, M., Fasola, G., Defferrari, C., Brianti, A., Bello, M.G.D., Follador, A., Sinaccio, G., Pronzato, P., Grossi, F.: Targeting the VEGF pathway: antiangiogenic strategies in the treatment of non-small cell lung cancer. Critical reviews in oncology/hematology **68**(3), 183–196 (2008)

35. Larsen, J.E., Minna, J.D.: Molecular biology of lung cancer: clinical implications. Clinics in chest medicine **32**(4), 703–740 (2011)

36. Yim, H.W., Slebos, R.J.C., Randell, S.H., Umbach, D.M., Parsons, A.M., Rivera, M.P., Detterbeck, F.C., Taylor, J.A.: Smoking is associated with increased telomerase activity in short-term cultures of human bronchial epithelial cells. Cancer letters **246**(1-2), 24–33 (2007)

37. Sakakibara-Konishi, J., Oizumi, S., Kikuchi, J., Kikuchi, E., Mizugaki, H., Kinoshita, I., Dosaka-Akita, H., Nishimura, M.: Expression of Bim, Noxa, and Puma in non-small cell lung cancer. BMC Cancer **12**, 286 (2012)

38. Pirie, K., Peto, R., Green, J., Reeves, G.K., Beral, V., Million Women Study Collaborators: Lung cancer in never-smokers. International Journal of Cancer (2016)

39. Finberg, K.E., Sequist, L.V., Joshi, V.A., Muzikansky, A., Miller, J.M., Han, M., Beheshti, J., Chirieac, L.R., Mark, E.J., Iafrate, A.J.: Mucinous differentiation correlates with absence of EGFR mutation and presence of KRAS mutation in lung adenocarcinomas with bronchioloalveolar features. The Journal of molecular diagnostics : JMD **9**(3), 320–326 (2007)

40. Woo, H.-J., Yoo, W.J., Bae, C.H., Song, S.-Y., Kim, Y.-W., Park, S.-Y., Kim, Y.-D.: Leptin up-regulates MUC5B expression in human airway epithelial cells via mitogen-activated protein kinase pathway. Experimental lung research **36**(5), 262–269 (2010)

41. Raina, D., Kosugi, M., Ahmad, R., Panchamoorthy, G., Rajabi, H., Alam, M., Shimamura, T., Shapiro, G.I., Supko, J., Kharbanda, S., Kufe, D.: Dependence on the MUC1-C oncoprotein in non-small cell lung cancer cells. Molecular Cancer Therapeutics **10**(5), 806–816 (2011)

42. Sakuma, Y., Matsukuma, S., Yoshihara, M., Nakamura, Y., Noda, K., Nakayama, H., Kameda, Y., Tsuchiya, E., Miyagi, Y.: 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. American journal of clinical pathology **128**(1), 100–108 (2007)

43. Iorio, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X., Egan, R.K., Liu, Q., Mironenko, T., Mitropoulos, X., Richardson, L., Wang, J., Zhang, T., Moran, S., Sayols, S., Soleimani, M., Tamborero, D., Lopez-Bigas, N., Ross-Macdonald, P., Esteller, M., Gray, N.S., Haber, D.A., Stratton, M.R., Benes, C.H., Wessels, L.F.A., Saez-Rodriguez, J., McDermott, U., Garnett, M.J.: A Landscape of Pharmacogenomic Interactions in Cancer. Cell (2016)

44. Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., Lopez-Bigas, N.: In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell **27**(3), 382–396 (2015)

45. Knijnenburg, T.A., Bismeijer, T., Wessels, L.F.A., Shmulevich, I.: A multilevel pan-cancer map links gene mutations to cancer hallmarks. Chinese journal of cancer **34**(10), 439–449 (2015)

46. Manié, E., Popova, T., Battistella, A., Tarabeux, J., Caux-Moncoutier, V., Golmard, L., Smith, N.K., Mueller, C.R., Mariani, O., Sigal-Zafrani, B., Dubois, T., Vincent-Salomon, A., Houdayer, C., Stoppa-Lyonnet, D., Stern, M.-H.: Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas. International Journal of Cancer **138**(4), 891–900 (2016)

47. Walsh, C.S.: Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. Gynecologic oncology **137**(2), 343–350 (2015)

48. Yu, X., Jiang, Y., Wei, W., Cong, P., Ding, Y., Xiang, L., Wu, K.: Androgen receptor signaling regulates growth of glioblastoma multiforme in men. Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine **36**(2), 967–972 (2015)

