www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 5 June 2017
Accepted: 16 April 2018
Published online: 30 April 2018

Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich

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

Cancer hallmarks are evolutionary traits required by a tumour to develop. While extensively characterised, the way these traits are achieved through the accumulation of somatic mutations in key biological pathways is not fully understood. To shed light on this subject, we characterised the landscape of pathway alterations associated with somatic mutations observed in 4,415 patients across ten cancer types, using 374 orthogonal pathway gene-sets mapped onto canonical cancer hallmarks. Towards this end, we developed SLAPenrich: a computational method based on population-level statistics, freely available as an open source R package. Assembling the identified pathway alterations into sets of hallmark signatures allowed us to connect somatic mutations to clinically interpretable cancer mechanisms. Further, we explored the heterogeneity of these signatures, in terms of ratio of altered pathways associated with each individual hallmark, assuming that this is reflective of the extent of selective advantage provided to the cancer type under consideration. Our analysis revealed the predominance of certain hallmarks in specific cancer types, thus suggesting different evolutionary trajectories across cancer lineages. Finally, although many pathway alteration enrichments are guided by somatic mutations in frequently altered high-confidence cancer genes, excluding these driver mutations preserves the hallmark heterogeneity signatures, thus the detected hallmarks' predominance across cancer types. As a consequence, we propose the hallmark signatures as a ground truth to characterise tails of infrequent genomic alterations and identify potential novel cancer driver genes and networks.

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[1–4]. Exploration of the genomic makeup of multiple cancer types has highlighted that driver somatic mutations typically involve a 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. 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 pharmacological or experimental way[8], and that can be used as therapeutic markers whose predictive ability is significantly improved when compared to that of genomic lesions in individual genes[9].

[1]European Molecular Biology Laboratory - European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK. [2]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SD, UK. [3]GlaxoSmithKline, Gunnels Wood Rd, Stevenage Herts, SG1 2NY, UK. [4]Joint Research Centre for Computational Biomedicine (JRC-COMBINE), RWTH Aachen University, Faculty of Medicine, MTZ Pauwelstrasse 19, Aachen, 52074, Germany. [5]Open Targets, Wellcome Genome Campus, Cambridge, CB10 1SD, UK. Correspondence and requests for materials should be addressed to F.I. (email: francesco.iorio@sanger.ac.uk) or J.S.-R. (email: saezrodriguez@gmail.com)

Additionally, this facilitates the stratification of cancer patients into informative subtypes[10], the characterisation of rare somatic mutations[11], and the identification of the spectrum of possible alterations underpinning a common evolutionarily successful trait acquired by a normal cell as it transforms itself into a precancerous cell and ultimately into a cancer. In two landmark papers[12,13] these traits have been summarised into a set of 11 principles, collectively referred as the *hallmarks of cancer*.

Here we propose a computational strategy, that we call SLAPenrich (Sample-population Level Analysis of Pathway Alterations Enrichments), for characterising the set of genomically altered pathways that might contribute 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). Similarly to other existing methods (such as PathScan and PathScore[14,15]), SLAPenrich aims to identify pathways that are consistently altered across the samples of a population, rather than pathways over-represented in the merged set of alterations in the population. Additionally, with respect to other existing tools, we go one step further by devising a metric to assess the predominance of alterations in pathways associated to the same canonical hallmark in each cancer type 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 mutations from the analysis. Thus we propose to use the obtained heterogeneity signatures of cancer hallmarks as a 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 this approach. Our method is implemented as an R package and publicly available at https://github.com/saezlab/SLAPenrich.

Results

Sample-population Level Analysis of Pathway Alterations Enrichments (SLAPenrich).

Problem definition and method overview. In the first step of our analysis we make use of SLAPenrich (Sample Level Analysis of Pathway alteration Enrichments): a computational method implementing an established statistical framework to perform pathway analyses of genomic datasets at the sample-population level. We have designed this tool as a means 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 categorised into two main classes[16]. The first class of approaches aims at identifying pathways whose constituent genes are significantly over-represented in the set of altered genes from all the samples of a dataset, compared with the background set of all studied genes. Many tools exist and are routinely used to perform this analysis[17–19], sometimes incorporating additional features, such as inter-gene dependencies and signal correlations[20], and also estimating single sample pathway deregulations based on transcriptional data[21]. 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 interest), the likelihood of their recurrence in the gene sets of interest is usually estimated. This is normally 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 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 mutually exclusive manner. On the basis of these cost functions, optimal sub-networks are identified and interpreted as novel cancer driver pathways[22–24]. However, at the moment there is no consensual method to rigorously define a mathematical metric for mutual exclusivity and compute its statistical significance, and a number of interpretations exist[22,23,25–27].

The problem we tackle here 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 than 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 population level (SLAPenriched) in relation to the whole cohort of patients. Therefore, 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 the mutation of a single gene of a pathway in an individual sample can be sufficient to deregulate the pathway activity. This allows pathways containing groups of genes with a tendency to be mutated in a mutually exclusive fashion (and therefore different individually mutated genes in different samples) to still be detected as enriched at the population level and further filtered based on this tendency, as additional evidence of positive selection[28]. Hence, SLAPenrich belongs roughly to the first class of computational methods described above, although it shares the mutual exclusivity consideration with the methods in the second class. 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 that expected by random chance. 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 prevents signal saturations due to hypermutated samples.

Statistical framework and implementation. 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 (Supplementary Figure S1A). 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 (Supplementary Figure S1B). The first analytical step of SLAPenrich consists in modeling the probability of such event for each individual sample. To this aim, for each sample, the likelihood of observing at least a mutation in the pathway under consideration by random chance is estimated, given the background mutation rate (for example, the number of observed mutations in the sample) and the total exonic block length of all the genes in the pathway. These individual probabilities are then aggregated in a collective test (detailed in the Methods) against the null hypothesis that the number of samples with at least one mutation in the pathway under consideration is that expected by random chance, therefore there is no association between that pathway and the disease represented by the analysed dataset.

