bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHOD

Dissecting the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich

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 a computational method, implemented in an open source R package, to explore how cancers from different tissue types might have acquired the same cancer hallmark (evolutionary successful trait) via preferential genomically altering different biological pathways. To this aim, we have curated a collection of 374 orthogonal pathway gene-sets encompassing thousands of genes (from public available resources) mapped to 10 canonical cancer hallmarks.

Using this curated data resource, we have characterised the landscape of pathway alterations putatively contributing to the acquisition of different cancer hallmarks via systematic analysis of somatic mutations in large cohorts of patients across 10 cancer types, from the Cancer Genome Atlas.

We assume that the heterogeneity of each hallmark in terms of number of corresponding enriched pathway alterations is reflective of its evolutionary fitness to the cancer type under consideration. A systematic evaluation of this heterogeneity across hallmarks and cancer types has resulted into a set of cancer hallmark heterogeneity signatures. These signatures quantitatively confirm the established predominance of certain hallmarks in determined cancer types and their clinical relevance, and they allow an easy data-driven comparison of cancer hallmark heterogeneity across different lineages.

We have found, as expected, that most of the pathway alteration enrichments and large hallmark heterogeneities are guided by somatic mutations in established, and highly frequently mutated, high-confidence cancer driver genes. However and most importantly, when excluding these variants from the analyses, we observe that the hallmark heterogeneity signatures, thus the level of predominance of the considered hallmarks, are strikingly preserved across cancer types. Therefore we propose to use the hallmark heterogeneity signatures as a ground truth to characterise long tails of infrequent genomic alterations, across cancer types, and we highlight a number of 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 a prior knowledge of the cellular mechanisms and biological pathways where the products of the mutated genes operate. Multiple methods exist based on this principle, reviewed in [8]. This reduces the dimensionality of large genomic datasets involving thousands of altered genes into a sensibly smaller set of altered mechanisms that are more interpretable, possibly actionable in a

*Correspondence: iorio@ebi.ac.uk; saezrodriguez@gmail.com
[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Cambridge, UK
Full list of author information is available at the end of the article
†Co-corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pharmacological or experimental way [9], and can be also used as therapeutic markers whose predictive ability is significantly improved when compared to that of genomic lesions in individual genes [10]. Additionally, this facilitates the stratification of cancer patients into informative subtypes [11], the characterisation of rare somatic mutations [12], and the identification of the spectrum of possible alterations underpinning a common evolutionary successful trait acquired by a normal cell as it transforms itself in a precancerous one and ultimately into cancer. In two landmark papers [13, 14] these traits have been summarised into a set of 11 principles, collectively referred as the *hallmarks of cancer*.

Here we propose a computational strategy to characterise the set of genomically altered pathways contributing to the acquisition of the canonical cancer hallmarks across 10 different cancer types, via a systematic analysis of 4,415 public available cancer patients' genomes (from the Cancer Genome Atlas). Then we devise a metric to assess the predominance of each hallmark in each cancer type quantitatively and in a data driven way. Finally, after verifying that the majority of these predominances are led by somatic mutations in established high-confidence cancer genes, we show that they are maintained when excluding these variants from the analysis. Thus we propose to use the obtained *heterogeneity signatures of cancer hallmarks* as the ground truth for functionally characterising long tails of infrequent genomic alterations, across cancer types. Finally we highlight a number of potential novel cancer driver genes and networks, identified with the proposed approach.

Results

Sample Level Analysis of Pathway Alterations Enrichments (SLAPenrich)

In the first step of our computational pipeline we make use of SLAPenrich (Sample Level Analysis of Pathway alteration Enrichments): a computational method implemented into an R package to perform pathway analyses of genomic datasets at the sample-population level. We have designed this tool on purpose 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.

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. Many tools exist and are routinely used to perform this analysis [15, 16, 17], sometimes incorporating additional features, such as inter-gene dependencies and signal correlations [18], and also estimating single sample pathway deregulations based on transcriptional data [19].

To identify pathways, gene sets and gene-ontology categories that are over-represented in a selected set of genes satisfying a certain property (for example, being differentially expressed when contrasting two biological states of interests), the likelihood of their recurrence in the gene set of interests is usually estimated. This is usually quantified through a *p-value* assignment computed through a hypergeometric (or Fisher's exact) test, against the null hypothesis that there is no association between the pathway under consideration and the biological state yielding the selected set of genes. The test fails (producing a non significant *p-value*) when the size of the overlap between the considered pathway and the set of genes of interests is close to that expected by random chance.

The problem we tackle with SLAPenrich is rather different: we want to test the hypothesis that, in a given cohort of cancer patients (or any population under evolutionary pressure), the number of samples harbouring a mutation in at least one gene belonging to a given pathway is significantly larger and divergent from its expectation (when considering the size of the measured cohort, the background mutation rate and the non-overlapping total exonic block lengths of all the genes). If this is the case, then the pathway under consideration is deemed as enriched at the sample population level (SLAPenriched) in relation to the whole cohort of patients.

The first step of our method consists in modeling the probability of observing at least a mutation in a single gene belonging to the pathway under consideration across the individual samples (one probability per sample, quantifying the likelihood of observing at least a mutation in that pathway by random chance, given the background mutation rate and the total exonic block length of the pathway). These individual probabilities are then aggregated in a collective test against the null hypothesis that there is no association between the pathway under consideration and

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the genomic alterations observed in the analysed cohort.To our knowledge there are only two other tools to conduct this type of analyses: PathScan [20] and PathScore [21]. SLAPenrich performs comparably to them, showing a slightly improved ability to rank as highly enriched pathways containing established cancer driver genes. Additionally, several aspects make SLAPenrich more suitable for the analyses described in this manuscript. These are discussed in the following section, together with a systematic comparison of SLAPenrich, PathScan and PathScore.

The second class of approaches aims at identifying novel pathways by mapping genomic alteration patterns on large protein interaction 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 [22, 23, 24]. 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 [22, 23, 25, 26, 27].

SLAPenrich does not require somatic mutations in a pathway to be statistically enriched among those detected in an individual sample nor the merged (or aggregated) set of mutations in the population. It assumes that a single mutation of a gene of a pathway in an individual sample can be sufficient to deregulate the pathway activity, potentially providing selective advantages. This allows pathways containing groups of genes with a tendency to be mutated in mutually exclusive fashion (therefore different individually mutated genes in different samples) to be still detected as enriched at the sample population level.
Subsequently, using a filter that accounts for this tendency, SLAPenrich can prioritize the pathways that are statistically enriched and also mutated in a mutually exclusive manner, as a further evidence of positive selection [28]. Hence, SLAPenrich belongs roughly to the first category described above, although it shares the mutual exclusivity consideration with the methods in the second.

More precisely, after modeling the probability of observing a genomic alteration in at least one member of a given pathway across the individual samples, SLAPenrich performs a collective statistical test 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, for example, the probability of the actual number of mutated genes) is that this prevent signal saturations due to hypermutated 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 harbours somatic mutations in at least one of the genes belonging to the pathway under consideration (Figure 1B).

The probability of success of each of these trials can be computed by either (i) a general hypergeometric model 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 [29]. 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 integrated in SLAPenrich.

