Identifying Cancer-Associated Genes across Seven Tumor Types using Topological Data Analysis

Udi E. Rubin

Submitted in partial fulfillment of the
requirements for the degree of
Masters of Arts
in the Graduate School of Arts and Sciences

Program in Biotechnology
Department of Biological Sciences

COLUMBIA UNIVERSITY
2016

ABSTRACT

Multidimensional data collected by large consortia such as The Cancer Genome Atlas Network (TCGA) provides sufficient statistical power for cross-sectional studies to identify cancer-associated genes with the aim to unravel complex cancer mechanisms involving clinically actionable targets. However, current methods employed in cross-sectional studies suffer from two main caveats. First, they do not account for the data's continuous structure when reducing its dimensionality, resulting in possible artifacts. Second, they rely on complex modeling of the mutational background to identify cancer-associate genes which might introduce systematic errors resulting in false positives. We have developed an orthogonal method involving topological data analysis and original statistics, which is not prone to the mentioned caveats. Using topological data analysis, specifically the Mapper algorithm, we map the global gene expression space into a two-dimensional space and apply to it a novel statistical algorithm which does not rely on complex modeling, but rather on mutations association with the disease phenotype, as captured in the two-dimensional space. We applied this method to 2,916 tumors spanning seven cancer types and identified in total 85 cancer-associated genes, including 38 novel candidates such as *FMN2* mutations in the p53/Rb pathway in lung adenocarcinoma and *PTPRD* mutations in urothelial bladder cancer.

TABLE OF CONTENTS

List of Figures

List of Tables

Acknowledgements

I would like to thank my thesis advisor Dr. Raul Rabadan, who kindly took me into his lab and supported me in completing this project by sharing his vast experience, acumen, and sage advice. I would also like to thank my program director, Dr. Lili Yamasaki, for supporting me in pursuing this research-based thesis and for always guiding me along the right path with a smile. Finally, I would like to thank Dr. Pablo G. Camara, who conceived the methodology and designed the analysis in this project, for mentoring me through the maturation of this work with utmost patience, intelligence, and kindness. Above all, I would like to acknowledge the trust he placed in me as a true collaborator, resulting in one of the intellectually richest periods of my life.

Dedications

I would like to dedicate this work firstly to my parents, Esther and Zvi (Julian) Rubin, who have always inspired and nourished my sense of curiosity - the driving force in pursuing my dreams.

To my mother and father-in-law, Yael and Shuki Goldwasser, whose unconditional support and generosity, made this dream, of a graduate education at Columbia University, a reality.

Ultimately, I would like to dedicate this work to my wife, Nitzan, the love of my life, without her by my side, all of this would have been meaningless to me.

BACKGROUND

General biology and epidemiology of cancer

Cancer is a group of more than one hundred diseases characterized by unregulated proliferation and spread of cells. If the abnormal growth, a.k.a the tumor, is not controlled, it can promote local or remote damage and might result in severe illness or death. Cancer can arise from different specialized cells within the body, including epithelial (known as carcinoma), mesenchymal (sarcoma), hematopoietic (lymphoma, leukemia, myeloma) and neuroectodermal (glioma, blastoma) tissues (*1*).

Alarmingly, according to the latest global report by the World Health Organization (*2*), cancer is one of the leading causes of mortality and morbidity worldwide with approximately 8.2 million cancer-associated deaths (not including non-melanoma skin cancer) and 14.1 million new cases in 2012 alone (*2*). By 2030, those rates are expected to rise to 20 million new cases and 13 million deaths simply due to aging and natural population growth. However, it is safe to assume that the numbers will grow even larger due to the adoption of lifestyles associated with cancer, such as smoking, poor diet, and fewer pregnancies, in developing countries (*2*). Global incidence and mortality rates per primary site are detailed below in Figure 1 (*2*).

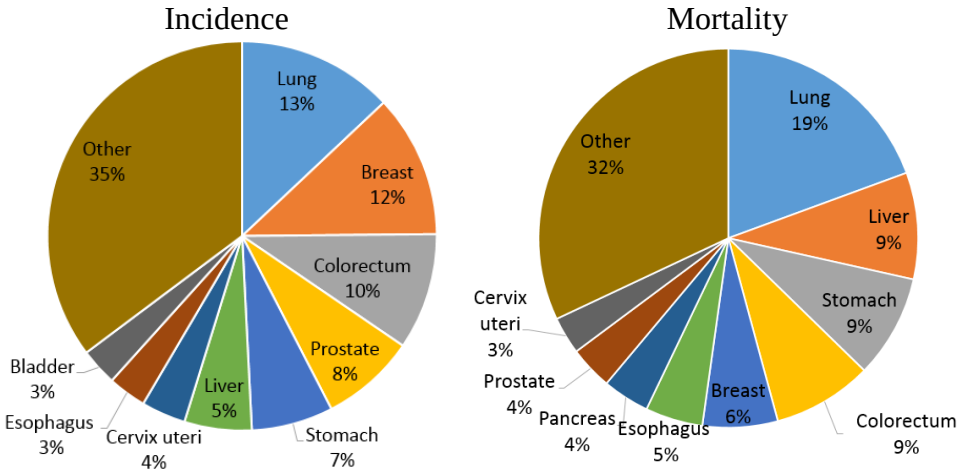


Figure 1 | Cancer incidence and mortality rates. The labels correspond to the primary site of origin and percentage among all cases and both sexes in 2012. Adapted from (*2*).

Cancer evolves in an evolutionary process initiated by acquired genomic mutations and natural selection that acts upon the phenotypic outcome (*3*). In other words, if a mutation endows the cell with a survival advantage over other cells in the tissue, cancer might develop. The mechanisms by which a cell acquires mutations varies, and to name a few, mutations are introduced by viral agents, exposure to environmental carcinogens, and through random errors introduced by DNA polymerase during cell replication. Among the common genomic alterations are point mutations, copy number variations and genomic fusions (*3*). While most of the acquired mutations have no functional impact, or can be an essential part of the development of a benign local mass (*3*), occasionally, a cell acquires a powerful set of advantageous capabilities that promotes cells spreading to nearby and remote tissues, rendering it malignant. These capabilities are induced by mutational processes known as the *"hallmarks of cancer"* as introduced in Table 1 (*4*).

Table 1 | Established hallmarks of cancer and mechanisms. Adapted from (4).

Cancer hallmark	Inducing mechanism
Sustaining proliferative signaling	Excessive binding of growth factors to cell surface receptors or disruption of growth negative feedback loops.
Evading growth suppressors	Silencing mutation of RB and TP53.
Resisting cell death	Evading apoptosis through silencing mutations of TP53.
Enabling replicative immortality	Circumventing telomere-induced senescence through activation of telomerase.
Inducing angiogenesis	Up-regulation of VEGF pathways
Activating invasion and metastasis	Expression of N-cadherin, a migration promoting molecule. Downregulation of E-cadherin, a cell adhesion molecule.

Although we have made progress in untangling the mysteries of this complex disease, we are still witnessing millions of cancer associated mortalities each year. Hence, a deeper understanding of the disease would be instrumental in improving patient's outcome. One way to accelerate this process is by conducting cross-sectional cancer studies using data collected from multiple tumor types and thousands of patients, as made available by large consortia such as The Cancer Genome Atlas (TCGA) (*5*) and the International Cancer Genome Consortium (ICGC) (*6*). This approach provides the required statistical power to generate a more comprehensive molecular profiling of the disease with the potential to unravel novel and clinically actionable mechanisms (*7*).

Cross-sectional studies and their importance in cancer research

Several consortia have undertaken the mission to assemble an integrative profile of cancer's molecular and clinical landscape with a goal to spur new therapeutic, preventive and diagnosis strategies for cancer patients. The multivariate data collected by large consortia such as The Cancer Genome Atlas (TCGA) (*5*) and the International Cancer Genome Consortium (ICGC) (*6*) includes, among others, genomic, epigenomic, transcriptomic, proteomic and clinical data from thousands of patients across dozens of tumor types. The resulting rich data provides the required statistical power to build an integrated picture of commonalities, differences and rising themes across tumor lineages, otherwise undetectable. Specifically, by integrating this data, we can now more easily identify new tumor subtypes, driver mutations (including rare somatic mutations), pathways, prognostic biomarkers, and increase our therapeutic repertoire either by identifying new therapeutic targets or potentially, by assigning treatment effective in one tumor type to another based on newly identified commonalities.

In this capacity, integration of gene expression levels and genomic events (Copy number variations and somatic mutations) of 200 Glioblastoma multiforme (GBM) patients in a cross-sectional study led by TCGA revealed clinically relevant subtypes and their associated genomic alterations. The four identified subtypes (Proneural, Neural, Classical, and Mesenchymal) differ regarding response to therapeutic treatment and survival patterns (*8*). Similarly, two independent studies of 230 lung adenocarcinoma (*9*) and 131 invasive urothelial bladder (*10*) carcinomas revealed activating mutations in the growth factor *ERBB2* (*HER2*) which had been reported implicated in HER2-positive metastatic breast (*11*) and gastric cancer (*12*). Indeed, a derivative of Trastuzumab, a monoclonal antibody that is already being used for treating HER2-positive gastric (*13*) and metastatic breast cancer (*14*), is in a recruiting phase of a clinical trial (ClinicalTrials.gov id: NCT02675829) for the treatment of other HER2 amplified or mutant cancers, including bladder and lung cancer.

Naturally, the increased availability of high-quality data invites more studies and raises demand for advanced analytical methods that can exploit the statistical power associated with large cohorts. Indeed, dozens of new methods with the aim to identify cancer-associated genes (*15*) or cancer-associated pathways (*16*) have been developed in the past decade.