49. Boland, C.R., Goel, A.: Microsatellite instability in colorectal cancer. (2010)

50. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Imielinsk, M., Jäger, N., Jones, D.T.W., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., López-Otín, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N.J., Valdés-Mas, R., van Buuren, M.M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P.A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R.: Signatures of mutational processes in human cancer. Nature **500**(7463), 415–421 (2013)

51. Syed, D.N., Khan, M.I., Shabbir, M., Mukhtar, H.: MicroRNAs in skin response to UV radiation. Current drug targets **14**(10), 1128–1134 (2013)

52. Zhang, X., Wan, G., Berger, F.G., He, X., Lu, X.: The ATM kinase induces microRNA biogenesis in the DNA damage response. Molecular cell **41**(4), 371–383 (2011)

53. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., Eder, J.P., Patnaik, A., Gubens, M., Horn, L., Carcereny, E., Ahn, M.-J., Felip, E., Lee, J.-S., Hellmann, M.D., Hamid, O., Goldman, J.W., Soria, J.-C., Dolled-Filhart, M., Rutledge, R.Z., Zhang, J., Lunceford, J.K., Rangwala, R., Lubiniecki, G.M., Roach, C., Emancipator, K., Gandhi, L., KEYNOTE-001 Investigators: Pembrolizumab for the treatment of non-small-cell lung cancer. The New England journal of medicine **372**(21), 2018–2028 (2015)

54. Robert, C., Schachter, J., Long, G.V., Arance, A., Grob, J.J., Mortier, L., Daud, A., Carlino, M.S., McNeil, C., Lotem, M., Larkin, J., Lorigan, P., Neyns, B., Blank, C.U., Hamid, O., Mateus, C., Shapira-Frommer, R., Kosh, M., Zhou, H., Ibrahim, N., Ebbinghaus, S., Ribas, A., KEYNOTE-006 investigators: Pembrolizumab versus Ipilimumab in Advanced Melanoma. The New England journal of medicine **372**(26), 2521–2532 (2015)

55. Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack, E.R., Castellano, D., Choueiri, T.K., Gurney, H., Donskov, F., Bono, P., Wagstaff, J., Gauler, T.C., Ueda, T., Tomita, Y., Schutz, F.A., Kollmannsberger, C., Larkin, J., Ravaud, A., Simon, J.S., Xu, L.-A., Waxman, I.M., Sharma, P., CheckMate 025 Investigators: Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. The New England journal of medicine **373**(19), 1803–1813 (2015)

56. Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., Biedrzycki, B., Donehower, R.C., Zaheer, A., Fisher, G.A., Crocenzi,

T.S., Lee, J.J., Duffy, S.M., Goldberg, R.M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R.H., Wood, L.D., Cuka, N., Pardoll, D.M., Papadopoulos, N., Kinzler, K.W., Zhou, S., Cornish, T.C., Taube, J.M., Anders, R.A., Eshleman, J.R., Vogelstein, B., Diaz, L.A.: PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. The New England journal of medicine **372**(26), 2509–2520 (2015)

57. Jess, T., Rungoe, C., Peyrin-Biroulet, L.: Risk of colorectal cancer in patients with ulcerative colitis: a meta-analysis of population-based cohort studies. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association **10**(6), 639–645 (2012)

58. West, N.R., McCuaig, S., Franchini, F., Powrie, F.: Emerging cytokine networks in colorectal cancer. Nature reviews. Immunology **15**(10), 615–629 (2015)

59. Lasry, A., Zinger, A., Ben-Neriah, Y.: Inflammatory networks underlying colorectal cancer. Nature immunology **17**(3), 230–240 (2016)

60. Poteet, E., Choudhury, G.R., Winters, A., Li, W., Ryou, M.-G., Liu, R., Tang, L., Ghorpade, A., Wen, Y., Yuan, F., Keir, S.T., Yan, H., Bigner, D.D., Simpkins, J.W., Yang, S.-H.: Reversing the Warburg effect as a treatment for glioblastoma. Journal of Biological Chemistry **288**(13), 9153–9164 (2013)

61. Pikor, L., Thu, K., Vucic, E., Lam, W.: The detection and implication of genome instability in cancer. Cancer metastasis reviews **32**(3-4), 341–352 (2013)

62. Grivennikov, S.I., Greten, F.R., Karin, M.: Immunity, inflammation, and cancer. Cell **140**(6), 883–899 (2010)

63. Deryugina, E.I., Quigley, J.P.: Matrix metalloproteinases and tumor metastasis. Cancer metastasis reviews **25**(1), 9–34 (2006)