Once these probabilities have been computed, the expected number of samples in the population harbouring 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 harbouring somatic mutations in P from its expectation, and its statistical significance quantified analytically (Figure 1C). Finally, the resulting statistically enriched pathways are 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 [30, 22, 23].
SLAPenrich includes a visualization/report framework allowing an easy exploration of the outputted enriched

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

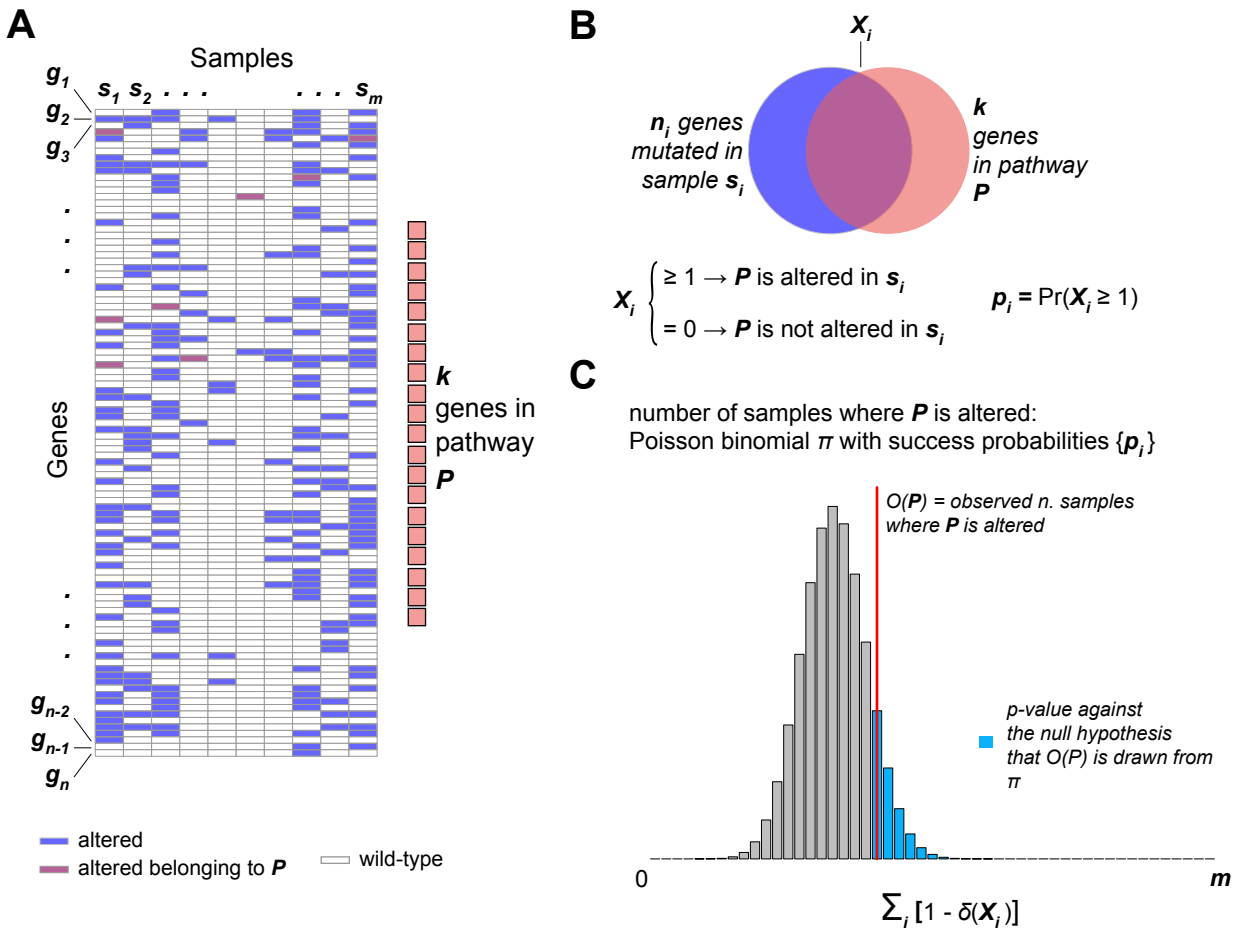

Figure 1 Schematic of the statistical framework underlying SLAPenrich: (A) The probability p_i of a pathway P being genomically altered in the individual sample s_i of the analyzed dataset is computed. This accounts for the somatic mutation rate of the sample and the sum of the total exonic length blocks 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, 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 π. Here the success probabilities are the likelihoods computed in A. δ 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 π evaluated at $O(P)$, which is the observed number of samples where P is genomically altered.

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.

A formal description of the statistical framework underlying SLAPenrich is provided in the Methods; further details are provided in the Supplementary File 1.

SLAPenrich: case study and comparisons with existing tools

To test the ability of SLAPenrich to recover pathways that are known to be associated to a given disease state and different clinico-pathological features, we have re-analysed (as detailed in the Supplementary File 1), a published dataset encompassing somatic mutations found in 188 lung adenocarcinoma (LUAD) patients, studied in [20].

This analysis yielded 48 significantly enriched pathways, at a FDR $< 5\%$ and a mutual exclusive coverage (EC) $> 50\%$ (Supplementary Table S1). Among these,

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we found pathways whose deregulation is known to be involved in lung cancer, such as *Tight Junction* (alteration score (AS) = 0.37, EC = 89%) [31] (Supplementary Figure S1A), *Gap Junction* (AS = 0.45, EC = 75%) [32], and several pathways previously found with other computational methods in LUAD (such as PathScan [20], among others [33]), and recently proposed as potential targets for lung cancer therapy (detailed results are included in Additional File 1). We found a significant agreement between our results and those obtained with PathScan on the same cohort of cancer patients (and reported in the Supplementary Table 1 of [20]) (Fisher's exact test (FET) p-value = 2.10×10^{-14}, Supplementary Tables S1 and S2 and Additional File 1). Additionally, we observed a significant correlation (R = 0.66, p = 0.0002) between the significance levels of the 26 commonly enriched pathways across the two methods (Figure 3A).

Similarly, we performed a comparison between the output obtained with SLAPenrich and PathScore [21] when analysing the LUAD dataset described above. To obtain comparable results we used the whole collection of 1,392 canonical pathway signatures from the Molecular Signature Database (MsigDB) [34], as this is the reference collection used by PathScore. We observed a significant overlap (181 pathways, FET p-value = 2.76×10^{-70}) between the enriched pathways outputted by SLAPenrich (at an FDR $< 5\%$) and those outputted by PathScore (adjusted *p-value* < 0.05) (Supplementary Table S3).

We also performed, with the same common collection of pathways as above, a systematic comparison between SLAPenrich and PathScore [21] on genomic datasets encompassing 4,415 patients across 10 different cancer types from The Cancer Genome Atlas (TCGA). Results confirmed that SLAPenrich and PathScore detect very similar sets of enriched pathways across all the different analysed cancer types (median $-\log_{10}$(FET p)= 119.2 ranging from 29 to 202, Supplementary Figure S3A). We observed a slightly better ability of SLAPenrich in ranking highly pathways that include at least one tissue-specific high-confidence cancer gene (HCG) [35]: median HCGs covered by the top 10 enriched pathway for SLAPenrich = 18% against 8% for PathScore; 21% and 14% for the top 20; 33% and 25% for the top 50; 45% and 34% for the 100% (Figure 3B, and Supplementary Figure S3BC). The median difference of HCGs covered by pathways enriched according to the two methods at the same significance level (5% FDR for SLAPenrich and adjusted $p < 0.05$ for PathScore) favoured PathScore for a 1%.