The probability of success in each of the modeled Bernoulli trials, i.e. each sample, 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 (Supplementary Figure S1A,B) 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, or sequencing coverage and mappability) can be used through user-defined functions that can be easily integrated into 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 (Methods). 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 (Supplementary Figure S1C). 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 additional evidence of positive selection[22,23,30].

A formal description of the statistical framework underlying SLAPenrich is provided in the Methods; further details are provided in the Supplementary Methods. SLAPenrich is implemented as a publicly available R package and is fully documented at https://github.com/saezlab/SLAPenrich/. It includes a visualization/report framework enabling easy exploration of outputted enriched pathways across the analyzed samples, in a way that highlights their mutual exclusivity mutation trends, and a module for the identification of core-component genes, shared by related enriched pathways. A brief description of the SLAPenrich exported functions is included in the Supplementary Note.

Unique features of SLAPenrich. To our knowledge, there are only two other tools enabling the type of analysis supported by SLAPenrich: PathScan[14] and PathScore[15]. SLAPenrich performs comparably to both of them, showing a slightly improved ability to rank pathways containing established cancer driver genes as highly enriched. Additionally, several aspects make SLAPenrich more suitable for the analyses described in this manuscript. Particularly, PathScan does not take possible mutual exclusivity trends between patterns of mutations of genes in the same pathway into account 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 as 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[31] for PathScan, and MsigDB[32] 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[33] (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 across datasets and reference pathway gene sets, to perform mutual exclusivity sorting of binary matrices, and to identify pathway core-components (i.e. subsets of genes leading the enrichment of different pathways).

These and other aspects are discussed in the Supplementary Methods, together with results from applying SLAPenrich to a case study Lung Adenocarcinoma Dataset to identify pathways that are differentially enriched across subpopulations of Smokers/non-smokers and mucinous/non-mucinous bronchioalveolar types, and from a systematic comparison of SLAPenrich, PathScan and PathScore (Supplementary Tables S1–S5 and Supplementary Figures S2–S4).

SLAPenrich analyses across different cancer types. Leveraging the capacities of SLAPEnrich, we set out to perform a systematic large-scale analysis of pathway alterations in cancer. To this aim we used a collection of pathways from the Pathway Commons data portal (v8, 2016/04)[33] (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[34], from 4,415 patients across 10 different cancer types, from publicly 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),

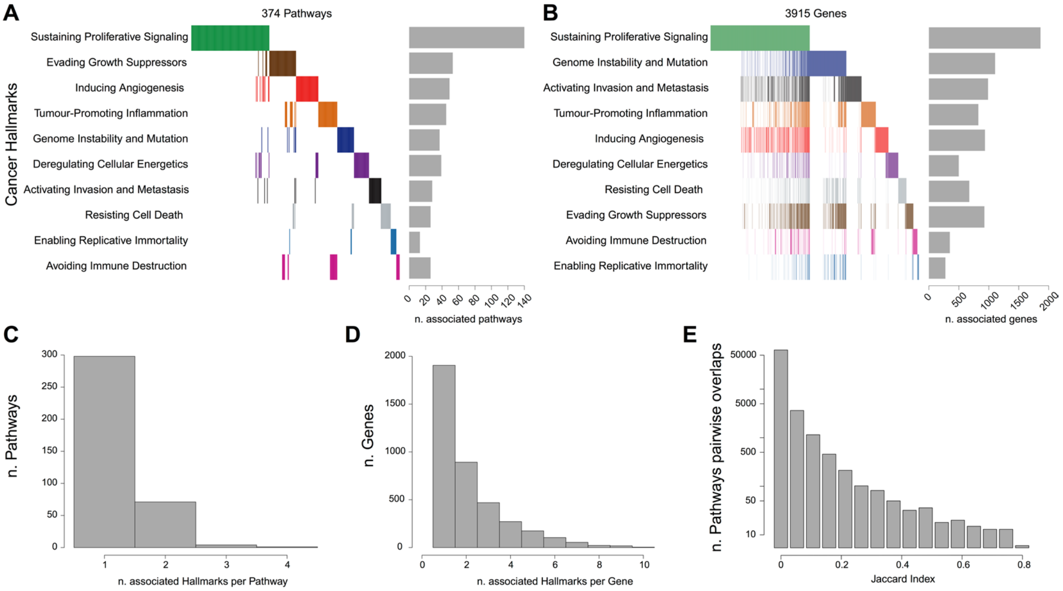

Figure 1. Manually curated mapping between genes, pathways and hallmarks. (**A**) Heatmap with cancer hallmarks on the rows, pathways gene sets on the columns. A coloured bar in position (i, j) indicates that the j-th pathway is associated with the i-th hallmark; bar diagram on the right shows the number 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.

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 Supplementary Table S6.

We tested the stability of SLAPenrich with respect to variations in mutation calling reliability, evaluating the effect of random noise and errors at the level of the SLAPenrich input matrices. To this aim, we increasingly introduced (respectively removed) uniformly distributed false-positives (respectively true-positives) mutations, in each of the 10 analysed genomic datasets. This was performed simulating a reduction of mutation call sensitivity (respectively, specificity) to 95, 80, 70, and 50%, producing 10 noise-inflated versions of the considered dataset for each reduction level. Subsequently, we executed a SLAPenrich analysis on each of these datasets and compared the sets of outputted pathways with those obtained when running SLAPenrich on the corresponding original datasets. Results from these analyses (detailed in the Supplementary Methods) are shown in Supplementary Figure S5. They highlight that the output of SLAPenrich is highly stable with respect to the introduced noise (median area under the Receiver Operating Characteristic (ROC) curves stably over 0.995 for all the tested ratios of Variants False Positives, and over 0.99 for all the tested ratios of Variants False Negatives).