Common methods for identifying cancer-associated genes and limitations

A common approach for identifying cancer-associated genes relies on the reasoning that if a gene is more frequently mutated than expected (as defined by a background distribution of the mutational landscape), it implies a sort of positive selection for the mutation in the overall oncogenic process. MutSigCV, by Broad Institute (*17*), and

Genome MuSiC by Washington University School of Medicine (*18*), take this approach and have been employed prevalently in prime cross-sectional studies (*7*).

In earlier versions of MutSigCV the background mutational process was averaged and unified genome-wide, and every gene that appeared to be mutated above a threshold (corresponding to the background), was deemed as significant. While this simple modeling performed well in earlier studies involving small cohorts (*19*)(*20*), it lacked the required sensitivity when applied to larger cohorts, resulting in many false positives (*17*). This was attributed to the increased mutational heterogeneity introduced by large cohorts that impose increasing difficulties in modeling the mutational background of the disease. For example, a substantial heterogeneity (up to 1,000 fold) exists in the mutational rate across patients with the same cancer type, mostly when there is an ongoing exposure to some carcinogens, such as smoking and UV radiation in lung adenocarcinoma and melanoma patients, respectively (Figure 2).

Additional heterogeneity in mutational rates exists across different genomic loci of the same patient. Two main factors explain to a large extent the genome-wide heterogeneity. The first factor is gene expression levels, which are known to be inversely correlated with mutational rates due to transcription-coupled repair mechanism (*21*). The second factor is the time of replication of the genomic loci during the cell cycle, as it has been shown that late-replicating regions accumulate more mutations than early ones, possibly due to temporal depletion of available nucleotides (*17*). The correlation between these two factors and the mutation rate is shown in Figure 3 for chromosome 14.

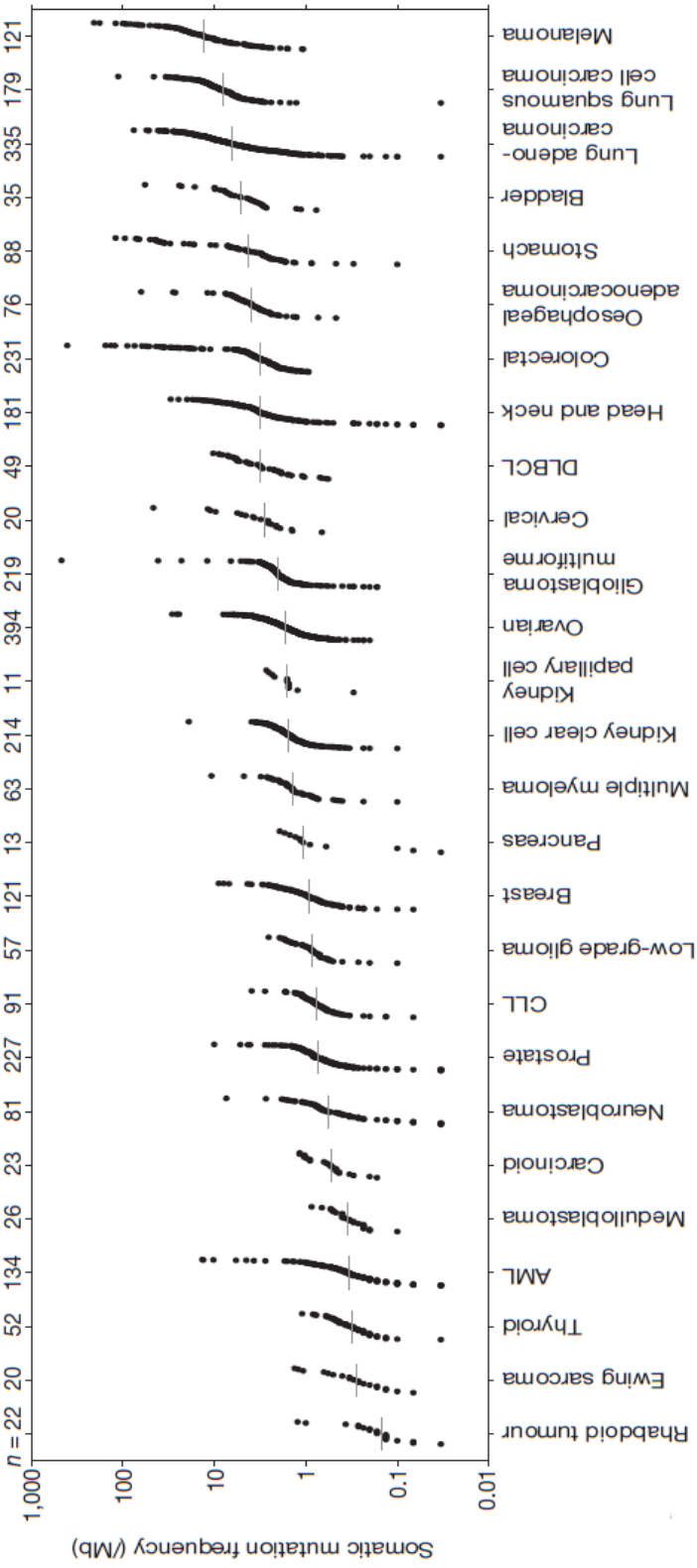


Figure 2 | Differential mutation frequencies among patients. "Each dot corresponds to a tumor–normal pair, with vertical position indicating the total frequency of somatic mutations in the exome" Adapted from (*17*).

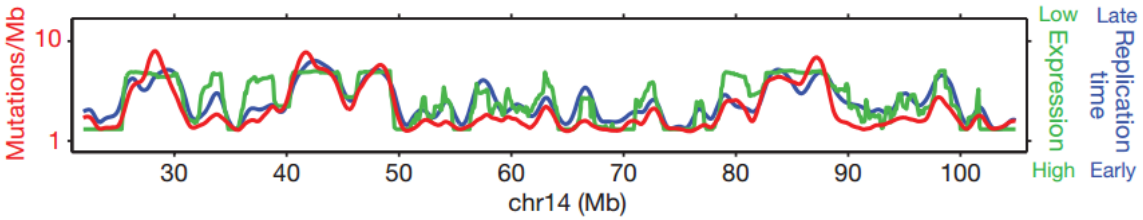


Figure 3 | Genome-wide differential mutations rates. Mutation rate (red), replication time (blue) and expression level (green) across chromosome 14. Adapted from (*17*)

To compensate for cross-patient and genome-wide heterogeneities, later versions of MutSigCV define a more stringent background modeling of the mutational landscape, considering genomic factors on a per-gene and per-patient basis, as opposed to a simple and averaged genome-wide mutational rates. These factors include gene expression levels, replication time, differential chromatin states, local GC content, gene density, and information about mutational rates of nearby genes in cases where the signal is weak (*17*). Similarly, MuSiC models the mutational background taking into account the mutated base, CpG islands properties, transversions and transitions rates (*18*). While employing simpler modeling than MutSigCV, MuSiC allows for a user-defined region of interest and further provides information about implicated mutated pathways, functional impact on the corresponding protein as well as clinical information from online databases such as OMIM (*18*).

As one can appreciate, modeling the mutational background is a complicated task that requires consideration of many factors. Therefore, the above methods are limited by the complexity involved in modeling the mutational background. Failure in generating an accurate mutational background would introduce systematic errors resulting in decreased sensitivity and specificity. Indeed, few factors such as recombination rates, evolutionarily conserved bases, and distance to the telomere, are not considered in current algorithms, although they all shown to be associated with mutation rates (*22*). Additionally, since the

above methods are recurrence-based (capture genes that are mutated above a certain threshold), they are not likely to identify very rarely mutated genes.

Given the limitations above, there is some added value in applying orthogonal methods which do not discriminate driver from passenger mutations based on complex modeling and/or recurrence. Oncodrive-FM is one such method that relies on the notion that there is a positive selection for mutations with high functional impact on the protein. (*23*). MutComFocal is another method which discerns driver from passenger mutations by integrating point mutations and copy number information (*24*). In this work, we have developed another orthogonal method that identifies cancer-associated genes based on global gene expression levels using topological data analysis.

Topology

Topology is a major branch of mathematics that is concerned with qualitative geometric properties of a shape or an object. These qualitative features are those that persist through continuous deformations, such as twisting, shrinking and bending without breaking or tearing the object apart (*25*), as exemplified in Figure 4. In other words, in topology, we care about how subcomponents of an object are interconnected, rather than their coordinates in a metric space. This connectedness property sits at the cornerstone of topology.

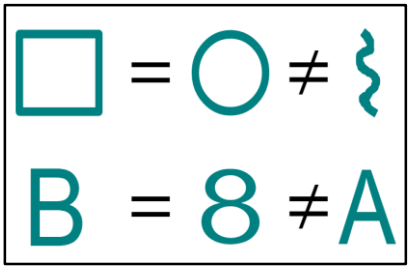
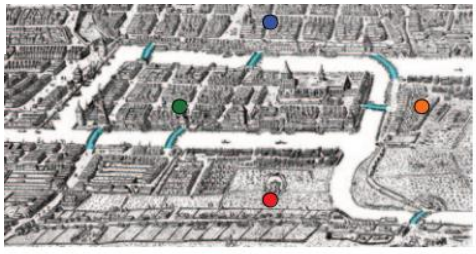


Figure 4 | Examples of deformations.

Leonard Euler laid the foundations of topology back in 1735 when he solved the mathematical problem of the "Seven Bridges of Königsberg"(*26*). In a nutshell, the problem is concerned with finding a walk through the four river-separated segments of the city, by crossing each of its seven bridges only once. On route to a solution, Euler recognized that the only information required to solve the problem is the qualitative features, a.k.a. topological features, of the city landscape, namely, the pattern in which the city segments are interconnected with bridges (Figure 5A). Metric features, such as the distances between segments were not important. By recognizing the importance of the topological features that are invariant under continuous deformations, Euler was able to reduce the problem to its core and find a solution (*27*) (Figure 5B).