Even if performing similarly to SLAPenrich a number of features of PathScan and PathScore make them unsuitable for the hallmark analyses presented here. PathScan does not take into account of possible mutual exclusivity trends between patterns of mutations of genes in the same pathway and, in more practical terms, it requires raw sequencing data (BAM files) as input: this is quite uncomfortable for large scale analyses where (as in our case) it is far more convenient to use available processed datasets represented through binary presence/absence matrices. PathScore uses the same mathematical framework of SLAPenrich, but the models for computing the individual pathway mutation probabilities are not fully customisable. More importantly, it is implemented as a web-application that restricts the number of individual analyses to a maximum of 10 per week.
Furthermore, both PathScan and PathScore make use of fixed pathway collections from public repositories (KEGG [36] for PathScan, and MsigDB [34] for PathScore). In contrast, the SLAPenrich R package allows users to define and use any collection of gene sets and by default it employs a large pathway collection from Pathway Commons [37] (including 2,794 pathways, covering 15,281 genes, 15 times the pathways and 3 times the genes considered by Pathscan, and twice the pathways and 1.72 times the genes of PathScore). Additionally, the SLAPenrich R package includes routines to update on the fly gene attributes and exonic lengths, to check and update gene nomenclatures in datasets and reference pathway gene sets, to perform mutual exclusivity sorting of binary matrices, and to identify core-components (i.e. sub-sets of genes leading the enrichment of different pathways). Furthermore, a unique feature of SLAPenrich with respect to PathScan and PathScore is that it allows to perform pathway differential enrichment analyses between sub-populations of the analysed cohort. Thus, it is able to associate pathway enrichments to different clinicopathological features. As an example, in the LUAD case study we were able to correctly identify pathways that are preferentially recurrently altered in never-smokers/current-smokers patients as well as mucinous/non-mucionous bronchioalveolar carcinomas (detailed results are reported in the Additional File 1, Supplementary Table S4 and S5, and Supplemtary Figures S1 and S2). Finally, SLAPenrich post-processes pathway collections for redundancy reduction: in this way pathways with large overlaps are merged together instead of being tested individually (Figure 2).

In summary, while other tools, specially PathScore, are based on similar assumptions and perform comparably, SLAPenrich provides a more flexible environ-

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ment enabling a wide range of possible large-scale analyses.

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

Leveraging the unique capacity of SLAPEnrich, we set out to perform a systematic large-scale analysis of pathway deregulation in cancer. We used a collection of pathways from the Pathway Commons data portal (v8, 2016/04) [37] (post-processed as detailed in the Methods), and we performed individual SLAPenrich analyses of 10 different genomic datasets containing somatic point mutations, preprocessed as described in [38], 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). In these analyses we used a Bernoulli model to define individual pathway alteration probabilities across the single samples (equation 5). With respect to the hypergeometric models (equations 3 and 4), this formulation upon full expansion sums the individual gene mutation probabilities, each accounting for the individual gene lengths.
The analysed 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).

Results from all these individual SLAPenrich analyses are contained in the Additional File 5.
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 (Supplementary Figure S4A), but a down-sampled analysis showed that our results are not broadly confounded by the sample sizes (see Methods and Supplementary Figure S4B).
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 [38, 35] (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) (Supplementary Figure S4C).
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 3C, as well as from the number of high confidence cancer driver genes (as predicted in [35], median = 100, range from 33 for THCA to 251 for SKCM, Figure 3D). 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 3C). 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 3C). 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 subtypes 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' [13, 14], that can be linked to gene mutations [39], 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 Supplementary Table S5) we were able to map 374 gene-sets (from the most recent release of pathway commons [37]) to 10 cancer hallmarks [13, 14] (Figure 2AB), for a total number of 3,915 genes (included in at least one gene set associated to at least one hallmark; Supplementary Table S5). The vast majority (99%, 369 sets) of the considered pathway gene-sets were mapped on two hallmarks

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

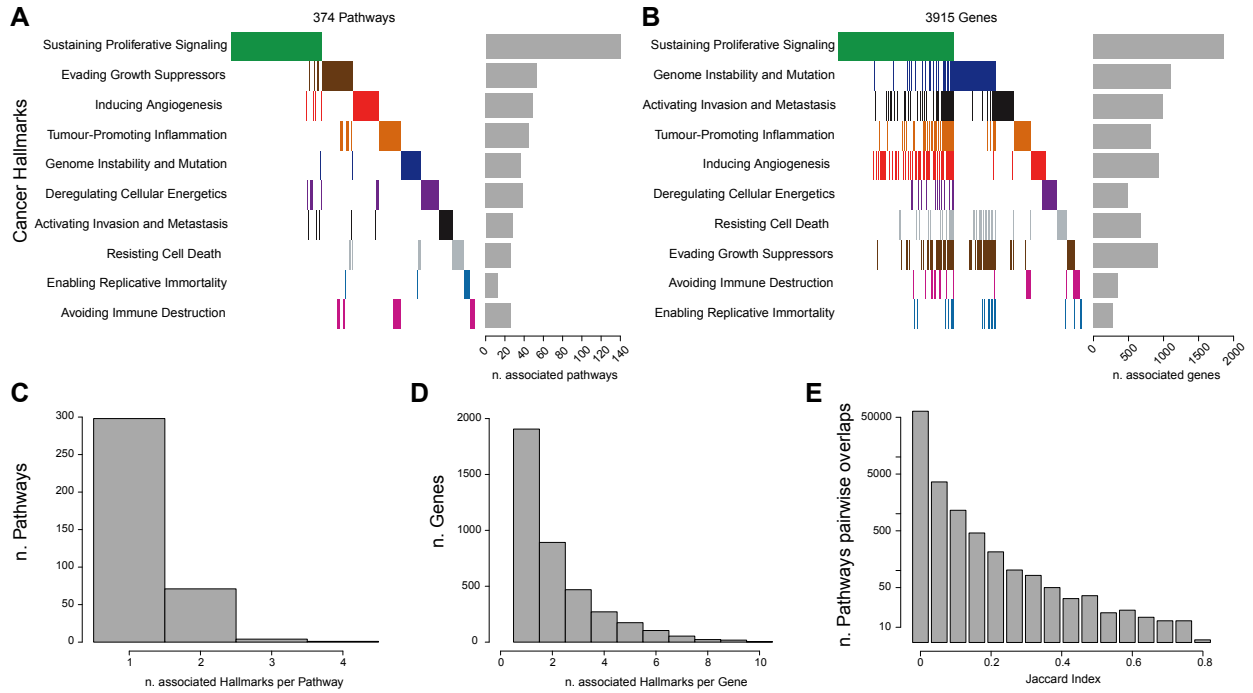

Figure 2 Manually curated mapping between genes, pathways and hallmarks (properties): (A) Heatmap with cancer hallmarks on the rows, pathways gene sets on the columns. A coloured bar in position (i, j) indicates that the j-th pathway is associated with the i-th hallmark; bar diagram on the right shows 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.

at most, and 298 of them (80%) was mapped onto one single hallmark (Figure 2C). Regarding the individual genes contained in at least one pathway gene-set, about half (49%) were associated with a single hallmark, 22% with two, 12% with three, and 7% with four (Figure 2D). Finally, as shown in 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 (Figure 4, and Supplementary Figure S5). 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 established tissue specific and high-confidence cancer genes (HCGs) [35], 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 Supplementary Figures S6 and S7. 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