Mapping pathway enrichments onto canonical cancer hallmarks. Subsequently, we reasoned that since the main role of cancer driver alterations is to enable cells to achieve a series of phenotypic traits termed the *cancer hallmarks*[12,13], that can be linked to gene mutations[35], it would be informative to group the pathways according to the hallmark they are associated to. Towards this end, through a computer-aided manual curation (see Methods and Supplementary Table S7) we were able to map 374 gene-sets (from the most recent release of pathway commons[33]) to 10 cancer hallmarks[12,13] (Figure 1AB), for a total number of 3,915 genes (included in at least one gene set associated to at least one hallmark; Supplementary Table S8). The vast majority (99%, 369 sets) of the considered pathway gene-sets were mapped on two hallmarks at most, and 298 of them (80%) was mapped onto one single hallmark (Fig. 1C). 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 (Fig. 1D). Finally, as shown in Fig. 1E, 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

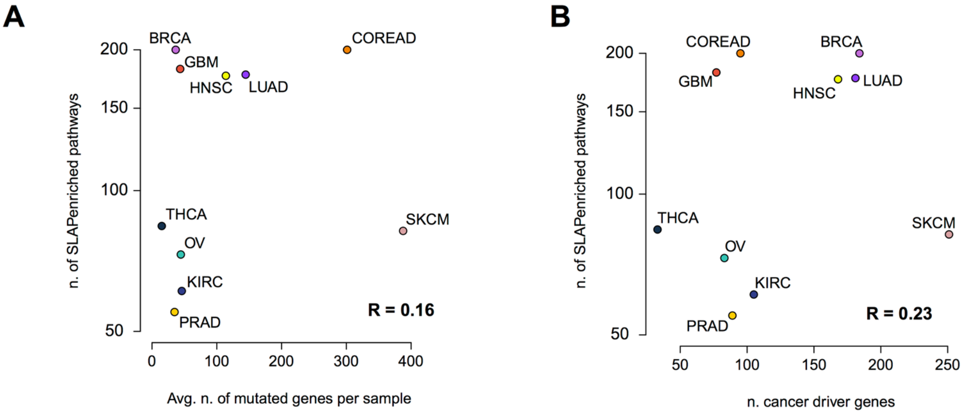

Figure 2. Number of SLAPenrichments versus mutation burdens and number of established cancer genes. (**A**) Number of pathways enriched at the population level across cancer types compared with the average number of mutated genes and (**B**) the average number of high confidence cancer driver genes.

pathway enrichments into canonical cancer hallmarks through this curation covered 46% of significant results on average across cancer types (Supplementary Figure S6A).

We observed a weak correlation ($R = 0.53$, $p = 0.11$) between the number of hallmark-associated (HMA) enriched pathways across the different analyses and the number of available samples in the analysed dataset (Supplementary Figure S6B), but a down-sampled analysis showed that our results are not broadly confounded by the sample sizes (see Methods and Supplementary Figure S6C).

We investigated how our HMA-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[34,36] (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 HMA-pathway, across the 10 different tissues analyses (median percentage = 63.5, range = 88.5%, for BRCA, to 28.7% for SKCM) (Supplementary Figure S6D).

Interestingly, we found that the number of HMA-SLAPenriched pathways per cancer type (median = 130, range = 55 for PRAD, to 200 for BRCA and COREAD) was independent of 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 $R = 0.16$ ($p = 0.65$), Fig. 2A, as well as from the number of high confidence cancer driver genes (as predicted in[36], median = 100, range from 33 for THCA to 251 for SKCM, Fig. 2B). Particularly, THCA has the lowest average number of mutations per sample (15.03), but there are 4 tissues with a lower number of HMA-pathways enriched. 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 (Fig. 2A). 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 HMA-pathways for this set of cancer types we can clearly distinguish two separate groups (Fig. 2A). The first group includes BRCA and GBM that seem to have a more heterogeneous set 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 also be indicative of different subtypes with dependencies on different pathways (and at least for BRCA this is expected) but could also be biased by the composition of the analysed cohorts being representative of selected subtypes only.

Genomic heterogeneity of cancer hallmarks' acquisition across cancer types.

Inspecting the sets of enriched HMA-pathways across the performed analyses allowed us to explore how different cancer types might acquire the same hallmark by selectively altering different pathways. Heatmaps in Fig. 3, and Supplementary Figure S7 (one per each hallmark) show a different level of enrichment of pathways associated with the same hallmark across different tissues, with clearly distinguishable patterns and well-defined clusters.

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[37]. 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[38].

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[39] is exclusively and preferentially altered in this cancer type.

The acquisition of the *Genome Instability and mutation* hallmark seems to be dominated in COREAD by alterations in the *HDR Through Single Strand Annealing (SSA)*, *Resolution Of D Loop Structures Through Synthesis Dependent*