A



B

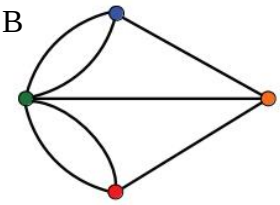


Figure 5 | The city of Königsberg. (A) Seven bridges connect the four parts of the city which are marked in different colors. (B) reduced representation of the city's topological features as offered by Euler. Adapted from (*27*)

Simplifying complex shapes or multidimensional spaces are often required in science. Algebraic topology builds upon Euler's earlier insights about the importance of the topological features, that is, how things are connected, and provides mathematical tools to replace the original object with a much simpler one called *simplicial complex* (*28*). A simplicial complex is a generalization of a network, consisting of nodes (or vertices) and edges with additional higher dimensional elements such as triangles and polyhedrons that accurately capture the topological features of the original space (for an example of a simplicial complex representing an annulus shape see section Mapper background below). In contrast to the original space which is usually defined by infinite continuity of points, the simplicial complex consists of a finite number of elements (nodes, edges, triangles, and higher dimensional elements) which satisfy specific mathematical properties that make it more amenable for algebraic operations (*28*). Consequently, one can perform algebraic operations on the simplicial complex to systematically obtain the topological features of the original shape, otherwise hard to extract. In simple words, a simplicial complex is a stripped down version of the original space that preserves its qualitative topological features, and its major advantage is that it serves as a proxy for operations from which one can extrapolate the original shape's topological features.

Although topology has been an area of intensive study for the last century (*29*), until recently, it has not grown outside of a pure mathematical realm (except a few abstract applications in mathematical physics (*30*)). The underlying reason is that topology, in its traditional formulation, mostly deals with infinite continuous spaces, which are very useful when approaching abstract problems in theoretical physics (e.g. field and string theories (*30*)), but are less helpful in addressing common and more practical problems.

In the past 10-15 years, however, a new branch of topology, termed topological data analysis (TDA) has emerged with strong applications to practical problems involving finite metric spaces, namely point cloud datasets. A point cloud can be imagined as a

sampled set of points from an underlying space or a shape, together with a notion of distance between points, hence defining a finite metric space (*31*). By applying topological notions to the point cloud, we can infer topological features of the underlying space without knowing the space itself. This type of approach resulted in an explosion of applications in many fields outside of pure mathematics, including medicine (*32*–*34*), anatomy (*35*), viral evolution (*36*), materials science (*37*) and image recognition (*38*).

Topological data analysis: a new approach to data analysis

In modern science, we often encounter multidimensional and complex datasets that are very hard to interpret, analyze and visualize. This is especially true in the context of genomics, where datasets are usually comprised of thousands of genes measured from a handful to thousands of samples. Several practical methods have been developed in the context of topological data analysis to analyze and visualize complex datasets. Persistent homology (*39*) and dimensional reduction using Mapper algorithm (*40*) are two popular examples. The latter has been used recently to study complex diseases such as diabetes (*33*) and breast cancer (*32*), and we build upon it in this work.

Non-TDA methods, such as principal component analysis (PCA), or in its more general form, multi-dimensional scaling (MDS), are commonly used algorithms to reduce the dimensionality of complex datasets. At the core of these methods, a projection of the multidimensional data into a lower dimensional space takes place in a way that explains the variance within the data (*41*). However, unlike TDA, these methods impose a projection without prioritizing the local relations (connectedness) of data points in the original multidimensional data (*31*). Hence, adjacent points in the projected space are not necessarily adjacent in the original data. Losing these local relations is a major drawback when handling continuous datasets that take a non-trivial structure with an intrinsic meaning about the data (*31*). As we will see, TDA is a well-suited approach for capturing

the non-trivial structure of the data. In this thesis, we present a novel method that we have developed building upon topological data analysis tools, specifically the Mapper algorithm, to find cancer-associated genes, integrating gene expression levels and somatic mutations.

Mapper background

The Mapper algorithm implements topological ideas based on partial clustering to reduce multi-dimensional data into a low-dimensional simplicial complex that captures some of the local relations within the original dataset, or point cloud (*40*). In the context of this work, the data consists of tumors from TCGA that are represented as points in a multidimensional gene expression space. By emphasizing the local relations between the points, the method can track patterns in the data that are lost when performing standard dimensional reduction techniques and traditional clustering (*40*). An additional feature of Mapper is that it allows analysis of the data through different scales by generating a series of simplicial complexes corresponding to different input parameters known as *gain* and *resolution* (described below). This multiscale analysis allows discovery of different patterns within the data and, therefore, adds robustness to the analysis, as implemented in our method.

Mapper implements a methodology similar to the Vietoris-Rips complex construction (*39*), a topological idea that is illustrated in Figure 6 and implemented as follows: for a given set of points S and a number $\varepsilon > 0$ the Vietoris-Rips complex C is a simplicial complex constructed by following two steps:

- Cover each point with a ball B of diameter ε (Figure 6A), and
- Connect pairs of points whose covering balls intersect with edges, connect three points whose covering balls intersect with a triangle, while points whose covering balls do not intersect at all remain disconnected (Figure 6B).

Repeating these two steps for a series of positive values of ε (Figure 6A, 6C) results in a series of simplicial complexes (Figure 6B, 6D). Intuitively, as the diameter ε increases (Figure 6C), more edges and multidimensional objects such as triangles (yellow) and polyhedrons (green) are added to the simplicial complex, and the latter becomes more connected (Figure 6D). Eventually, at a critical (large) value of ε, all data points are connected to each other. Persistent homology relates topological features of the simplicial complexes at different values of ε.

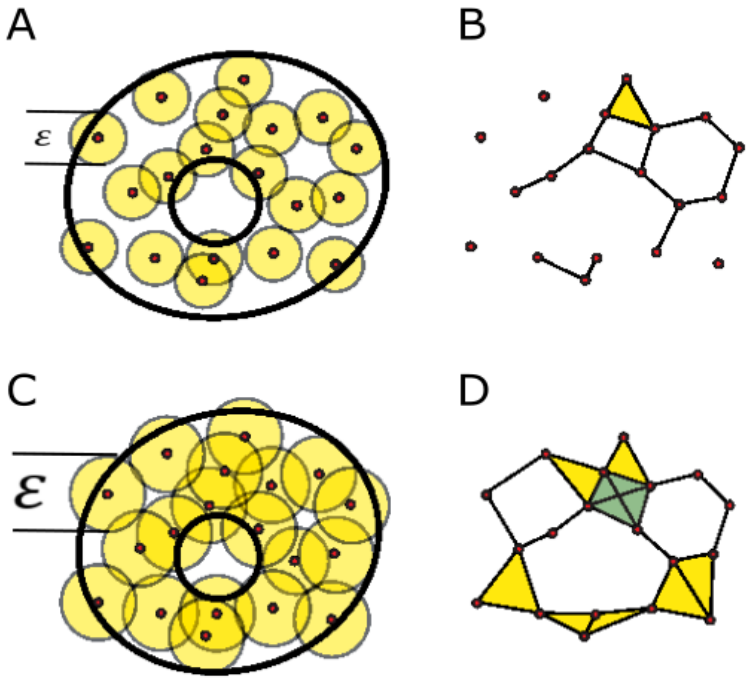


Figure 6 | Vietoris-Rips complex. (A) a point cloud (red dots) representing an annulus is covered with balls of diameter ε. (B) a simplicial complex corresponding to the point cloud and covering balls. Edges, connect pairs of points whose covering balls intersect, triangles (yellow) connect three points whose covering balls intersect. Points, whose covering balls do not intersect at all, remain disconnected in the network. (C) the same point cloud is now covered with balls of a larger diameter ε. The resulting simplicial complex (D) is more connected with more edges, triangles (yellow) and a tetrahedron (green) which connect four points whose covering balls intersect. Adapted from (*39*)

In resemblance to the Vietoris-Rips complex, the Mapper algorithm builds covering regions related to data points and uses their intersection to generate a simplicial complex. However, the Mapper algorithm does not use covering in the form of balls around individual points. Instead, it uses a user-defined filter function and other parameters known as gain and resolution, to define covered regions of the data (*35*).

Mapper algorithm

The input for the Mapper algorithm is a point cloud (namely, a set of points endowed with a metric), one or more filter functions, and two parameters (gain and resolution). One can use different combinations of filter function and metric based on the question and data at hand. Construction of a simplicial complex using Mapper takes place in four steps.

The first step is to map the points in the point cloud to the real line R through the filter function f. One can use several filter functions, among them: Gaussian density, nearest neighbor, etc. For simplicity we demonstrate the Mapper algorithm with a simple Y-coordinate filter function that maps each point in the point cloud to the real line R according to its Y coordinate (Figure 7A).

The second step is to define a covering of the real line R with overlapping intervals whose size (number of points contained) is determined by the resolution parameter. The extent of overlap between intervals is determined by the gain parameter. The resolution parameter takes values ranging from 1 to the total number of data points. With a minimal resolution value, the interval covers the entire real line, whereas with a maximum resolution value, each interval covers individual data points. The gain parameter ranges from 1 to 10 which corresponds to 10 percent to 100 percent overlap between the intervals. Practically, the higher the resolution the more vertices the simplicial complex has, and the higher the gain the more edges it gets. Once a covering of the real line R is set, the inverse

function f^{-1} is used to define a covering of the actual point cloud. These covering patches are called “bins” (Figure 7B). Next, Mapper clusters points within each bin using single-linkage clustering (Figure 7C). The clusters formed in this step correspond to the vertices of the simplicial complex.