64. Rabbani, S.A., Mazar, A.P.: The role of the plasminogen activation system in angiogenesis and metastasis. Surgical oncology clinics of North America **10**(2), 393–415 (2001)

65. Kumari, S., Malla, R.: New Insight on the Role of Plasminogen Receptor in Cancer Progression. Cancer growth and metastasis **8**, 35–42 (2015)

66. Zarling, J.M., Shoyab, M., Marquardt, H., Hanson, M.B., Lioubin, M.N., Todaro, G.J.: Oncostatin M: a growth regulator produced by differentiated histiocytic lymphoma cells. Proceedings of the National Academy of Sciences of the United States of America **83**(24), 9739–9743 (1986)

67. Tartour, E., Dorval, T., Mosseri, V., Deneux, L., Mathiot, C., Brailly, H., Montero, F., Joyeux, I., Pouillart, P., Fridman, W.H.: Serum interleukin 6 and C-reactive protein levels correlate with resistance to IL-2 therapy and poor survival in melanoma patients. British journal of cancer **69**(5), 911–913 (1994)

68. Lacreusette, A., Nguyen, J.-M., Pandolfino, M.-C., Khammari, A., Dreno, B., Jacques, Y., Godard, A., Blanchard, F.: Loss of oncostatin M receptor beta in metastatic melanoma cells. Oncogene **26**(6), 881–892 (2007)

69. Caffarel, M.M., Coleman, N.: Oncostatin M receptor is a novel therapeutic target in cervical squamous cell carcinoma. The Journal of pathology **232**(4), 386–390 (2014)

70. Faraone, D., Aguzzi, M.S., Toietta, G., Facchiano, A.M., Facchiano, F., Magenta, A., Martelli, F., Truffa, S., Cesareo, E., Ribatti, D., Capogrossi, M.C., Facchiano, A.: Platelet-derived growth factor-receptor alpha strongly inhibits melanoma growth in vitro and in vivo. Neoplasia (New York, NY) **11**(8), 732–742 (2009)

71. Yamazaki, D., Kurisu, S., Takenawa, T.: Regulation of cancer cell motility through actin reorganization. Cancer science **96**(7), 379–386 (2005)

72. Bid, H.K., Roberts, R.D., Manchanda, P.K., Houghton, P.J.: RAC1: an emerging therapeutic option for targeting cancer angiogenesis and metastasis. Molecular Cancer Therapeutics **12**(10), 1925–1934 (2013)

73. Bailey, C.L., Kelly, P., Casey, P.J.: Activation of Rap1 promotes prostate cancer metastasis. Cancer research **69**(12), 4962–4968 (2009)

74. Lee, J.-W., Ryu, Y.-K., Ji, Y.-H., Kang, J.H., Moon, E.-Y.: Hypoxia/reoxygenation-experienced cancer cell migration and metastasis are regulated by Rap1- and Rac1-GTPase activation via the expression of thymosin beta-4. Oncotarget **6**(12), 9820–9833 (2015)

75. Matin, R.N., Chikh, A., Chong, S.L.P., Mesher, D., Graf, M., Sanza, P., Senatore, V., Scatolini, M., Moretti, F., Leigh, I.M., Proby, C.M., Costanzo, A., Chiorino, G., Cerio, R., Harwood, C.A., Bergamaschi, D.: p63 is an alternative p53 repressor in melanoma that confers chemoresistance and a poor prognosis. The Journal of experimental medicine **210**(3), 581–603 (2013)

76. Costanzo, A., Pediconi, N., Narcisi, A., Guerrieri, F., Belloni, L., Fausti, F., Botti, E., Levrero, M.: TP63 and TP73 in cancer, an unresolved "family" puzzle of complexity, redundancy and hierarchy. FEBS letters **588**(16), 2590–2599 (2014)

77. Perez-Llamas, C., Lopez-Bigas, N.: Gitools: analysis and visualisation of genomic data using interactive heat-maps. PLoS ONE **6**(5), 19541 (2011)

78. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N.: Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling **6**(269), 1 (2013)

79. Ciriello, G., Cerami, E., Aksoy, B.A., Sander, C., Schultz, N.: Using MEMo to discover mutual exclusivity modules in cancer. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 8**, 8–17 (2013)

80. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America **100**(16), 9440–9445 (2003)

81. Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B., Gillette, M., Paulovich, A., Pomeroy, S., Golub, T., Lander, E.: 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 **102**(43), 15545 (2005)

82. Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L., D'Eustachio, P.: The Reactome pathway knowledgebase. Nucleic Acids Research **42**(Database issue), 472–7 (2014)

83. Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H., Thomas, P.D.: The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Research **33**(Database issue), 284–8 (2005)

84. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C.A., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D.S., Karp, P.D.: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research **44**(D1), 471–80 (2016)

85. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the Pathway Interaction Database. Nucleic Acids Research **37**(Database issue), 674–9 (2009)

86. Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C., Xia, J., Liang, Y., Shrivastava, S., Wishart, D.S.: SMPDB: The Small Molecule Pathway Database. Nucleic Acids Research **38**(Database issue), 480–7 (2010)

87. Grondin, C.J., Davis, A.P., Wiegers, T.C., King, B.L., Wiegers, J.A., Reif, D.M., Hoppin, J.A., Mattingly, C.J.: Advancing Exposure Science through Chemical Data Curation and Integration in the Comparative Toxicogenomics Database. Environmental health perspectives (2016)

88. Yamamoto, S., Sakai, N., Nakamura, H., Fukagawa, H., Fukuda, K., Takagi, T.: INOH: ontology-based highly structured database of signal transduction pathways. Database : the journal of biological databases and curation **2011**, 052 (2011)

89. Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E.L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, S.R., Miller, R., Coort, S.L., Cirillo, E., Smeets, B., Evelo, C.T., Pico, A.R.: WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Research **44**(D1), 488–94 (2016)

90. Kandasamy, K., Mohan, S., Raju, R., Keerthikumar, S., Kumar, G.S.S., Venugopal, A.K., Telikicherla, D., Navarro, D.J., Mathivanan, S., Pecquet, C., Gollapudi, S.K., Tattikota, S.G., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H.K.C., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R.,

Ramachandra, Y.I., Rahiman, A., Keshava Prasad, T.s., Lin, J.-X., Houtman, J.C.D., Desiderio, S., Renauld, J.-C., Constantinescu, S., Ohara, O., Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G.D., Sander, C., Leonard, W.J., Pandey, A.: NetPath: a public resource of curated signal transduction pathways. Genome Biology **11**(1), 3 (2010)

91. Chou, C.-H., Chang, N.-W., Shrestha, S., Hsu, S.-D., Lin, Y.-L., Lee, W.-H., Yang, C.-D., Hong, H.-C., Wei, T.-Y., Tu, S.-J., Tsai, T.-R., Ho, S.-Y., Jian, T.-Y., Wu, H.-Y., Chen, P.-R., Lin, N.-C., Huang, H.-T., Yang, T.-L., Pai, C.-Y., Tai, C.-S., Chen, W.-L., Huang, C.-Y., Liu, C.-C., Weng, S.-L., Liao, K.-W., Hsu, W.-L., Huang, H.-D.: miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Research **44**(D1), 239–47 (2016)

92. Wain, H.M., Bruford, E.A., Lovering, R.C., Lush, M.J., Wright, M.W., Povey, S.: Guidelines for human gene nomenclature. Genomics **79**(4), 464–470 (2002)

93. Johnson, D., Krishnan, S., Chhugani, J., Kumar, S., Venkatasubramanian, S.: Compressing large boolean matrices using reordering techniques. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30. VLDB '04, pp. 13–23. VLDB Endowment, ??? (2004). http://dl.acm.org/citation.cfm?id=1316689.1316693

94. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical review. E, Statistical, nonlinear, and soft matter physics **69**(6 Pt 2), 066133 (2004)

95. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C.Y., Jia, M., De, T., Teague, J.W., Stratton, M.R., McDermott, U., Campbell, P.J.: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research **43**(Database issue), 805–11 (2015)

96. Gonzalez-Perez, A., Lopez-Bigas, N.: Functional impact bias reveals cancer drivers. Nucleic Acids Research **40**(21), 169 (2012)

97. Tamborero, D., Gonzalez-Perez, A., Lopez-Bigas, N.: OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics **29**(18), 2238–2244 (2013)

**Additional Files**

**Additional figure 1 — Differential pathway enrichment analysis results: non-mucinous vs mucinous bronchioloalveolar LUAD patients**
Differential SLAPenrich analysis obtained contrasting two sub-pulations of LUAD patients based on their bronchioloalveolar type (non-mucinous vs. mucinous).