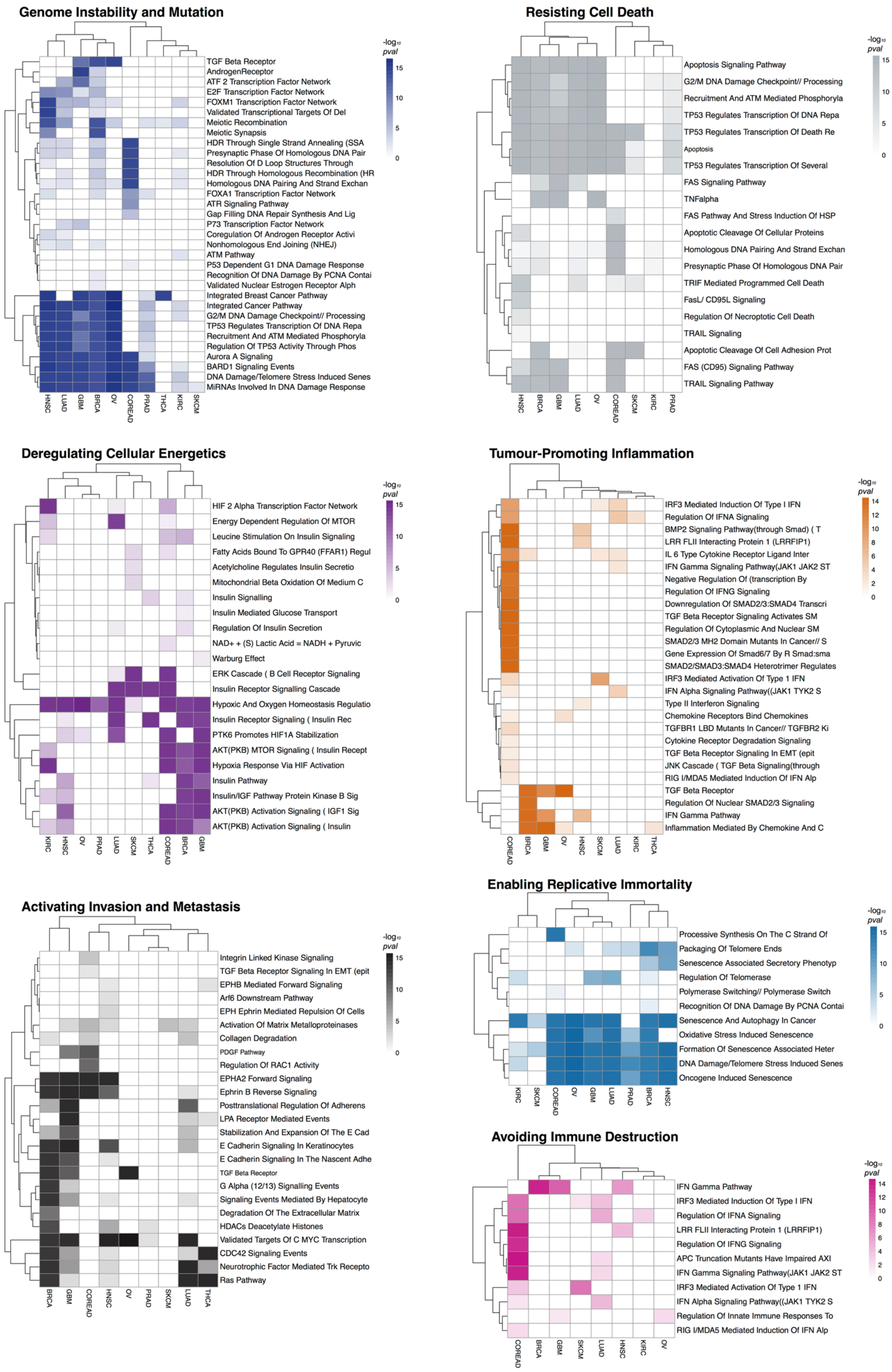

Figure 3. Heterogeneity of hallmark acquisition across cancer types. Heatmaps showing pathways enrichments at the population level across cancer types for individual hallmarks (representative cases). Color intensities correspond to the enrichment significance. Cancer types and pathways are clustered using a correlation metric. See also Supplementary Figure 7.

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[40].

Finally, the heatmap for *Genome Instability and Mutation* shows nearly no enriched pathways associated to the acquisition of this hallmark in 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, genomic instability in SKCM originates from cell extrinsic processes such as UV light exposure[41].

The maintenance of genomic integrity is guarded by a network of damage sensors, signal transducers, and mediators, and 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[42]. Our analysis strikingly detects *MiRNAs Involved In DNA Damage Response* as the unique pathway associated to *Genome instability and mutation* in SKCM. This suggests that mutations in this pathway, involving ATM (as the most recurrently mutated gene, and known to induce miRNA biogenesis following DNA damage[43]), impair the ability of melanocytes to properly respond to insults from UV light and may have a significant role in the tumourigenesis of melanoma.

The *Avoiding Immune destruction* heatmap (Fig. 3) highlights a large number of pathways selectively enriched in COREAD, whereas 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[44], melanoma[45] or renal-cell carcinoma[46]. 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[47].

Moreover, in the context of COREAD, the *Tumor-promoting inflammation* heatmap (Fig. 3) also highlights several pathways predominantly and very specifically 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[48]. Indeed, a number of clinical trials evaluating the utility of inflammatory and cytokine-modulatory therapies are currently underway in colorectal cancer[49,50]. 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 (Fig. 3). 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[51]. 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 altered pathways mapped to each cancer hallmark in a given tumor type, via a cumulative heterogeneity score (CHS). The CHS of a hallmark is computed as the proportion of the pathways associated to that hallmark that are significantly enriched. We hypothesize that a large CHS points to the exploitation of many evolutionary trajectories pursued to acquire a defined hallmark. This might suggest that the hallmark with a higher CHS is more advantageous evolutionary than others for the cancer type under consideration.

The pattern of CHSs per cancer hallmark in a cancer type gives its *hallmark heterogeneity signature* (Fig. 4). 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[52], for *Tumour-promoting inflammation* and *Avoiding immune-destruction* in COREAD[53]. Lastly and as expected, for *Sustaining proliferative-signaling* and *Enabling replicative immortality*, the key hallmarks in cancer initiation[12], high CHSs are observed across the majority of the analysed cancer types.

Taken together, these results show the potential of our pipeline 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 putative acquisition of the cancer hallmarks across tissue types, confirming several known relations between cancer types, and pinpointing preferentially altered pathways.

Hallmark heterogeneity analysis points at novel cancer driver genes and networks. To investigate the potential of our computational method in identifying novel cancer driver genes and networks, we evaluated first to what extent the identified enriched HMA-pathways were dominated by somatic mutations in established high-confidence cancer genes (HCGs)[36] 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 samples with mutations in HCGs in *P* and the number of samples with mutations in any gene in *P*. Results of this analysis are shown in Supplementary Figures S7 and S8. We observed a median of 15% of pathway enrichments, across hallmarks, with a 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 did not involve somatic mutations in HCGs (range from 0.25% for *Evading Growth Suppression* to 15% for *Avoiding Immune Destruction*). Across all the hallmarks, the cancer type with the lowest median HCG-dominance was KIRC (33%), whereas that with the highest was THCA (91%).

Subsequently, we re-analysed the TCGA data excluding all the variants involving HCGs from each cancer type (from now the *filtered analysis*). Results from this exercise (Fig. 5, Supplementary Table S9 and Supplementary Figure S10), 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 Fig. 5 and Supplementary Figure S10) was 21%, (range from 2.1% for GBM to 56.2% for COREAD). However, several HMA-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 (Fig. 5, Supplementary Figure S10).