Finally, to complete the simplicial complex construction, edges are formed between vertices (clusters) that share at least one data point (Figure 7D). Similarly, if three vertices share a data point, they are connected with triangles and so on. In our analysis, we discard multidimensional elements such as triangles and polyhedrons (see Figure 6D) and consider only one-dimensional elements of the simplicial complex, namely nodes and edges.

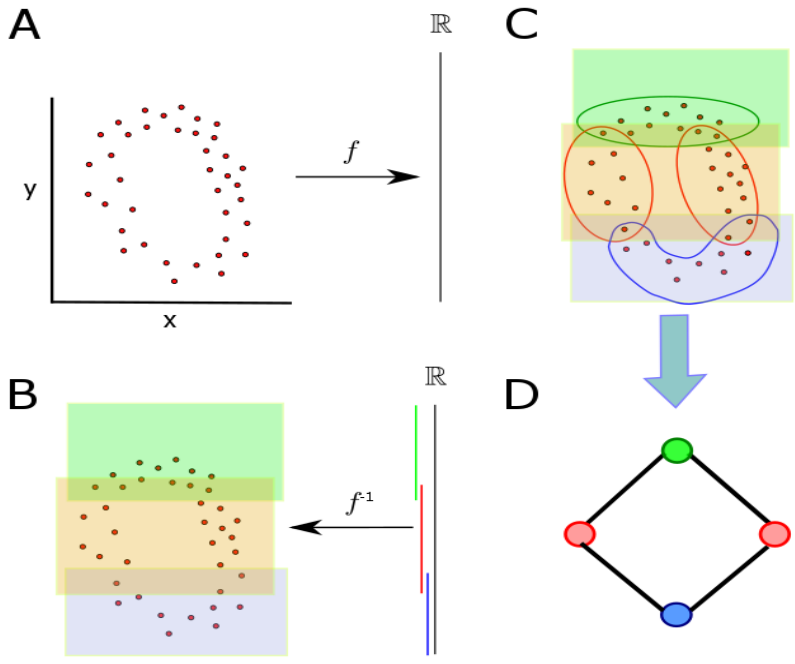


Figure 7 | Simplicial complex construction with Mapper. (A) A point cloud (in this work each point represents a differnet cancer patient) is mapped to the real line R using a filter function. (B) A covering of the real line is defined by gain and resolution parameters and implies a covering of the point cloud. (C) Clustering takes place within each bin. (D) Simplicial complex - clustered points are reduced to nodes, edges connect nodes that share at least one data point. Figure courtesy of Pablo G. Camara.

STRATEGY

Motivation for TDA and global gene expression levels as a phenotype

Global gene expression patterns are a strong and indicative phenotype for tumors. Indeed, many cross-sectional studies have identified oncogenic molecular pathways and tumor subtypes based on gene expression patterns extracted from RNA-seq data (*9*, *10*, *42–46*). Notwithstanding, these studies, based on traditional clustering methods, might be incomplete since the transcriptional pattern across large cohorts usually takes a continuous structure rather than a discrete one. In other words, there exists a continuum of samples between each reported subtype that is lost when applying standard dimensional reduction techniques and traditional clustering methods. Topological data analysis, on the other hand, is a viable alternative approach, since it achieves dimensional reduction without losing local relations within the data that can be used in the downstream analysis (Methods). For instance, the simplicial complexes generated by the Mapper algorithm from gene expression levels of 142 glioblastoma multiforme (GBM) patients represent a continuum of samples between classical and proneural GBM subtypes, characterized by elevated expression levels of *EGFR* and *PDGFRA*, respectively (*8*) (Figure 8).

Calling cancer-associated mutations from gene expression levels is an orthogonal approach to standard methods such as MutSigCV and MuSiC. These methods rely on modeling the background mutation rate to discern driver from passenger mutations. Our method offers the advantage to pick up novel genes, undetectable by those recurrences-based methods, such as rarely mutated genes or very long genes, as long as they have an effect on global gene expression patterns. Since our method is non-parametric, it is not vulnerable to systematic errors involved in complex modeling of a background mutation rate.

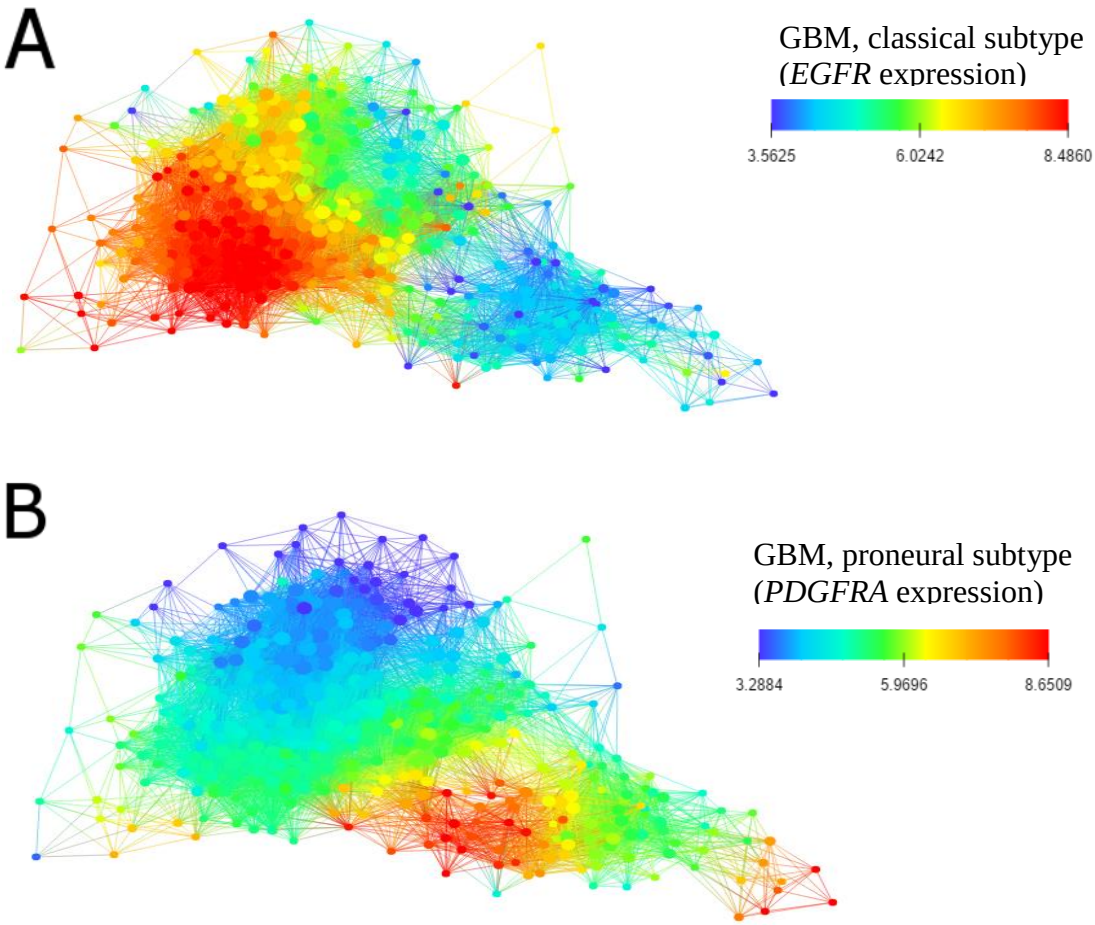


Figure 8 | GBM classical and proneural networks. The simplicial complex was generated by Mapper from the global expression levels of 142 GBM patients. Each node consists of several samples clustered together based on similar gene expression levels. Edges connect nodes that share at least one sample. Nodes are colored by the average expression levels of (A) *EGFR* and (B) *PDGFRA,* corresponding to markers of classical and proneural GBM subtypes, respectively (*8*). There is no sharp separation between patients based on expression levels of *EGFR* and *PDGFRA*, but rather a continuum of samples between the two subtypes.

Statistical Analysis - Connectivity

In this work, we have developed and applied a statistical method (*connectivity analysis*) in conjunction with topological representations of the data (simplicial complex representations generated using the Mapper algorithm) to identify cancer-associated genes. We nominate a mutation as *cancer-associated* if it is statistically associated with global gene expression commonalities between tumors carrying that mutation. Our statistical analysis assesses this association by exploiting the local relations within the simplicial complex representation of the data. More specifically, our analysis calculates, for every mutated gene, a "connectivity value" which reflects how much the mutation is connected or localized across the simplicial complex compared to random chance (Methods). To understand the rationale behind this approach, it is useful to consider a simplicial complex as a network that captures the similarities between transcriptional programs of different patients. Intuitively, features (such as specific mutations) localized or connected in a non-random fashion across the network are associated with the expression commonalities of a group of samples. The connectivity analysis calculates the statistical significance of the feature's localization within the simplicial complex, thereby identifying features that are associated with the structural layout of the simplicial complex. In our case, the features we are testing for are mutated genes and the simplicial complex represents the local relations between gene expression levels of the tumors in the cohort. Appropriately, we nominate candidate cancer-associated genes if they are deemed statistically significant by this connectivity analysis algorithm. For example, a statistically significant localization of *CIC* and *PTEN* mutations over a simplicial complex generated from global gene expression levels of 513 lower-grade glioma patients are shown in Supplemental Figure 1.