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

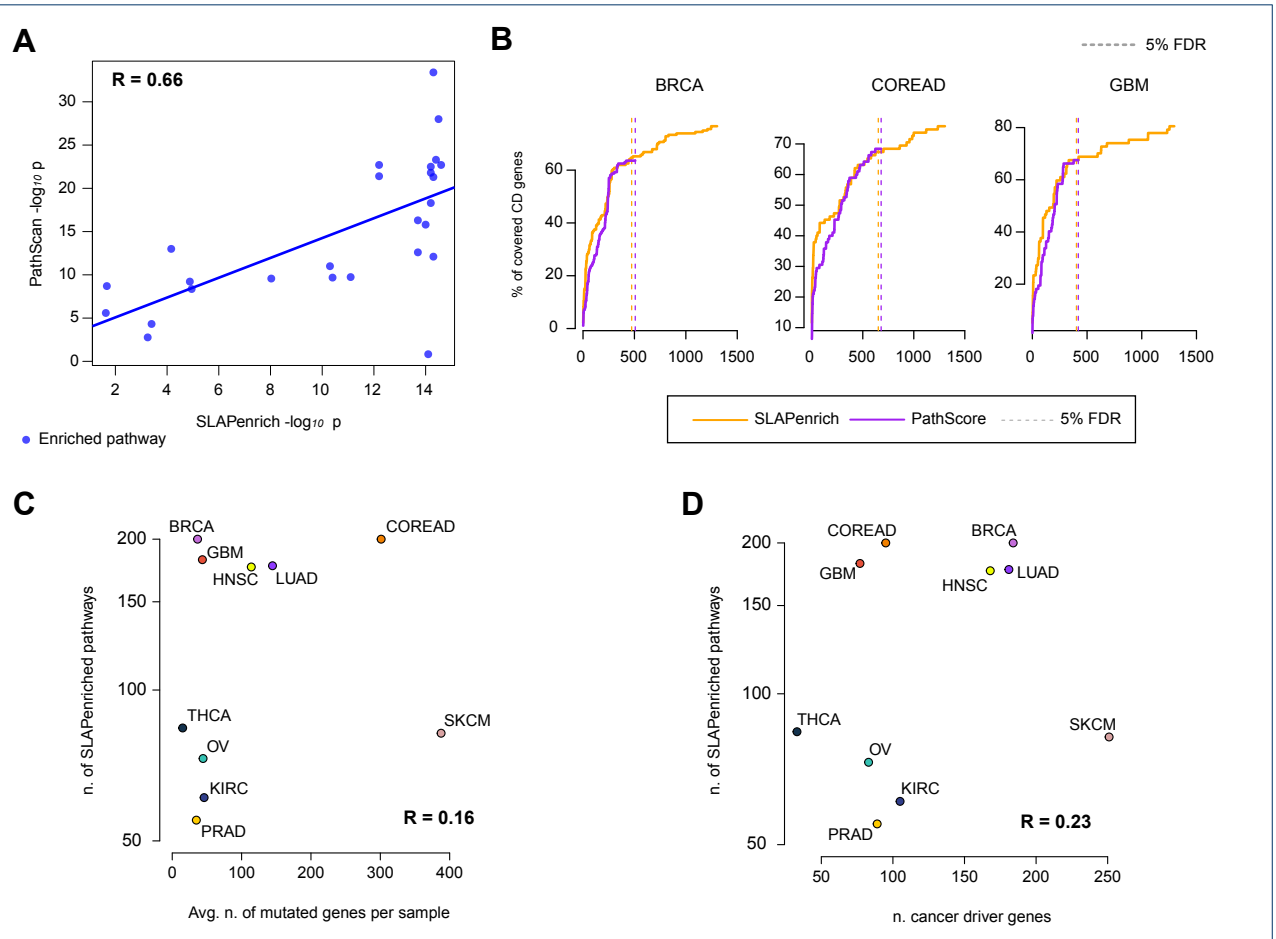

Figure 3 SLAPenrich comparison with existing methods and comparison between number of enrichments versus mutation burdens and number of established cancer genes: (A) Comparison between the significance levels of the enriched pathways (blue dots) identified with both SLAPenrich (x-axis) and PathScan (y-axis) on the LUAD dataset; (B) Percentages of tissue specific high-confidence cancer driver genes included in the top k enriched pathways according to SLAPenrich and PathScore. For SLAPenrich, all the possible k values are considered; PathScore does not output results for all the tested pathways but only for the significantly enriched one, therefore in this case k ranges from 1 to the least significantly enriched pathway. Results for three representative cancer types are reported. All the other results are included in Supplementary Figure S3C; (C) Number of pathway enriched at the population level across cancer types compared with the average number of mutated genes and (D) the average number of high confidence cancer driver genes.

THCA (91%).

Patterns and well defined clusters can be clearly distinguished in the heatmaps of Figure 4 and Supplementary Figure S5. As an example, 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 [40]. 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 [41].

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 [42] 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 Anneal-*