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 the removal of known drivers, then the filtered analyses might have uncovered novel cancer driver networks composed by infrequently mutated genes. In fact, these new gene modules are typically composed by groups of functionally

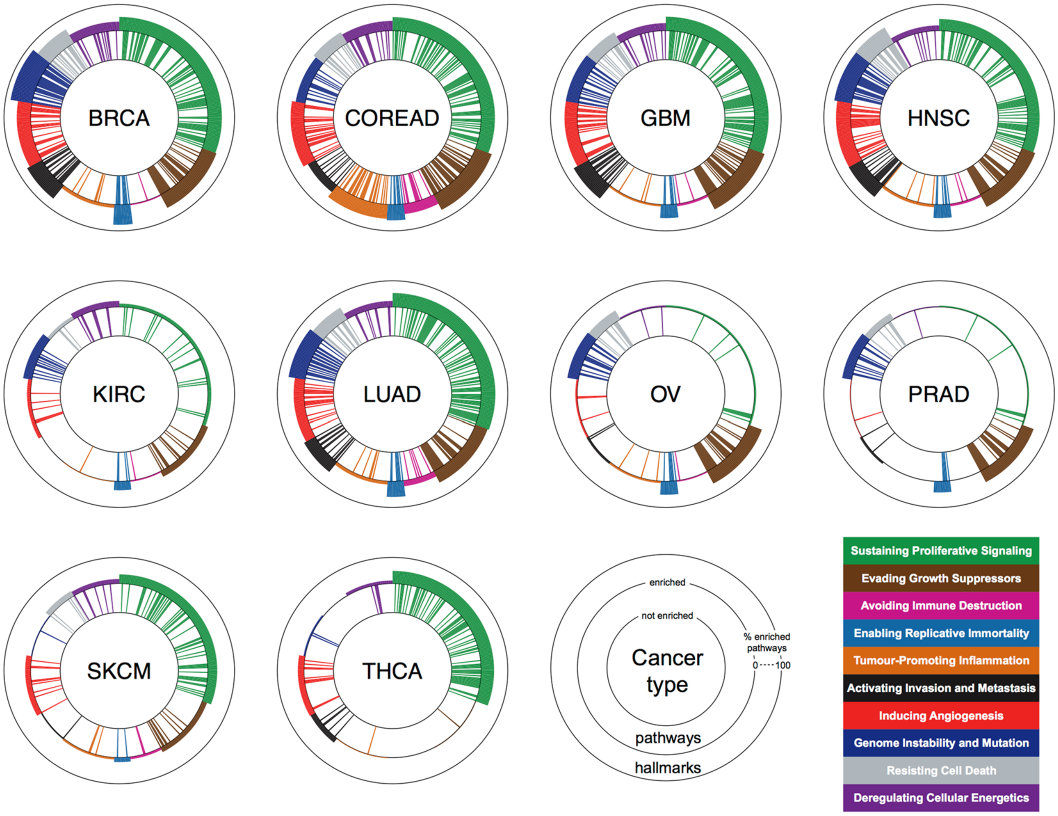

Figure 4. Cancer hallmark heterogeneity signatures. Each cancer hallmark signature plot is composed of three concentric circles. Bars between the inner and middle circles indicate pathways, bars between the middle and external circle indicate cancer hallmarks. Different colors indicate different cancer 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 of enriched pathways over those associated with that hallmark.

interconnected and very lowly frequently mutated genes (examples are shown in Fig. 6 and the whole bulk of identified network is included in the Supplementary Results).

An example is given 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%) (Fig. 6A), 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[54]. 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[36]) 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[55]. 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[56]. The remaining altered component of this pathway is mostly made of a network of very lowly frequently mutated (and in a highly mutually 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* (Fig. 6B). IL-6-type cytokines have been observed to modulate cell growth of several cell types, including melanoma[57]. Increased IL-6 blood levels in melanoma patients correlate with disease progression and lower response to chemotherapy[58]. Importantly, studies proposed OSMR, an IL-6-type cytokine receptor, to play a role in the prevention of melanoma progression[59], and as a novel potential target in other cancer types[60]. Consistent with these findings, OSMR is the member of this pathway with the largest number of mutations in the SKCM cohort (Fig. 6B), complemented by a large number of other lowly frequently mutated genes (most of which are interleukins).

In the context of melanoma, we observed two other highly enriched pathways in the filtered analysis: *PDGF receptor signaling network* (FDR = 2.7%) (Fig. 6C) and *Neurophilin Interactions with VEGF And VEGFR* (0.21%) (Fig. 6D), both associated with the *Inducing angiogenesis* hallmark. Mutations in all of 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