**Additional figure 2 — Manually curated mapping between genes, pathways and hallmarks (properties):**
(A) Heatmap with cancer hallmarks on the rows, pathways gene sets on the columns. A coloured bar in position $(i, j)$ indicates that the $j$-th pathway is associated with the $i$-th hallmark; bar diagram on the right shows the number of pathways associated with each hallmark. (B) Heatmap with cancer hallmarks on the rows and genes on the columns. A coloured bar in position $(i, j)$ indicates that the $j$-th gene is contained in at least one pathway associated with the $i$-th hallmark (thus associated with the $i$-th hallmark); bar diagram on the right shows the number of genes associated with each hallmark. (C) Number of associated hallmarks per pathways: the majority of the pathways is associated with 1 hallmark. (D) Number of associated hallmarks per gene: the majority of the genes is associated with less than 3 hallmarks. (E) Distribution of Jaccard similarity scores (quantifying the extent of pair-wise overlaps) computed between pairs of pathway gene sets.

**Additional figure 3 — Enriched pathways versus sample size, downsampled analyses, and covered known cancer genes:**
(A) Number of significantly enriched pathway at the population versus the number of samples available in the analysed cohorts, across cancer type. (B) Number of significantly enriched pathway at the population level across 5 different cancer types (with more than 350 samples), indicated by different colors, and down-sampled trials. In each of this trials, for each cancer type and 50 different iterations, a set of $n$ samples is randomly selected and a SLAPenrich analysis is performed on this sub-set of data. Average number of SLAPenriched pathway (and standard deviations) are reported. n = 800, 400, and 250 for BRCA and n = 250 for the other four cancer type. For four of the tested tissues there is no tendency for increased number of samples to produce more SLAPenriched pathways. A mild dependency trend is observable for BRCA only, with a continuously increasing average number of enriched pathways as a function of sample size up to 800 samples, that plateaus above this size, with a very similar number of enriched pathways when analysing 1,132 samples or across 50 analysis on 800 pathways. (C) Each bar quantifies the ratio of high-confidence cancer genes (as predicted in [44]) contained in at least one pathway enriched at the population level (covered pathways), across cancer types. Different contained colored bars indicate the ratio of the genes included in covered pathways associated to different hallmarks, one colored bar per hallmark. The white bar at the top indicate the ratio of genes included in covered pathways associated to multiple hallmarks.

**Additional figure 4 — Hallmark heterogeneity across cancer types:**
Heatmaps showing pathways enrichments at the population level across cancer types for individual hallmarks. Color intensities correspond to the enrichment significance. Cancer types and pathways are clustered using a correlation metric. See also figure 4.

**Additional figure 5 — Impact of known cancer genes' mutations on the results:**
Heatmaps showing, for each enriched-pathway/cancer-type, the ratio between the number samples harbouring mutations in known cancer genes belonging to the pathway under consideration and the total number of samples harbouring mutations in any gene belonging to the pathway under consideration. See also figure 4.

**Additional figure 6 — Impact of known cancer genes' mutations on the results:**
Heatmaps showing, for each enriched-pathway/cancer-type, the ratio between the number samples harbouring mutations in known cancer genes belonging to the pathway under consideration and the total number of samples harbouring mutations in any gene belonging to the pathway under consideration. See also figure 4.

**Additional figure 7 — Hallmark signature analysis to discover new cancer driver networks:**
In each row, first circle plots shows pathway enrichments at the population levels when considering all the somatic variants (bars on the external circle) and when considering only variants not involving known high-confidence cancer driver genes; second circle plot shows similarly a comparison between the hallmark signatures resulting from SLAPenrich analysis including (bars on the external circle) or excluding (bars on the internal circle) the variants involving known high-confidence cancer genes. The bar plot shows a comparison, in terms of true-positive-rate (TPR) and positive-predicted-value (PPV), of the SLAPenriched pathways across the two analysis and, finally, the scatter plots on the right shows a comparison between the resulting hallmark signatures.

**Additional file 1 — KEGG pathways enriched in the LUAD case study dataset**
**Additional file 2 — Differential enrichment analysis comparing Smokers vs. Non-Smokers: Results**
**Additional file 3 — Differential enrichment analysis comparing Mucinus vs. Non-Mucinus BAC types: Results**
**Additional file 4 — SLAPenrich analysis results across 10 different cancer types**
**Additional file 5 — List of keywords used to curate the mapping between genes, pathways and hallmarks**
**Additional file 6 — Manually curated mapping between genes, pathways and hallmarks**
**Additional file 7 — Overview of the exposed function of the SLAPenrich R package**