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

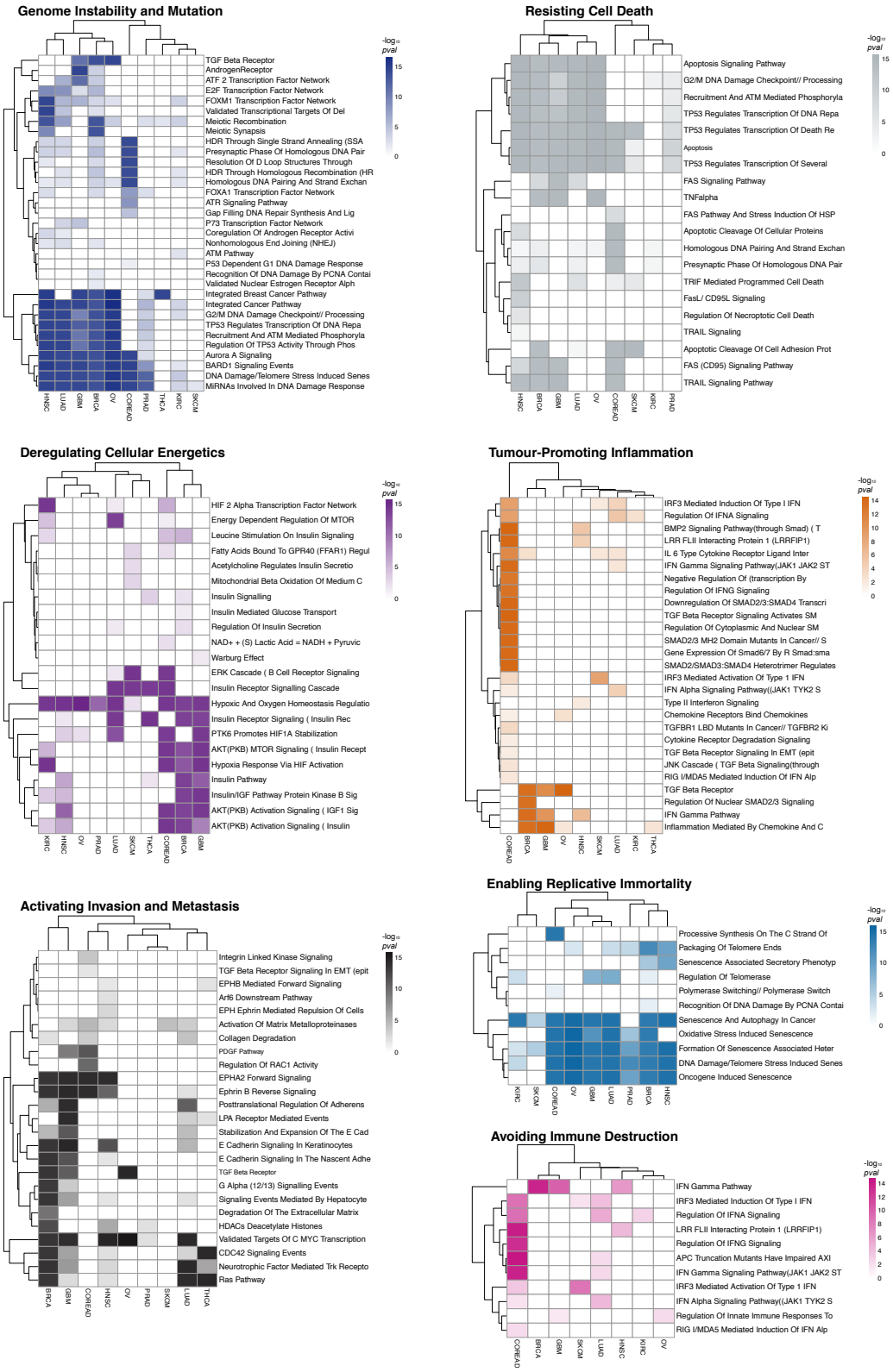

Figure 4 Heterogeneity of hallmark acquisition across cancer types: Heatmaps showing pathways enrichments at the population level across cancer types for individual hallmarks (representative cases). Color intensities correspond to the enrichment significance. Cancer types and pathways are clustered using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

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ing (SSA), *Resolution Of D Loop Structures Through Synthesis Dependent Strand Annealing (SDSA)*, *Homologous 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 [43].
Finally, the heatmap for *Genome Instability and Mutation* shows nearly no enriched pathways associated to the acquisition of this hallmark for SKCM. This is consistent with the high burden of mutations observed in melanoma being the effect of this hallmark rather than leading its acquisition. In fact, the genomic instability of SKCM originates from cell extrinsic processes such as UV light exposure [44].
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 [45]. 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 [46]), 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 [47], melanoma [48] or renal-cell carcinoma [49]. 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 [50].
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 [51]. Indeed, a number of clinical trials evaluating the utility of inflammatory and cytokine-modulatory therapies are currently underway in colorectal cancer [52, 53]. 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 [54]. 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 significantly 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* (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 [55], for *Tumour-promoting inflammation* and *Avoiding immune-destruction* in COREAD [56]. Lastly, and as expected for *Sustaining proliferative-signaling* and *Enabling replicative immortality*, the key hallmarks in cancer initiation [13], 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.

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

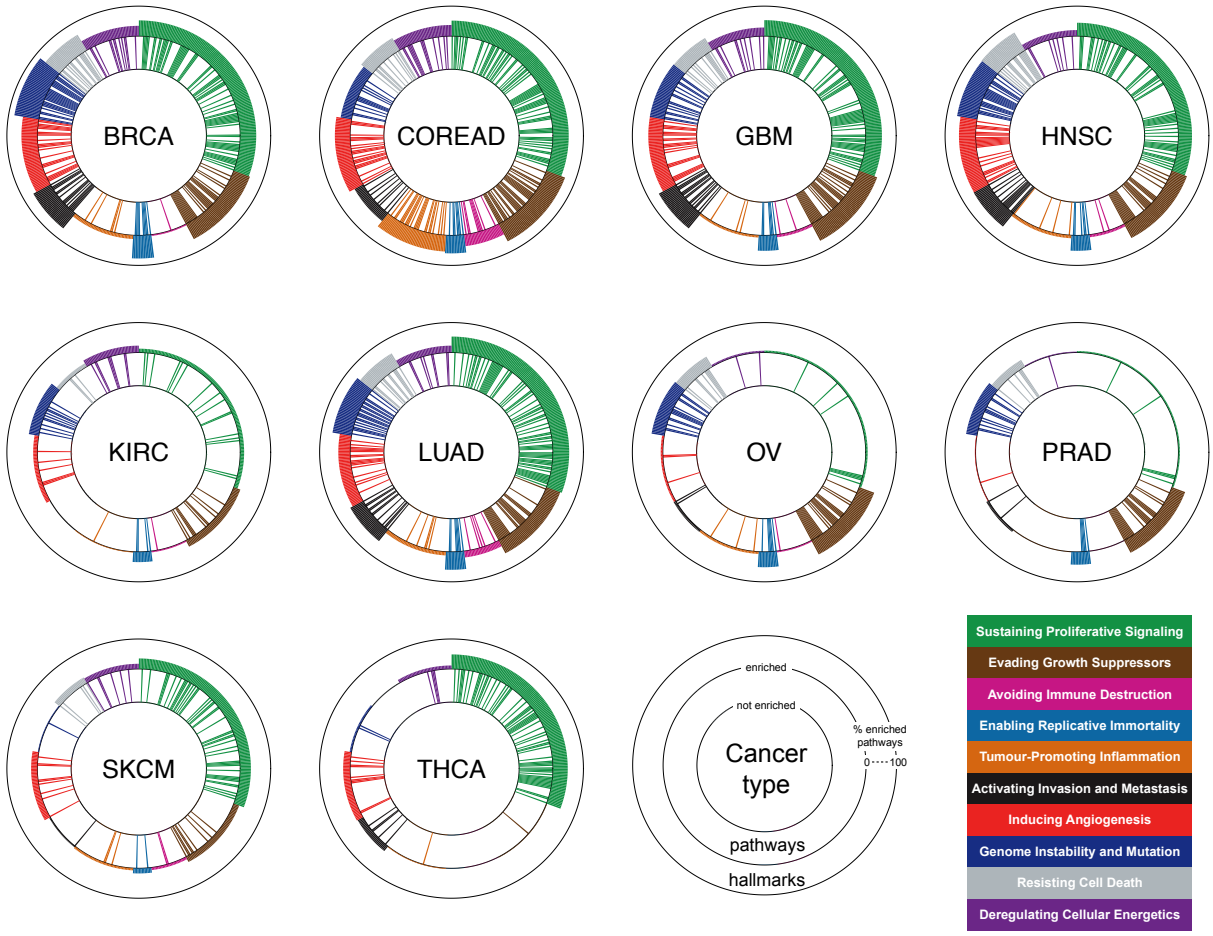

Figure 5 Cancer hallmark heterogeneity signatures: Each cancer hallmark signature plot is composed of three concentric circles. Bars between the inner and middle circles indicate pathways, bars between the middle and external circle indicate cancer hallmarks. Different colors indicate different cancer 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.

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 considered cancer type, the corresponding HCGs. Results from this exercise (Figure 6, Additional File 6 and Supplementary Figure S8), 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 (maintained enrichments (MA) in Figure 6 and Supplementary Figure S8) 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 across the filtered and unfiltered analyses for most of the cancer types, with coincident top fitting hallmarks and significantly high over-all correlations (Figure 6, Supplementary Figure S8). 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 uncov-

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ered 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 (examples are shown in Figure 7 and the whole bulk of results is included in the Additional File 7).

An example is represented by the pathway *Activation Of Matrix Metalloproteinases* associated with the *Invasion and metastasis* hallmark and highly enriched 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 [57]. 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 Metallopeptidase 2 - MMP2 - has been reported as harbouring cancer driving alterations in LUAD [35]) 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 [58]. 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 [59]. 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 [60]. Increased IL-6 blood levels in melanoma patients correlate with disease progression and lower response to chemotherapy [61]. Importantly, studies proposed OSMR, a IL-6-type of cytokine receptor, to play a role in the prevention of melanoma progression [62], and as a novel potential target in other cancer types [63]. 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 [64].