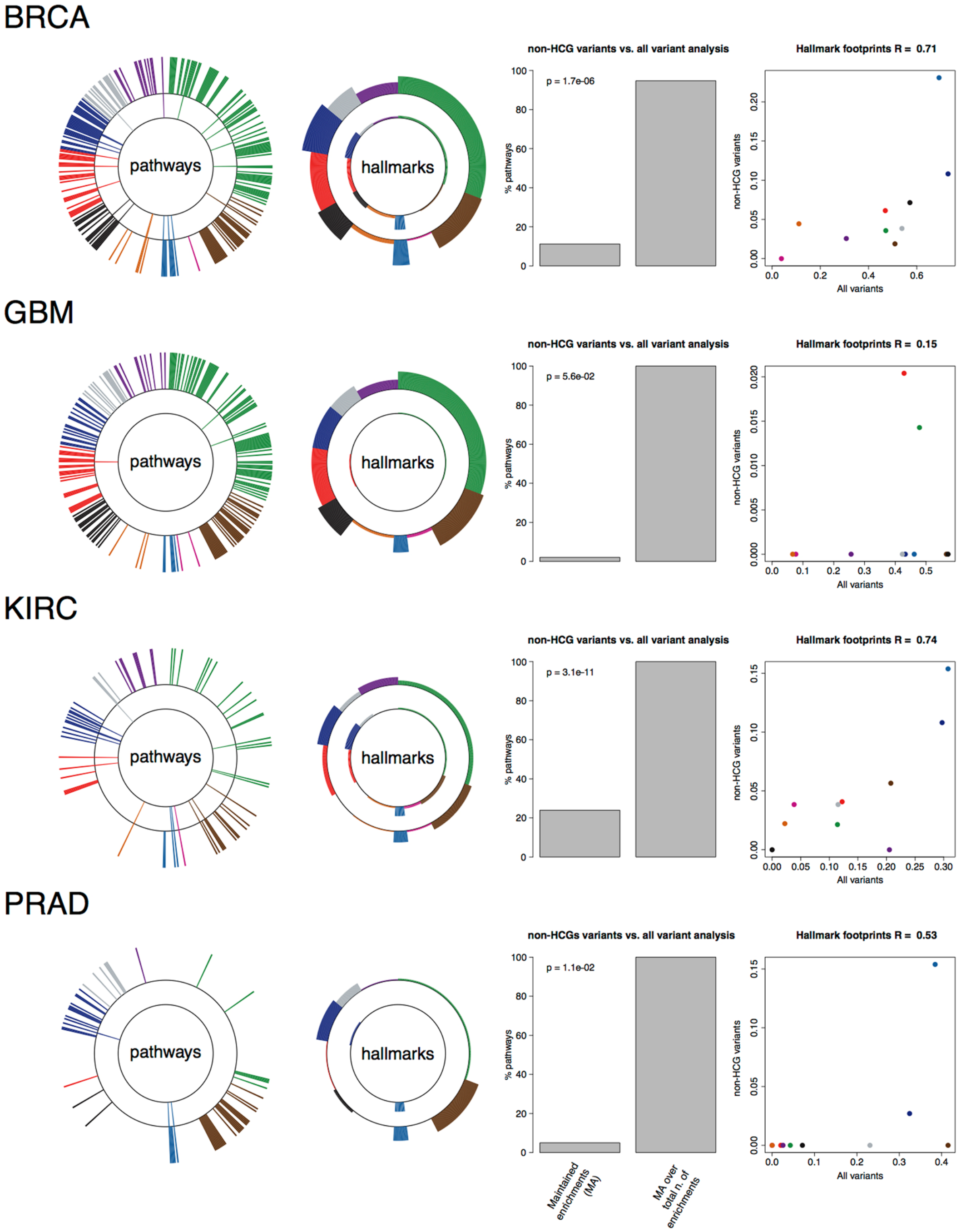

Figure 5. Hallmark heterogeneity signature analysis including and not including known cancer driver genes. In each row, the first circle plot show pathway enrichments at the population level when considering all the somatic variants (bars on the external circle) and when considering only variants not 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-predictive-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.

these two pathway enrichments is characterised by patterns of highly mutually exclusive somatic mutations in Platelet-derived growth factor (PDGF) genes, and their corresponding receptors: a network that has been recently proposed as an autocrine endogenous mechanism involved in melanoma proliferation control[61].

A final example is given by the enriched pathway *Regulating the activity of RAC1* (associated with the *Activating Invasion and Metastasis* hallmark) in COREAD (Fig. 6E). 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[62]. Previous *in vitro* and *in vivo* studies in prostate cancer demonstrated a marked increase in RAC1 activity in cell