RESULTS

We analyzed in total seven tumor types, spanning 2,916 patients from TCGA, identifying 85 cancer-associated genes, including 38 genes (Table 2) that were not previously reported in the literature of the corresponding cancer type. We retrieved global RNA-seq gene expression levels and somatic mutation data for all tumors using TCGA (*5*) and Broad's firehose pipeline (*47*), respectively (see Methods section for somatic mutation criteria). We reduced the multidimensional global gene expression space into a two-dimensional network without losing its continuous structure using the Mapper algorithm. The nodes in the network comprise samples with similar gene expression patterns. Edges connect nodes that share at least one data point (c.f. example in Figure 7). For robustness, and to maximize our statistical power, we generated a series of topological representations for each tumor by scanning systematically over the parameter space of the Mapper algorithm (gain and resolution). We reasoned that mutations that are localized in the resulting networks are associated with the global gene expression levels in the patients that harbor those mutations, and, therefore, with the disease. We quantified statistically the localization of all non-synonymous mutations that passed initial filtering thresholds, based on mutation frequency and the ratio of non-synonymous versus synonymous mutations (Methods), using our connectivity analysis.

To test for artifacts coming from the observed inverse correlation between expression and neutral somatic mutations (*21*), and potential deviations from it due to positive selection, we also computed the Jensen-Shannon distance between the distribution of mutations and expression in the network, for all mutated genes with a statistically significant connectivity in the network. Additionally, as a further check on consistency, we tested for batch effects and for the association between mutational load and global expression patterns, subsampling mutations and/or removing outliers (samples with less than 10 mutations) when required. A flow diagram of our pipeline is provided in

Supplemental Figure 2**,** and the parameters used for each tumor are provided in Supplemental Table 2.

Aggregating results from the seven tumor types that we have analyzed (bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), glioblastoma (GBM), low-grade glioma (LGG), colon (COAD), lung (LUAD), and stomach adenocarcinoma (STAD)), we have identified 85 distinct (in at least one tumor) cancer-associated genes. Among those genes, 38 genes are reported here for the first time as being associated with the corresponding cancer type (23 of them in lung adenocarcinoma). Another 39 genes were previously reported in the literature (*9*, *10*, *42–46*) and 8 more (one in stomach and the rest in colon adenocarcinoma) are reported by MutSig2CV analysis by Broad institute's Firehose pipeline (*47*), in the same cohort of patients (Table 2). A complete list of the results, organized by cancer type, including one pseudogene and 12 genes that were considered as artifacts due to low expression levels (TPM<2), batch effects, or because they declared as false discoveries by our positive selection test are provided in Supplemental Table 1.

Table 2 | Cancer-Associated Genes Identified using Topological Data Analysis
In parentheses the number of cancer types associated to the mutated gene. A large fraction of the detected genes (underlined) were associated with lung adenocarcinoma.

Novel candidates			Literature			MutSig2CV
FMN2 (2)	EPHB1	PLXNA4	TP53 (6)	ATM	NOTCH1	AKAP13
ADAMTS12	ESRRA	POLQ	EGFR (3)	CBFB	PIK3R1	ARHGAP5
AFF2	FAT3	PTPRC	ATRX (2)	CIC	PTEN	ARFGEF1
ANK2	GPR158	PTPRD	CDH1 (2)	CTCF	RB1	VPS13B
ANO4	HSPG2	RP1L1	COL6A3 (2)	ELF3	RNF43	FLT3
CACNA2D1	HUWE1	SATB2	IDH1 (2)	FGFR3	RUNX1	NEFH
CALN1	LRP2	SCN2A	KRAS (2)	FUBP1	SMAD4	CCDC141
CCDC129	MED13	SLC8A1	NF1 (2)	GATA3	SOX9	STK11
CDH12	MYOM2	SLITRK4	PIK3CA (2)	IDH2	STK11	
CHD5	NCOR1	SYNE1	SMARCA4 (2)	KEAP1	TCF7L2	
CPED1	NOTCH2	TSSC2	AKAP9	KMT2C	TCF12	
CROCCP2	PCDHB4	UNC13C	APC	MAP2K4	NIPBL	
DGKB	PEG3		ARID1A	MAP3K1	ZBTB20	

Finding previously reported genes set confidence in our novel method, especially considering hypermutated tumors such as colon adenocarcinoma (*48*) which confer a

significant number of passenger mutations. Canonical cancer-associated genes (such as *TP53*, *APC*, *PIK3CA*, *KRAS*) were deemed significant (JSD q-value<0.15) in our positive selection test (see Methods) across all tumors (Figures 9C,11-16C) and none of the previously reported genes were declared as a false positive (JSD q-value >0.85). Our method also proved capable of identifying 18 rare mutations occurring in less than 5 percent of the patients. In what follows, we discuss our results in the context of each cancer type.

Lung adenocarcinoma

Despite the considerable heterogeneity in mutation rates across lung adenocarcinoma patients (Figure 2), the mutational load distribution is uniform (Figure 9A top), and it is not associated with mutational load effects, as assessed by our connectivity analysis across a series of networks generated using the Mapper algorithm (Figure 9A, bottom). After running our connectivity analysis for a filtered list of 350 genes over a coarse range of networks (see Methods), we identified an optimal range of networks (Figure 9B), which is enriched for significant genes, does not overlap with mutational load effects, and contains a large number of samples. We used this range of networks for a secondary connectivity analysis with finer resolution and gain parameters (Supplemental Figure 3A) and identified in total 35 cancer-associated genes in lung adenocarcinoma (Figure 9C, bottom), of which 23 are novel candidates that account for almost half of all the novel candidates we have found across the seven tumors collectively (Table 2). Positively, all of the previously reported genes (*STK11, EGFR, KEAP1, KRAS, ATM, TP53, SMARCA4* (*9*) and *AKAP9* (*49*)) have a significant JSD q-value (q<0.15, Figure 9C, middle), which implies a signal for positive selection (Methods). This raises a particular interest in the 12 novel candidates that also present a signal for positive selection and removes interest in two genes, *KLHL4,* and *PSG8*, which demonstrate non-significant JSD q-value (q>0.85), and are therefore possibly artifacts.

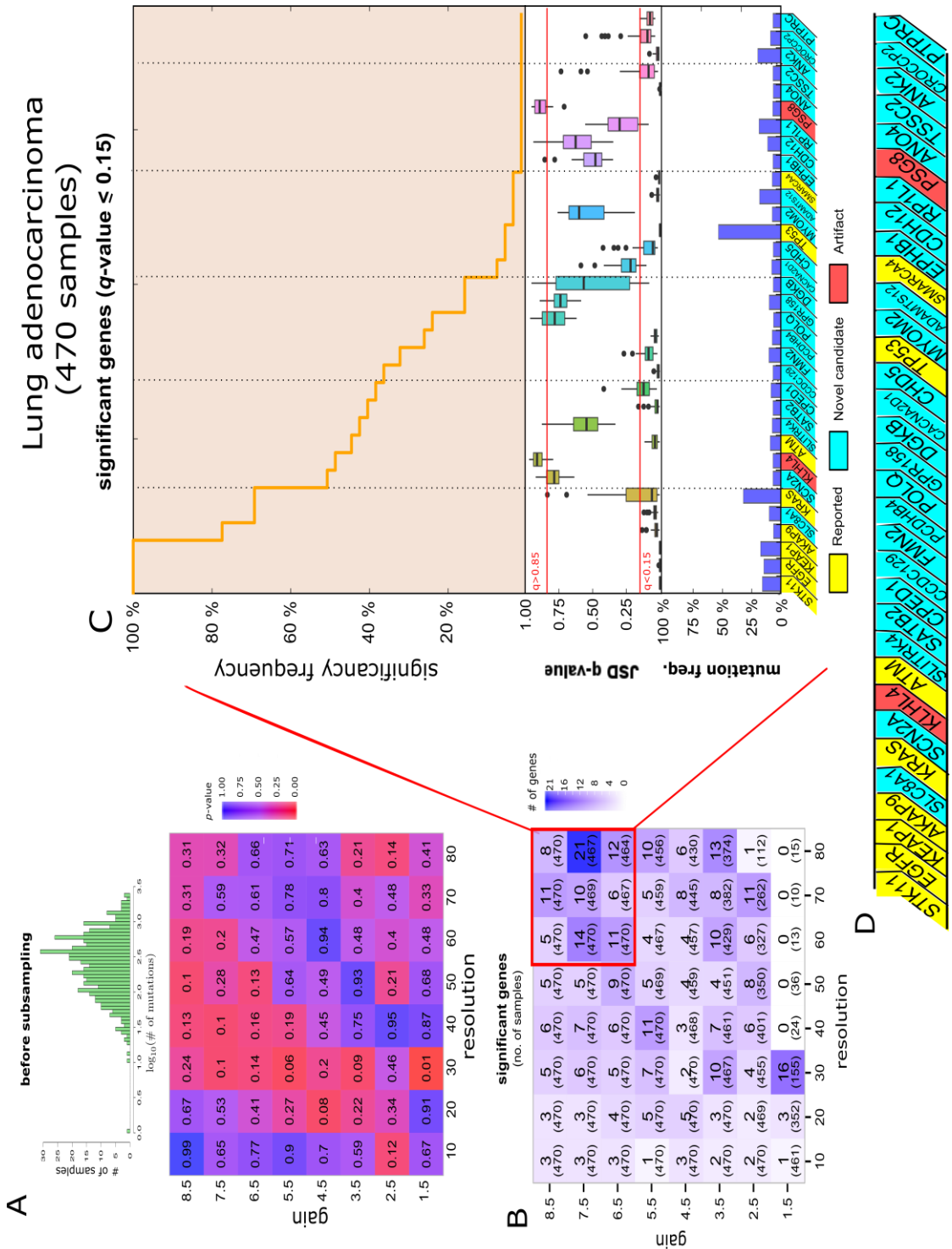

Lung adenocarcinoma
(470 samples)
A
before subsampling
# of samples
log10(# of mutations)
p-value
gain
resolution
B
significant genes
(no. of samples)
# of genes
C
significant genes (q-value ≤ 0.15)
significancy frequency
JSD q-value
mutation freq.
q>0.85
q<0.15
Reported
Novel candidate
Artifact
D
STK11
EGFR
KEAP1
NFE2L2
SOX17
KRAS
SCN2A
KLHL4
ATM
SLITRK4
SATB2
CPED1
CDC42
FMN2
PCDHB4
POLQ
GPR158
DGKB
CACNA2D1
CHD5
TP53
MYOM2
ADAMTS12
SMARCA4
EPHB1
CDH12
RPL4
PSG8
ANO4
TSSC4
ANK2
CROCCP2
PTPRC

Figure 9 | Identification of Lung Adenocarcinoma Associated Mutations using TDA. (A) Top: mutational load distribution in a logarithmic scale across all samples. Bottom: statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan, artifacts in red. Genes are ordered by significance frequency across the networks, as presented in the top step function. Boxplots represent JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort. (D) Enlarged list of identified genes.