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 pivotal for cell motility [65]. 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 [66, 67]. 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 [68].

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 [69], dissecting the functional relations between them is still widely considered a formidable challenge [70]. 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 show the effectiveness of SLAPenrich in identifying potential novel cancer driver genes and cancer driver networks composed by lowly frequently mutated genes.

Discussion

We have presented a computational pipeline, with a paired statistical framework implemented in an open

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

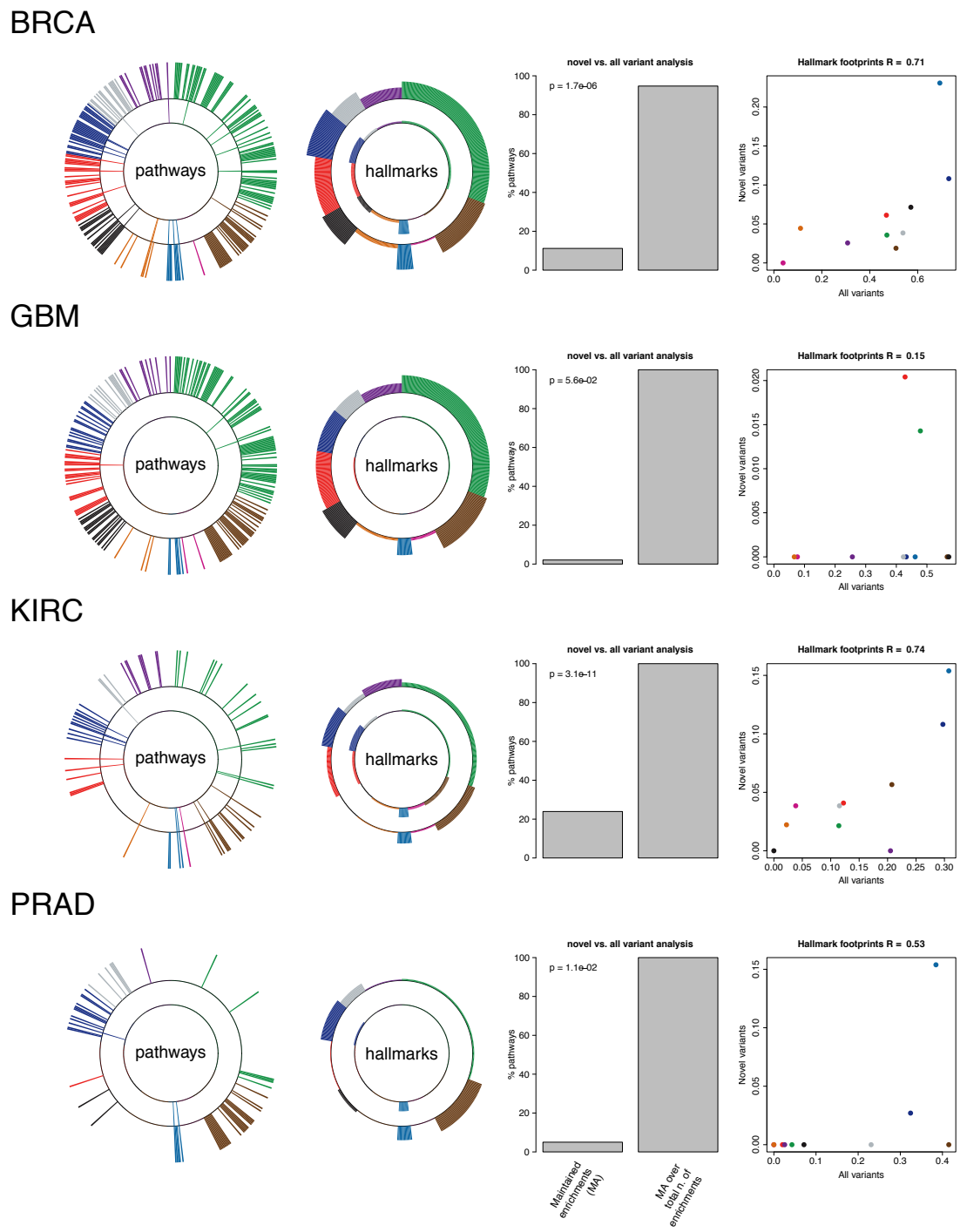

Figure 6 Hallmark heterogeneity signature analysis with and without known drivers: In each row, the first circle plot show pathway enrichments at the population level when considering all the somatic variants (bars on the external circle) and when considering only variants not involving known high-confidence 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.

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

source R package (SLAPenrich) to identify genomic alterations in biological pathways contributing in the acquisition of the canonical cancer hallmarks. Our statistical framework 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 mutations and other genomic alterations from publicly available or user defined datasets via a browser accessible software suite (e.g. GiTools [71] and cBioPortal [72]) or as a result of combinatorial pattern analysis (such as MEMo [73] and Dendrix [23]). 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.
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 high level of abstraction. As a consequence, our computational pipeline 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 large cohorts of publicly available cancer genomes patient data that is publicly available in the TCGA. However, SLAPenrich is of great utility in other scenarios such as for characterizing genomic data generated upon chemical mutagenesis to identify somatic mutations involved in acquired drug resistance (an application has been recently published in [74]). 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.

Conclusions

We have performed a large-scale comparative analysis of the mutational landscape of different cancer types at the level of Cancer Hallmarks. Our results represent a first data-driven landmark exploration of the hallmarks of cancer showing that they might be acquired through preferential genomic alterations of heterogenous sets of pathways across cancer types. This has confirmed the established predominance of certain hallmark in defined cancer types, and has highlighted peculiar patterns of altered pathways for several cancer lineages. Finally, by using the identified hallmark signatures as a ground truth signal, we have devised an approach to detect novel cancer driver genes and networks.

A number of possible limitations could hamper deriving definitive conclusions from our study, 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. Possible future developments of our method could integrate different *omics*, such as transcriptional data, to better refine the set of functionally impacting variants considered in the analysis. Additionally further refinements will account for structural variants such as small indels and copy number alterations, known to play an important role in cancer. Nevertheless, we have provided the community with a useful tool for the analysis of large genomic datasets, whose results could open a wide range of novel *in silico* investigations.

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

$$f(g_i, s_j) = \{1 \text{ if } g_i \text{ is mutated in } s_j \text{ and } 0 \text{ otherwise}\}. \quad (1)$$

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

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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_j if there is a gene g_i belonging to G such that g_i is a member of P and $f(g_i, s_j) = 1$, i.e. at least one gene in the pathway P is altered in the j-th sample (Figure 1B). To quantify how likely it is to observe at least one gene belonging to P altered in sample s_j, we introduce the variable $X_j = |\{g_i \in G : g_i \in P \text{ and } f(g_i, s_j) = 1\}|$, accounting for the number of genes in P altered in sample s_j. Under the assumption of both a gene-wise and sample-wise statistical independence, the probability of X_j assuming a value greater or equal than 1 is given by:

$$p_j = \Pr(X_j \geq 1) = \sum_{x=1}^{k} H(x, N, k, n_j), \tag{2}$$

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

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

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_j probabilities (of observing at least one genes in the pathway P altered in sample s_j) as follows:

$$p_j = \Pr(X_j \geq 1) = \sum_{x=1}^{k} H(x, N', k', n'_j), \tag{4}$$

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'_j is the total number of individual point mutations involving genes belonging to P in sample s_j, and H is defined as in equation 3, but with parameters x, N', k', and n'_j. Similarly, the p_j 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 [29], as follows:

$$p_j = \Pr(X_j \geq 1) = 1 - \exp(-\rho k'), \tag{5}$$

where k' is defined as for equation 4 and ρ 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)[29].
If considering the event "the pathway P is altered in sample s_j" as the outcome of a single test in a set of Bernoulli trials $\{j\}$ (with $j = 1, \ldots, M$) (one for each sample in S), then each p_j can be interpreted as the success probability of the $j - th$ 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:

$$E(P) = \sum_{j=1}^{M} p_j. \tag{6}$$

On the other hand, if we consider a function ϕ 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. $\phi(X) = \{1 \text{ if } X > 0, \text{ and } 0 \text{ otherwise}\}$, 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:

$$O(P) = \sum_{j=1}^{M} \phi(X_j). \tag{7}$$

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:

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

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 $\{j\}$ independent Bernoulli trials where the success probabilities p_j are not identical (with $j = 1, \ldots, M$). In our case, the j-th Bernoulli trial accounts for the event "the pathway P is altered in the sample s_j" and its success probability is given by the $\{p_j\}$ introduced above (and computed with one amongst 2, 4, or 5). The parameters of such B distribution are then the probabilities $\pi = \{p_j\}$, and its mean is given by Equation 6. The

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

probability of the event $O(P) = y$ can be then written as

$$\Pr(O(P) = y) = B(\pi, y) = \sum_{A \in F_y} \prod_{k \in A} p_k \prod_{h \in A^c} (1 - p_h), \tag{9}$$

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 π. Such *p-value* can be derived for an observation $O(P) = z$, with $z \leq M$, as (Figure 1C):

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

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 [75]. SLAPenrich is implemented as an R packag publicly available and fully documented at (https://github.com/saezlab/SLAPenrich/). An overview of the exposed function of this package is also provided in the Additional File 8.

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 [36] pathway gene sets from MsigDB [34], 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) [37] (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, 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 [76] 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:

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

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 Supplementary Table S6. Results of this data curation are reported in the Supplementary Table S7.

Mutual exclusivity coverage

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 selec-

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tion. To this aim, for a given enriched pathway P, an exclusive coverage score $C(P)$ is computed as

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

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 threshold value pass this final filter.

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 [38] (http://www.cancerrxgene.org/gdsc1000/). This resource (available at http://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Data/suppData/TableS2B.xlsx) encompasses variants from sequencing of 6,815 tumor normal sample pairs derived from 48 different sequencing studies [35] and reannotated using a pipeline consistent with the COSMIC database [77] (Vagrent: https://zenodo.org/record/16732#.VbeVY2RViko).
Lists of tissue specific high-confidence cancer genes [35] 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 [78, 79] and further filtered as described in [38] (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/saezlab/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[37], post-processed for redundancy reduction as explained in the previous sections, and embedded in the SLAPenrich package as R data object: `PATHCOM_HUMAN_nr_i_hu_2016.RData`.
A pathway in this collection was considered significantly enriched, and used in the following 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, i.e. BRCA, COREAD, GBM, HNSC, and 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.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and material
R code and data-objects are available at:
https://github.com/saezlab/SLAPenrich/.
Pre-processed data sources are specified in the Methods.

Funding
OpenTargets funds JSR (Projects OpenTargets15 and OpenTargets16).

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
FI designed the statistical framework underlying SLAPenrich, conceived the hallmark heterogeneity analysis, and designed the other heuristic algorithms, conceived the visualization framework, implemented the R package, and wrote the manuscript; LGA contributed to the implementation of the visualization functions, tested and contributed to implementing the R package, curated data, and contributed to manuscript writing and revising; JB contributed to testing the R package, interpreted results and findings, contributed to manuscript writing and revising; IM contributed to the design of the validation analyses, read and edited the manuscript; DRW contributed to the design of the statistical framework and supervised its mathematical formalization; UM contributed to the interpretation of results; JSR supervised the study and contributed to the manuscript writing and revising.

Acknowledgements
We would like to thank Jorge Buendia, Mathew Garnett, and Annalisa "Lilla" Mupo for a number of insightful discussions, David Tamborero and Nuria Lopez-Bigas for critical feedback on the manuscript.

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. Pe'er, D., Hacohen, N.: Principles and strategies for developing network models in cancer. Cell **144**(6), 864–873 (2011)
10. Shi, W., Jiang, T., Nuciforo, P., Hatzis, C., Holmes, E., Harbeck, N., Sotiriou, C., Peña, L., Loi, S., Rosa, D.D., Chia, S., Wardley, A., Ueno, T., Rossari, J., Eidtmann, H., Armour, A., Piccart-Gebhart, M., Rimm, D.L., Baselga, J., Pusztai, L.: Pathway level alterations rather than mutations in single genes predict response to HER2-targeted therapies in the neo-ALTTO trial. Annals of Oncology, 434 (2016)
11. Hofree, M., Shen, J.P., Carter, H., Gross, A., Ideker, T.: Network-based stratification of tumor mutations. Nature Methods **10**(11), 1108–1115 (2013)
12. Leiserson, M.D.M., Vandin, F., Wu, H.-T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M.S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G.A., Lopez-Bigas, N., Getz, G., Ding, L., Raphael, B.J.: Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics **47**(2), 106–114 (2015)
13. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell **144**(5), 646–674 (2011)
14. 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)
15. 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)
16. 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)
17. 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)
18. Wu, D., Smyth, G.K.: Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Research **40**(17), 133–133 (2012)
19. Drier, Y., Sheffer, M., Domany, E.: Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences of the United States of America **110**(16), 6388–6393 (2013)
20. 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)
21. Gaffney, S.G., Townsend, J.P.: PathScore: a web tool for identifying altered pathways in cancer data. Bioinformatics (2016)
22. Ciriello, G., Cerami, E., Sander, C., Schultz, N.: Mutual exclusivity analysis identifies oncogenic network modules. Genome Research **22**(2), 398–406 (2012)
23. Vandin, F., Upfal, E., Raphael, B.J.: De novo discovery of mutated driver pathways in cancer. Genome Research **22**(2), 375–385 (2012)
24. Schubert, M., Iorio, F.: Exploiting combinatorial patterns in cancer

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genomic data for personalized therapy and new target discovery. Pharmacogenomics **15**(16), 1943–1946 (2014)
25. 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)
26. 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)
27. Constantinescu, S., Szczurek, E., Mohammadi, P., Rahnenführer, J., Beerenwinkel, N.: TiMEx: a waiting time model for mutually exclusive cancer alterations. Bioinformatics (2015)
28. Yeang, C.H., McCormick, F., Levine, A.: Combinatorial patterns of somatic gene mutations in cancer. The FASEB Journal **22**(8), 2605–2622 (2008)
29. Youn, A., Simon, R.: Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics **27**(2), 175–181 (2011)
30. 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)
31. Soini, Y.: Tight junctions in lung cancer and lung metastasis: a review. International journal of clinical and experimental pathology **5**(2), 126–136 (2012)
32. Guy, S., Geletu, M., Arulanandam, R., Raptis, L.: Stat3 and gap junctions in normal and lung cancer cells. Cancers **6**(2), 646–662 (2014)
33. 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)
34. 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)
35. 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)
36. 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)
37. 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)
38. 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)
39. 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)
40. 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)
41. 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)
42. 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)
43. Boland, C.R., Goel, A.: Microsatellite instability in colorectal cancer. (2010)
44. 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)
45. 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)
46. 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)
47. Garon, E.B., Rizvi, N.A., Hui, R., Leighl, N., Balmanoukian, A.S., Eder, J.P., Patnaik, A., Aggarwal, C., 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)
48. 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)
49. Motzer, R.J., Escudier, B., McDermott, D.F., George, S., Hammers, H.J., Srinivas, S., Tykodi, S.S., Sosman, J.A., Procopio, G., Plimack,