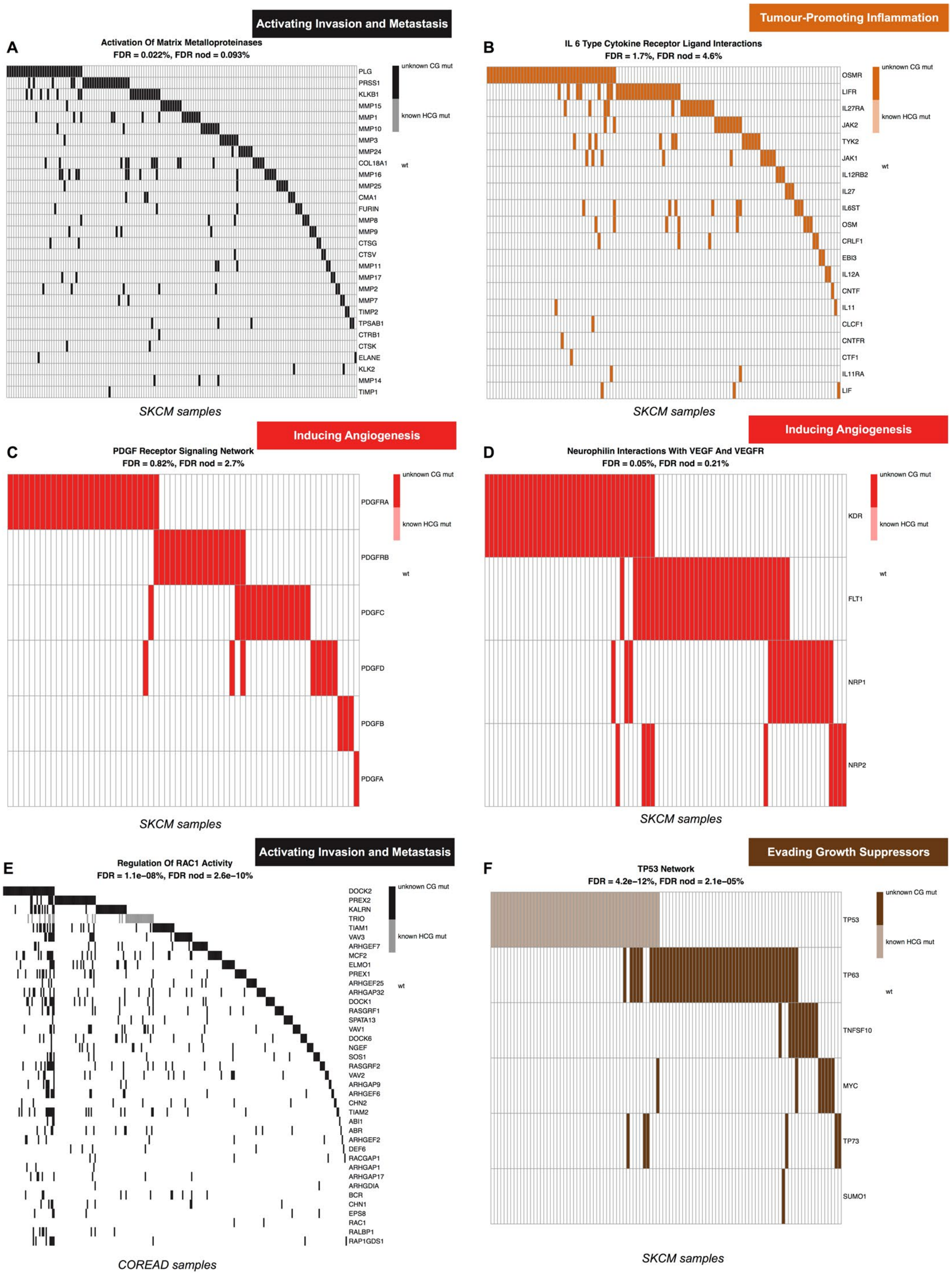

Figure 6. Example of potential novel cancer genes and networks. Picked examples of novel putative cancer driver genes and networks. The first FDR value refers to the unfiltered analysis, whereas the second FDR refers to the filtered one (in which variants involving high confidence and highly frequently mutated cancer driver genes have been removed).

migration and invasion, and that RAC1 inhibition immediately stopped these processes[63,64]. 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[65].

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 (Fig. 6F). Whereas alterations in some nodes of this network are known to be an alternative to p53 repression, conferring chemoresistance and poor prognosis[66], dissecting the functional relations between them is still widely considered a formidable challenge[67]. 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 our approach in identifying potential novel 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-source R package (SLAPenrich) to identify genomic alterations in biological pathways, which putatively contribute to the acquisition of the canonical cancer hallmarks. Our statistical framework does not seek 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 mutually exclusive manner across samples.

With this method 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 heterogeneous sets of pathways across cancer types. This has confirmed the established predominance of certain hallmarks in defined cancer types, and it 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 the derivation of 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, the limitation of our knowledge of pathways, and the possibility that pathways that we were not mapped onto cancer hallmarks in our curation could correspond to specific capabilities of cancer cell in certain tumour types. 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 could account for structural variants such as small indels and copy number alterations, known to play an important role in cancer.

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 that are publicly available from 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, as reported in a recent publication[68]. More generally, it can be used to characterize, at the pathway level, any type of biological dataset that can be modeled as a presence/absence matrix, where genes are on the rows and samples are on the columns.

Methods

Formal description of the SLAPenrich statistical framework.

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

$$f(g_i,\ s_j) = \{1 \text{ if } g_i \text{ harbours a somatic mutation in } s_j \text{ and } 0 \text{ otherwise}\}. \tag{1}$$

Given the set of all the genes whose products belong to the same pathway P, we aim at assessing if there is a statistically significant tendency for the samples in S to carry mutations in P. Importantly, we do not require the genes in P to be significantly enriched in those that are altered in any individual sample nor in the subset of G composed by all the genes harbouring at least one somatic mutation in at least one sample. In what follows P will be used to indicate the pathway under consideration as well as the corresponding set of genes, interchangeably. We assume that P is altered in sample s_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 (Supplementary Figure S1B). 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. $\phi(X) = \{1$ if $X > 0$, and 0 otherwise$\}$, then summing the $\phi(X_i)$ across all the samples in S, gives the observed number of samples harbouring a mutation in at least one gene belonging to P:

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

A pathway alteration index, quantifying the deviance of $O(P)$ from its expectation, and thus how unexpected 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 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, …, 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 (Supplementary Figure S1C):

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

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[69].

SLAPenrich is implemented as an R package 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. The hallmark signature analyses were performed on a large collection of pathway gene sets from the Pathway Commons data portal (v8, 2016/04)[33] (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[70] 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:

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

To guarantee results' comparability with respect to previously published studies, for the case study analysis on the LUAD dataset we downloaded and used the whole collection of KEGG[31] pathway gene sets from MsigDB[32], encompassing 189 gene sets for a total number of 5,224 genes included in at least one set.

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 with 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 keyword 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 S7. Results of this data curation are reported in the Supplementary Table S8.

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

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

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[34] (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[36] and reannotated using a pipeline consistent with the COSMIC database[71] (Vagrent: https://zenodo.org/record/16732#.VbeVY2RViko).

Lists of tissue-specific high-confidence cancer genes[36] 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[72,73] and further filtered as described in[34] (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[33], 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 mutually 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.

References

1. Weinstein, J. N. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).
2. Consortium, T. I. C. G. *et al*. PERSPECTIVES. *Nature* **464**, 993–998 (2010).
3. Garnett, M. J. *et al*. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575 (2012).
4. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
5. Lawrence, M. S. *et al*. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
6. Garraway, L. A. & Lander, E. S. Lessons from the Cancer Genome. *Cell* **153**, 17–37 (2013).
7. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
8. Pe'er, D. & Hacohen, N. Principles and strategies for developing network models in cancer. *Cell* **144**, 864–873 (2011).
9. Shi, W. *et al*. Pathway level alterations rather than mutations in single genes predict response to HER2-targeted therapies in the neo-ALTTO trial. *Annals of Oncology* mdw434 (2016).
10. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nature Methods* **10**, 1108–1115 (2013).
11. Leiserson, M. D. M. *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics* **47**, 106–114 (2015).
12. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
13. Vogelstein, B. *et al*. Cancer Genome Landscapes. *Science (New York, NY)* **339**, 1546–1558 (2013).
14. Wendl, M. C. *et al*. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27**, 1595–1602 (2011).
15. Gaffney, S. G. & Townsend, J. P. PathScore: a web tool for identifying altered pathways in cancer data. *Bioinformatics* (2016).
16. Creixell, P. *et al*. Pathway and network analysis of cancer genomes. *Nature Methods* **12**, 615–621 (2015).
17. Reimand, J., Arak, T. & Vilo, J. g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research* **39**, W307–W315 (2011).
18. 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**, 48 (2009).
19. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**, 44–57 (2008).
20. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research* **40**, e133–e133 (2012).
21. 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**, 6388–6393 (2013).
22. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research* **22**, 398–406 (2012).
23. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. *Genome Research* **22**, 375–385 (2012).
24. Schubert, M. & Iorio, F. Exploiting combinatorial patterns in cancer genomic data for personalized therapy and new target discovery. *Pharmacogenomics* **15**, 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**, 57–63 (2016).
26. Lu, S. *et al*. Identifying Driver Genomic Alterations in Cancers by Searching Minimum-Weight, Mutually Exclusive Sets. *PLoS computational biology* **11**, e1004257 (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**, 2605–2622 (2008).
29. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**, 175–181 (2011).
30. Thomas, R. K. *et al*. High-throughput oncogene mutation profiling in human cancer. *Nature genetics* **39**, 347–351 (2007).
31. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–62 (2016).
32. Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545 (2005).
33. Cerami, E. G. *et al*. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* **39**, D685–90 (2011).
34. Iorio, F. *et al*. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* (2016).
35. 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**, 439–449 (2015).
36. Rubio-Perez, C. *et al*. *In silico* prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382–396 (2015).
37. Manié, E. *et al*. Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas. *International Journal of Cancer* **138**, 891–900 (2016).
38. Walsh, C. S. Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecologic oncology* **137**, 343–350 (2015).
39. Yu, X. *et al*. Androgen receptor signaling regulates growth of glioblastoma multiforme in men. *Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine* **36**, 967–972 (2015).
40. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. (2010).
41. Alexandrov, L. B. *et al*. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
42. Syed, D. N., Khan, M. I., Shabbir, M. & Mukhtar, H. MicroRNAs in skin response to UV radiation. *Current drug targets* **14**, 1128–1134 (2013).
43. 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**, 371–383 (2011).