An especially interesting novel candidate is *FMN2*, that encodes a member of the Formin homology proteins family, Formin-2. *FMN2* is mutated in 10% of the samples (Figure 9C, histogram), and 30% of the networks (Figure 9C, top) and is suspected to act as a tumor suppressor in the p53/Rb pathway via positive regulation of *CDKN1A*, that encodes the Cdk inhibitor, p21 (*50*). We, therefore, reasoned that *FMN2* mutations, if associated with lung adenocarcinoma, should be mutually exclusive with other mutant tumor suppressors or mutually exclusive with amplified oncogenes in the p53/Rb pathway (This is especially true if CDKN1A is the only critical target of FMN2). Indeed, we found *FMN2* mutations to be mutually exclusive with *TP53* mutations both in lung adenocarcinoma $(p < 5.3x10^{-6})$ and bladder urothelial carcinoma $(p < 0.022)$, using Fisher's exact test. Additionally, in lung adenocarcinoma, *FMN2* mutations are mutually exclusive with copy number amplification of *CCNE1* ($p < 0.055$) and *E2F3* ($p < 0.054$), two oncogenes in the pathway that pushes cell cycle forward (Figure 10). These observations reinforce a potential role for *FMN2* in the p53/Rb pathway as a tumor suppressor and encourage further investigation into *FMN2* implications in lung and bladder cancer.

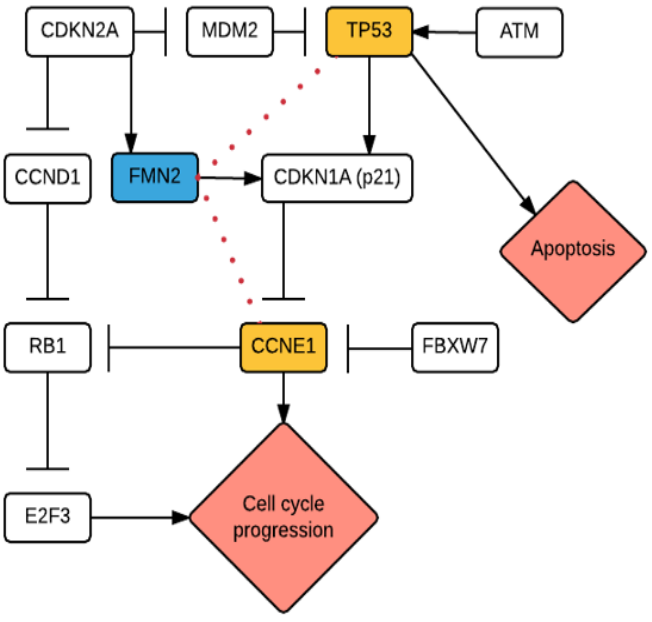


Figure 10 | *FMN2* and p53/Rb pathway. *FMN2* is induced by *CDKN2A* and positively regulates *CDKN1A* by preventing its degradation (*10*, *50*). *FMN2* mutations are mutually exclusive with *TP53* and *CCNE1* (red points) in lung adenocarcinoma.

Additionally, gene ontology enrichment analysis using Enrichr (*51*) indicated a group of six genes involved in mechanism and structure of ion channels. These include the previously reported *AKAP9* (*49*), as well as *ANO4*, *SLC8A1*, *ANK2* and *SCN2A* and *CACNA2D* (Figure 9C bottom). Aberrant ion channels activity is known to participate in various oncogenic processes (*52*) and specifically in lung adenocarcinoma (*53*). While these results are a subject for further validation, they might guide future research about ion channels and their involvement in cancer and their potential as biomarkers for prognosis and diagnosis, a direction already under exploration (*53*). Other interesting novel candidates are *ADAMTS12 and PCDHB4* which are involved in cell-cell adhesion processes (*54*, *55*), an important step towards cellular invasion and metastases (*4*). Finally, we report *PTPRC,* a member of the protein tyrosine phosphatase (PTP) family, which was previously associated with stomach adenocarcinoma (*45*) as a novel candidate in our lung adenocarcinoma analysis. Interestingly, *PTPRD*, another family member of the PTP family, that is known to play various roles in cancer (*56*), is reported by us as a novel candidate in urothelial bladder carcinoma as detailed below.

Urothelial bladder carcinoma

Similar to the analysis of lung adenocarcinoma, we did not observe mutational load effects involved in the bladder cancer analysis (Figure 11A). Our analysis over a coarse grid, followed by a fine grid of networks (Figure 11B, Supplemental Figure 3B) identified in total 9 cancer-associated genes in the urothelial bladder carcinoma cohort, four of them were previously reported (*FGFR3*, *RB1*, *ELF3* and *TP53* (*10*)), another four genes (*PTPRD*, *MED13*, *FMN2*, *HSPG2*) are novel candidates and one of them (*MUC17*) has high JSD q-value and is a probable false positive (Figure 11C).

Besides *FMN2* (mutated in 6.3% of the samples) which has been already discussed above, *PTPRD* (mutated in 6.9% of the samples) is another interesting candidate. *PTPRD* has a near significant JSD q-value, and it is found to be statistically significant in 60% of the networks, therefore being robustly associated with global expression patterns in the cohort. *PTPRD* gene encodes a member of the Protein Tyrosine Phosphatase family, which are reported to play a role in cancer either as tumor suppressors or oncogenes (*56*). Specifically, it was reported to have a tumor suppressor role in glioblastoma through dephosphorylation of STAT3, an oncogenic transcription factor which is also implicated in bladder cancer development (*57*). Considering the above, and the possible involvement of other PTP family members (*PTPRC*) in lung adenocarcinoma (see previous section), *PTPRD* mutations and *PTPRD-STAT3* interactions are an interesting line of further investigation.