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)

50. 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)
51. 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)
52. West, N.R., McCuaig, S., Franchini, F., Powrie, F.: Emerging cytokine networks in colorectal cancer. Nature reviews. Immunology **15**(10), 615–629 (2015)
53. Lasry, A., Zinger, A., Ben-Neriah, Y.: Inflammatory networks underlying colorectal cancer. Nature immunology **17**(3), 230–240 (2016)
54. 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)
55. 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)
56. Grivennikov, S.I., Greten, F.R., Karin, M.: Immunity, inflammation, and cancer. Cell **140**(6), 883–899 (2010)
57. Deryugina, E.I., Quigley, J.P.: Matrix metalloproteinases and tumor metastasis. Cancer metastasis reviews **25**(1), 9–34 (2006)
58. 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)
59. Kumari, S., Malla, R.: New Insight on the Role of Plasminogen Receptor in Cancer Progression. Cancer growth and metastasis **8**, 35–42 (2015)
60. 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)
61. 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)
62. 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)
63. 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)
64. 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)
65. Yamazaki, D., Kurisu, S., Takenawa, T.: Regulation of cancer cell motility through actin reorganization. Cancer science **96**(7), 379–386 (2005)
66. 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)
67. Bailey, C.L., Kelly, P., Casey, P.J.: Activation of Rap1 promotes prostate cancer metastasis. Cancer research **69**(12), 4962–4968 (2009)
68. 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)
69. 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)
70. 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)
71. Perez-Llamas, C., Lopez-Bigas, N.: Gitools: analysis and visualisation of genomic data using interactive heat-maps. PLoS ONE **6**(5), 19541 (2011)
72. 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)
73. 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)
74. Brammeld, J.S., Petljak, M., Martincorena, I., Williams, S.P., Alonso, L.G., Dalmases, A., Bellosillo, B., Robles-Espinoza, C.D., Price, S., Barthorpe, S., Tarpey, P., Alifrangis, C., Bignell, G., Vidal, J., Young, J., Stebbings, L., Beal, K., Stratton, M.R., Saez-Rodriguez, J., Garnett, M., Montagut, C., Iorio, F., McDermott, U.: Genome-wide chemical mutagenesis screens allow unbiased saturation of the cancer genome and identification of drug resistance mutations. Genome Research (2017)
75. 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)
76. 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)
77. 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)
78. Gonzalez-Perez, A., Lopez-Bigas, N.: Functional impact bias reveals cancer drivers. Nucleic Acids Research **40**(21), 169 (2012)
79. 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 file 1 — Supplementary Methods
Additional file 2 — Legends of Supplementary Figures and Tables
Additional file 3 — Supplementary Figures S1 to S8
Additional file 4 — Supplementary Tables S1 to S7
Additional file 5 — SLAPenrich results across 10 cancer types
Additional file 6 — SLAPenrich results across 10 cancer types excluding cancer driver genes
Additional file 7 — Novel cancer driver genes and networks

bioRxiv preprint doi: https://doi.org/10.1101/077701; this version posted March 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

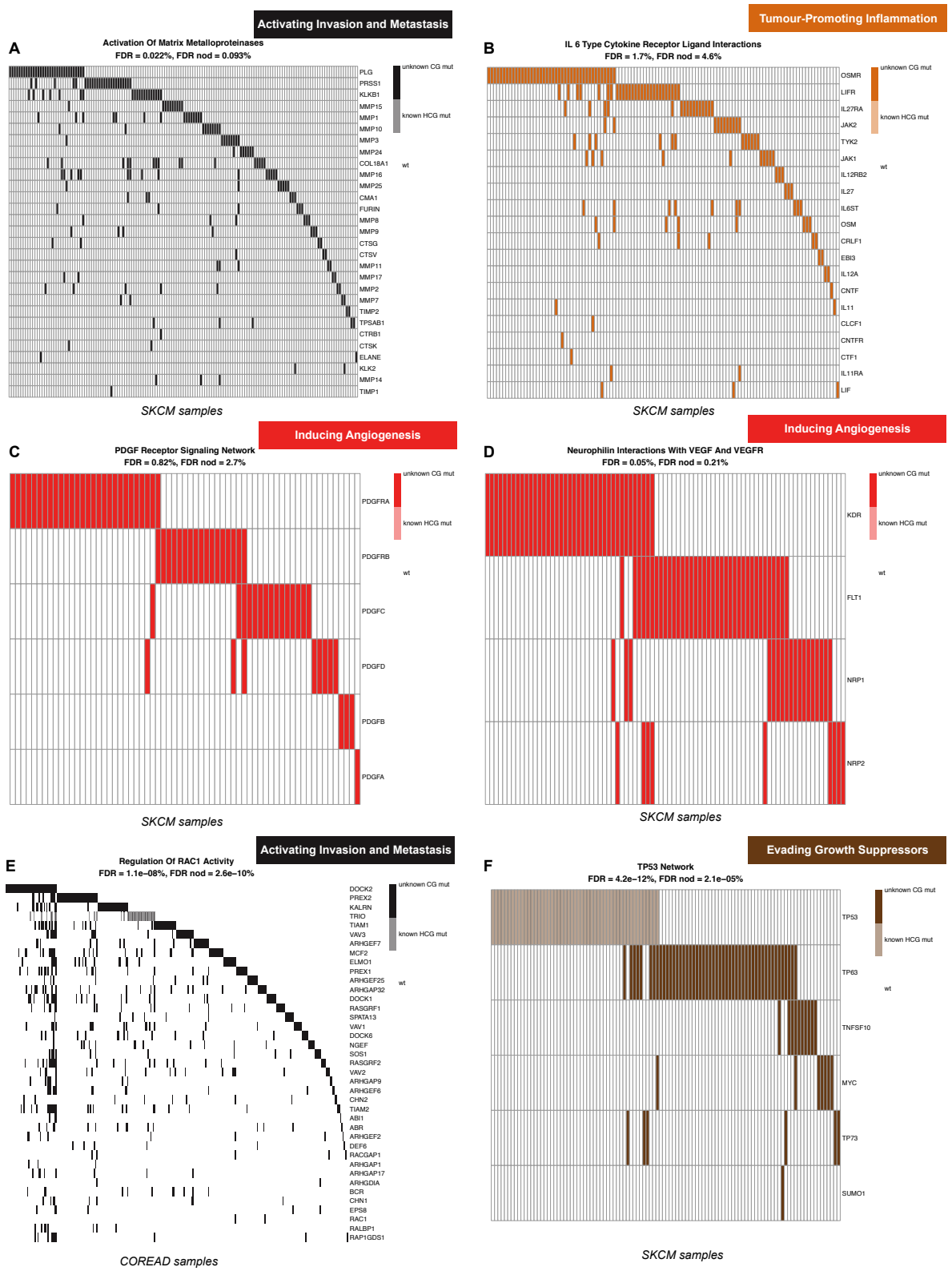

Figure 7 Example of putative novel cancer genes and networks identified by SLAPenrich Picked examples of novel putative cancer driver genes and networks. The first FDR value refers to the unfiltered analysis, whereas the second FDR refers to the filtered one (in which variants involving high confidence and highly frequently mutated cancer driver genes have been removed).