44. Garon, E. B. *et al*. Pembrolizumab for the treatment of non-small-cell lung cancer. *The New England journal of medicine* **372**, 2018–2028 (2015).
45. Robert, C. *et al*. Pembrolizumab versus Ipilimumab in Advanced Melanoma. *The New England journal of medicine* **372**, 2521–2532 (2015).
46. Motzer, R. J. *et al*. Nivolumab versus Everolimus in Advanced Renal-Cell Carcinoma. *The New England journal of medicine* **373**, 1803–1813 (2015).
47. Le, D. T. *et al*. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *The New England journal of medicine* **372**, 2509–2520 (2015).
48. 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**, 639–645 (2012).
49. West, N. R., McCuaig, S., Franchini, F. & Powrie, F. Emerging cytokine networks in colorectal cancer. *Nature reviews. Immunology* **15**, 615–629 (2015).
50. Lasry, A., Zinger, A. & Ben-Neriah, Y. Inflammatory networks underlying colorectal cancer. *Nature immunology* **17**, 230–240 (2016).
51. Poteet, E. *et al*. Reversing the Warburg effect as a treatment for glioblastoma. *Journal of Biological Chemistry* **288**, 9153–9164 (2013).
52. Pikor, L., Thu, K., Vucic, E. & Lam, W. The detection and implication of genome instability in cancer. *Cancer metastasis reviews* **32**, 341–352 (2013).
53. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140**, 883–899 (2010).
54. Deryugina, E. I. & Quigley, J. P. Matrix metalloproteinases and tumor metastasis. *Cancer metastasis reviews* **25**, 9–34 (2006).
55. Rabbani, S. A. & Mazar, A. P. The role of the plasminogen activation system in angiogenesis and metastasis. *Surgical oncology clinics of North America* **10**, 393–415-x (2001).
56. Kumari, S. & Malla, R. New Insight on the Role of Plasminogen Receptor in Cancer Progression. *Cancer growth and metastasis* **8**, 35–42 (2015).
57. Zarling, J. M. *et al*. 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**, 9739–9743 (1986).
58. Tartour, E. *et al*. 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**, 911–913 (1994).
59. Lacreusette, A. *et al*. Loss of oncostatin M receptor beta in metastatic melanoma cells. *Oncogene* **26**, 881–892 (2007).
60. Caffarel, M. M. & Coleman, N. Oncostatin M receptor is a novel therapeutic target in cervical squamous cell carcinoma. *The Journal of pathology* **232**, 386–390 (2014).
61. Faraone, D. *et al*. Platelet-derived growth factor-receptor alpha strongly inhibits melanoma growth *in vitro* and *in vivo*. *Neoplasia (New York, NY)* **11**, 732–742 (2009).
62. Yamazaki, D., Kurisu, S. & Takenawa, T. Regulation of cancer cell motility through actin reorganization. *Cancer science* **96**, 379–386 (2005).
63. 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**, 1925–1934 (2013).
64. Bailey, C. L., Kelly, P. & Casey, P. J. Activation of Rap1 promotes prostate cancer metastasis. *Cancer research* **69**, 4962–4968 (2009).
65. 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**, 9820–9833 (2015).
66. Matin, R. N. *et al*. p63 is an alternative p53 repressor in melanoma that confers chemoresistance and a poor prognosis. *The Journal of experimental medicine* **210**, 581–603 (2013).
67. Costanzo, A. *et al*. TP63 and TP73 in cancer, an unresolved "family" puzzle of complexity, redundancy and hierarchy. *FEBS letters* **588**, 2590–2599 (2014).
68. Brammeld, J. S. *et al*. Genome-wide chemical mutagenesis screens allow unbiased saturation of the cancer genome and identification of drug resistance mutations. *Genome Research* (2017).
69. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445 (2003).
70. Wain, H. M. *et al*. Guidelines for human gene nomenclature. *Genomics* **79**, 464–470 (2002).
71. Forbes, S. A. *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43**, D805–11 (2015).
72. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Research* **40**, e169 (2012).
73. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).

Acknowledgements

OpenTargets funding to JSR (Projects OpenTargets15 and OpenTargets16). We would like to thank Jorge Buendia, Marco Ranzani, David Adams, 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 Contributions

F.I. 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; L.G.A. 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; J.B. contributed to testing the R package, interpreted results and findings, contributed to manuscript writing and revising; I.M. contributed to the design of the validation analyses, read and edited the manuscript; D.R.W. contributed to the design of the statistical framework and supervised its mathematical formalization; U.M. contributed to the interpretation of results; J.S.R. supervised the study and contributed to the manuscript writing and revising.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-25076-6.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.