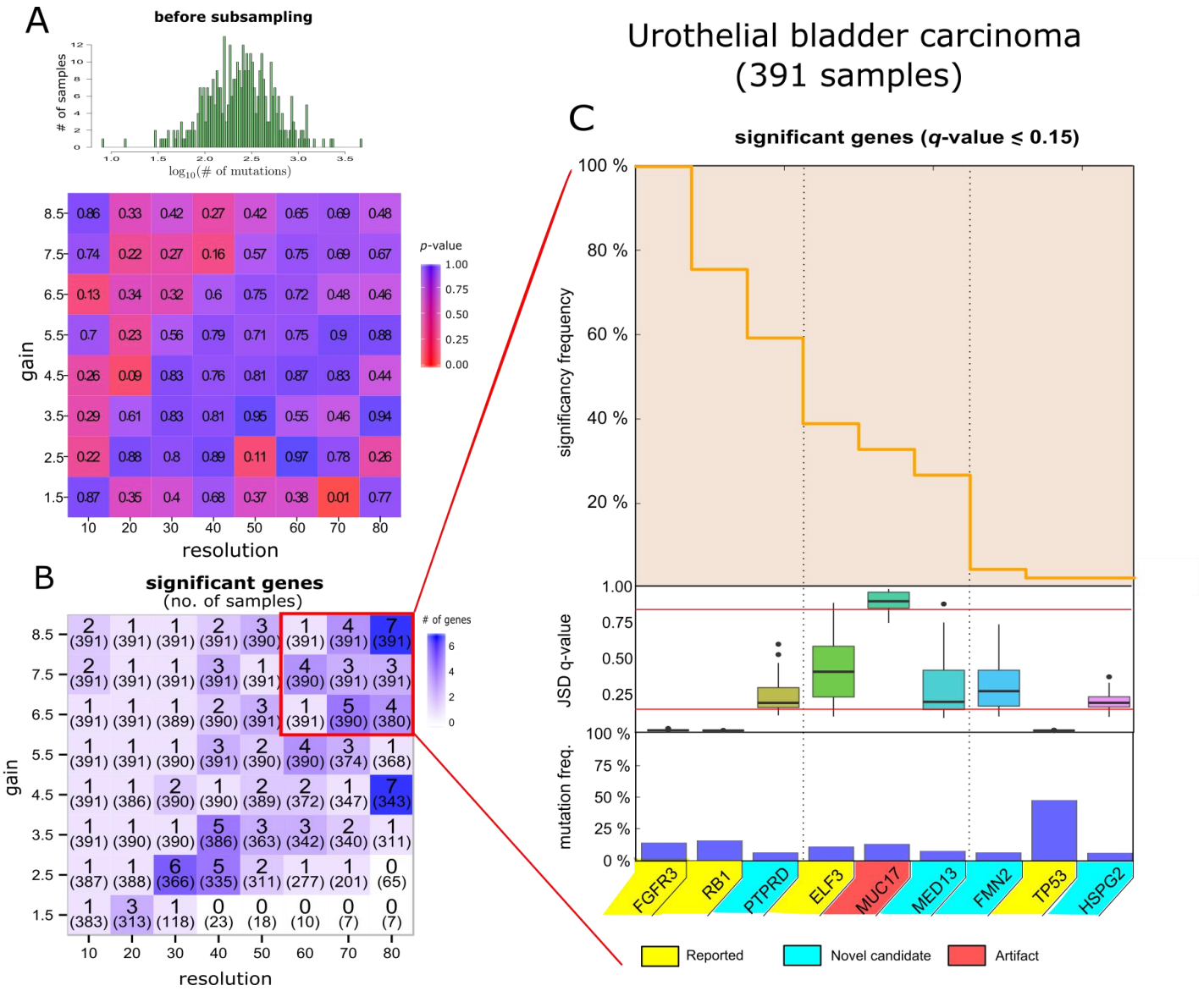

A
before subsampling
# of samples
log10(# of mutations)
p-value
gain
resolution
B
significant genes
(no. of samples)
# of genes
gain
resolution
Urothelial bladder carcinoma
(391 samples)
C
significant genes (q-value ≤ 0.15)
significancy frequency
JSD q-value
mutation freq.
FGFR3
RB1
PTPRD
ELF3
MUC17
MED13
FMN2
TP53
HSPG2
Reported
Novel candidate
Artifact

Figure 11 | Identification of Urothelial Bladder Cancer Associated Mutations using TDA. (A) Top: mutational load distribution in a logarithmic scale across all samples. Bottom: no statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan, artifacts in red. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represent JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort.

Lower-grade glioma and glioblastoma multiforme

The low-grade glioma cohort included two hypermutated samples which resulted in significant mutational load effects in some of the generated networks. Subsequently, in order to remove the mutational load effects, we removed mutations from hypermutated cases in a subsampling process (see Methods section for more information) (Figure 12A). In the lower-grade glioma cohort, connectivity analysis over a selected range of networks (Figure 12B, Supplemental Figure 3C) revealed highly enriched results for previously reported genes (Figure 12C) (*43*). The only novel candidate is *SYNE1* (mutated in 2.3% in the samples), a very long gene encoding for 8,797 amino acids protein. Identifying *SYNE1* as a novel candidate emphasizes the power of our method to detect long and rarely mutated genes which are usually undetected by standard recurrence based techniques, such as MutSig2CV. Since our method does not rely on any prior modeling of this kind, but rather on association with an indicative phenotype (gene expression levels), we are able to identify it as cancer associated, despite its length and low frequency. Other interesting candidates in lower-grade glioma are NIPBL (mutated in 3.7% of the samples) and *COL6A3* (mutated in 2.1% of the samples), which has just been recently reported as cancer-associated using MutSig2CV analysis on a joint low-grade glioma and glioblastoma cohort of 1,122 patients (*58*). NIPBL1 has a role in chromatin cohesion, a potential oncogenic process if malfunctioned (*58*). Our results further support its implication in the disease. *COL6A3* was found to be significant both in low-grade glioma and glioblastoma (mutated in 6.3% of the samples). In the GBM cohort, on top of *COL6A3,* we have identified five cancer-associated genes, four of them (*IDH, ATRX, EGFR,* and *NF1*) are previously reported (*8*), and one of them, *CALN1* (mutated in 4.9 % of the samples), is a novel candidate with a significant JSD q-value, a sign for positive selection. *CALN1* encodes a calcium ion binding protein. Glioblastoma multiforme results are summarized in Figure 13 and Supplemental Figure 3G.

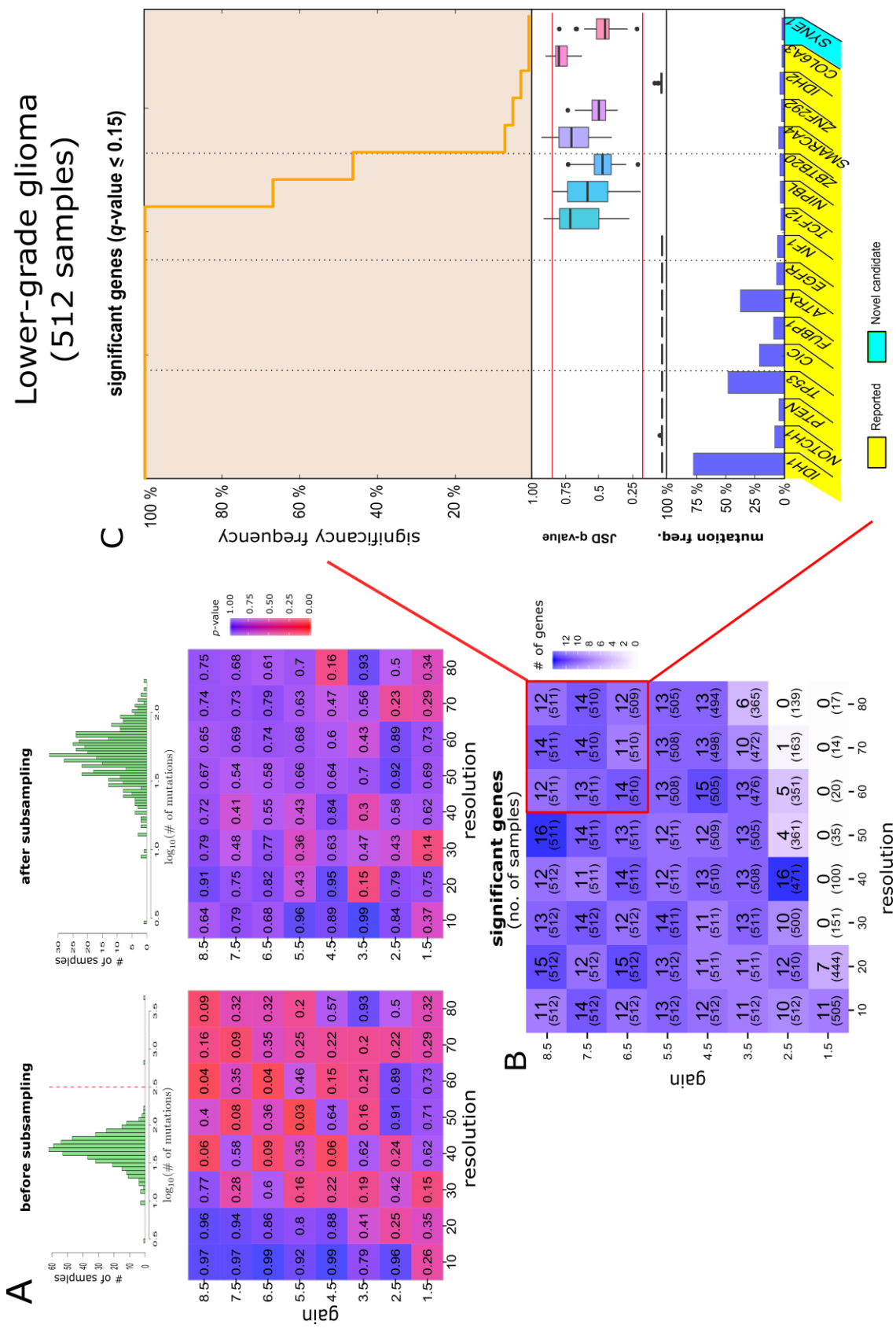

Lower-grade glioma
(512 samples)
A
before subsampling
after subsampling
# of samples
log10(# of mutations)
gain
resolution
p-value
B
significant genes
(no. of samples)
# of genes
gain
resolution
C
significant genes (q-value ≤ 0.15)
significance frequency
JSD q-value
mutation freq.
IDH1
NOTCH1
PTEN
TP53
CIC
FUBP1
ATRX
EGFR
NF1
TCF12
NIPBL
ZBTB20
SMARCA4
ZNF292
IDH2
COL6A3
SYNE1
Reported
Novel candidate

Figure 12 | Identification of Lower-grade Glioma Associated Mutations using TDA. (A) Top left: mutational load distribution in a logarithmic scale across all samples. Before subsampling (Methods), there are two hypermutated cases. The dashed red line represents a subsampling threshold. Grid below the distributions summarizes the statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). In most networks the mutational load is significant ($p<0.05$). After subsampling (Methods) as per the threshold, the new mutational load distribution is centered around the mutational load median of the non-hypermutated cases. The mutational load is no longer significant in most networks. (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represent JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort. (D) Enlarged list of identified genes.

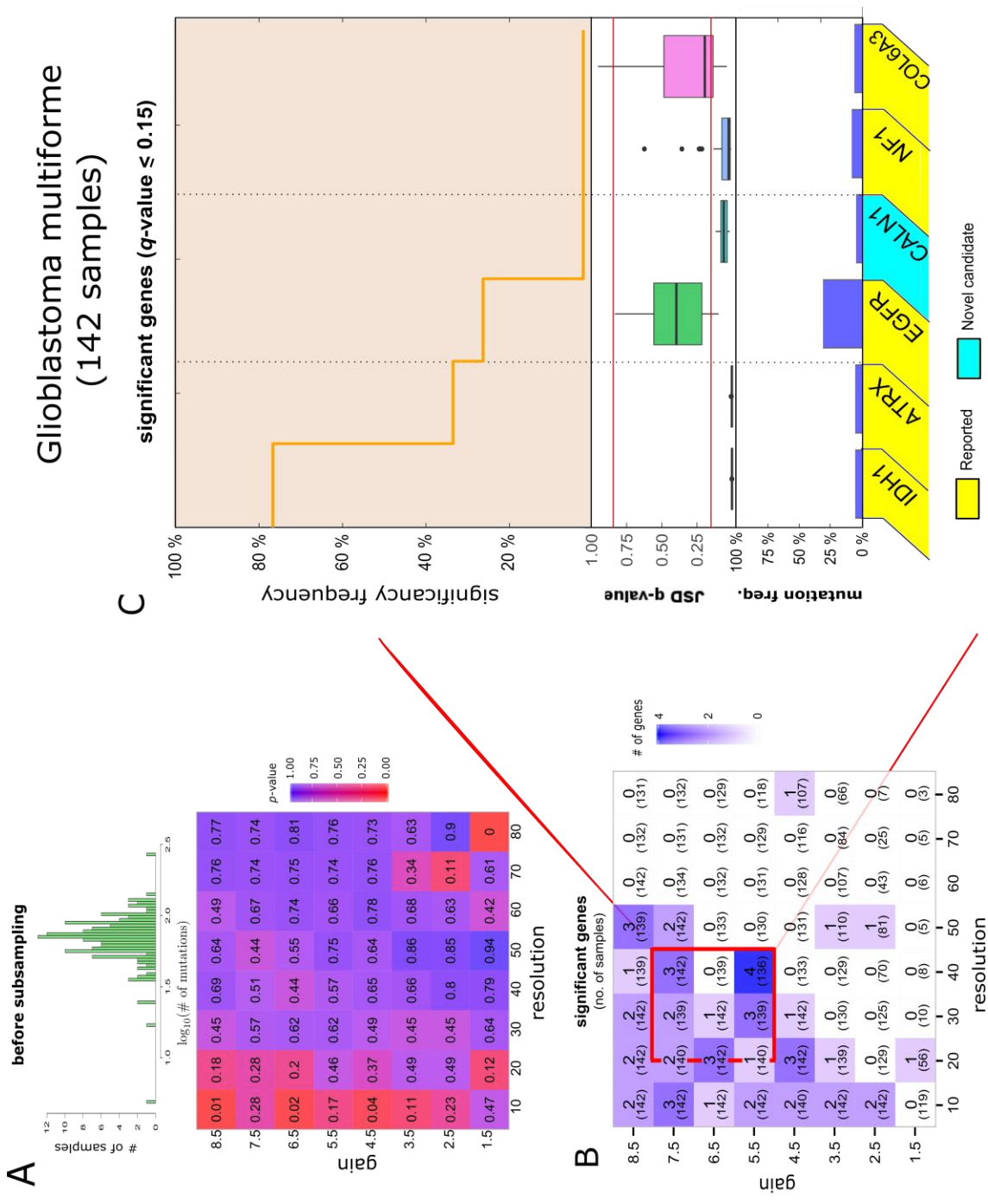

A
before subsampling
# of samples
log10(# of mutations)
gain
resolution
p-value
B
significant genes
(no. of samples)
# of genes
gain
resolution
C
Glioblastoma multiforme
(142 samples)
significant genes (q-value ≤ 0.15)
significancy frequency
JSD q-value
mutation freq.
IDH1
ATRX
EGFR
CALN1
NF1
COL6A3
Reported
Novel candidate

Figure 13 | Identification of Glioblastoma Multiforme Associated Mutations using TDA. (A) Top: mutational load distribution in a logarithmic scale across all samples. Bottom: no statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes. To avoid inclusion of networks with mutational load effects (A), the number of tiles selected is smaller than in other cases (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represents JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort.

Invasive breast carcinoma

The breast cancer cohort is the largest one we analyzed, encompassing 930 samples after removal of a few outliers (Figure 14A) that contained less than 10 mutations. Mutational load effects were not completely removed, resulting in a narrow range of networks that were used for the connectivity analysis (Figure 14B, Supplemental Figure 3F). Nevertheless, we identified 9 previously reported genes (*44*) and 4 novel candidates (Figure 14C). Interestingly, 8 out of the 13 cancer associated genes are found mutated in less than 4 percent of the samples, emphasizing the method's ability to identify rarely mutated genes.

One of the novel candidates, *HUWE1* (mutated in 2.5% of the samples), is a core player in various oncogenic pathways, including ubiquitination of *TP53* and *BRCA1* (*59*), two important tumor suppressors prominent in breast cancer (*44*). While elevated expression levels of *HUWE1* have been reported associated with breast cancer development (*59*), it has never been reported associated in its mutant form. Similarly, aberrant expression levels of another rarely mutated gene, *NOTCH2* (mutated in 2% of the samples), showed association with breast cancer progression (*60*), however not in its mutant form. Given the fact that *NOTCH2* showed a significant signal for positive selection (JSD q-value<0.15) and *HUWE1* has not be declared as false discovery, we find them to be interesting for further research, especially since there is a growing evidence supporting the Notch pathway and *HUWE1* as therapeutic targets (*61*)(*62*). *LRP2*, which encodes, LDL receptor protein-2, is another novel candidate. Interestingly, *LRP1B*, a paralog of *LRP2* (according to Ensembl ver.84 (*63*)), is reported to be associated with endocervical carcinoma (*64*). While a deeper investigation is needed in order to identify *LRP2's* role in breast cancer, the fact that other LDL receptor, *LDLR*, is therapeutically targeted (*65*), makes *LRP2* an interesting candidate for further investigation.

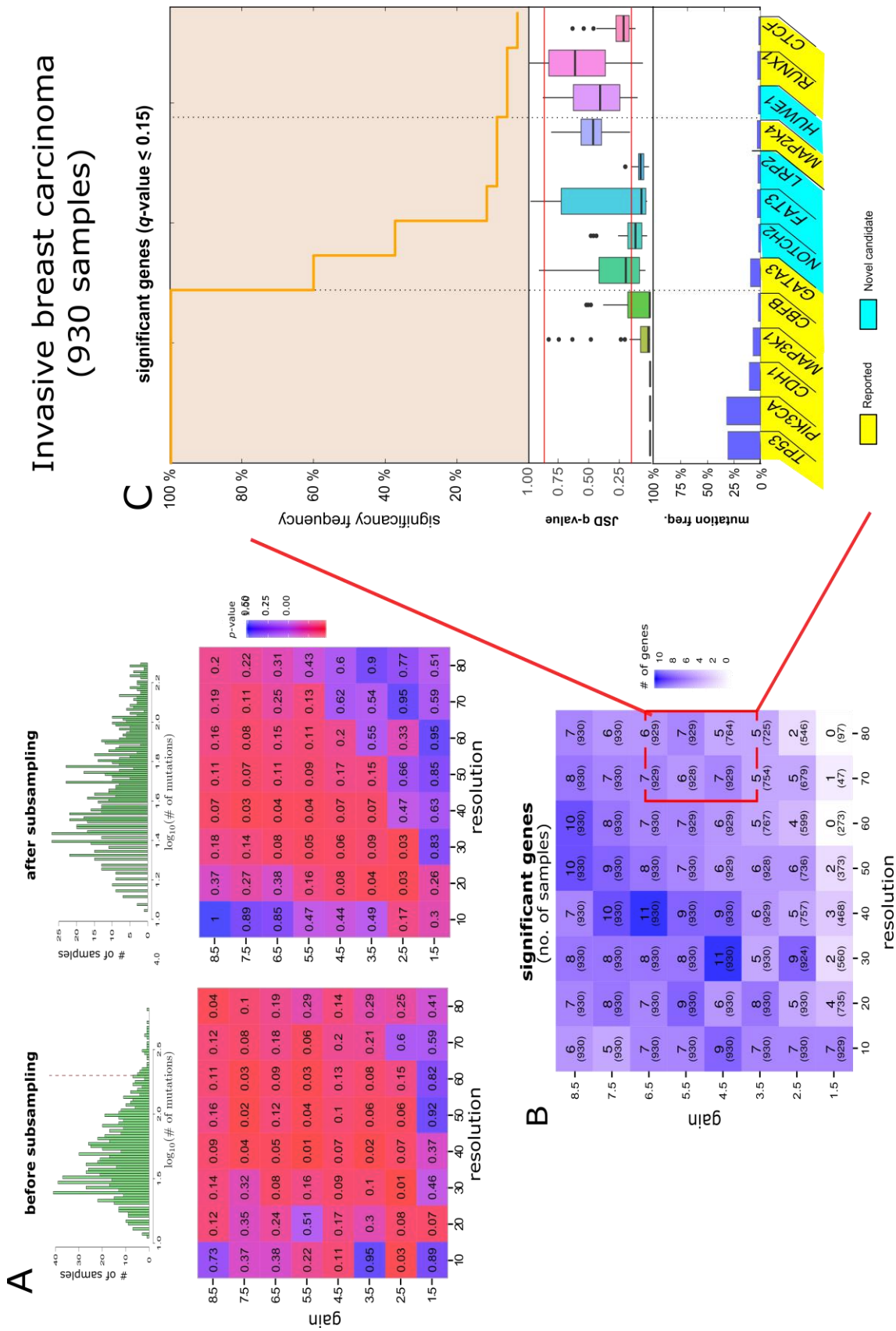

Invasive breast carcinoma
(930 samples)
A
before subsampling
after subsampling
# of samples
log10(# of mutations)
gain
resolution
p-value
B
significant genes
(no. of samples)
# of genes
C
significant genes (q-value ≤ 0.15)
significance frequency
JSD q-value
mutation freq.
TP53
PIK3CA
CDH1
MAP3K1
CBFB
GATA3
NOTCH2
FAT3
LRP2
MAP2K4
HUWE1
RUNX1
CTCF
Reported
Novel candidate

Figure 14 | Identification of Invasive Breast Carcinoma Associated Mutations using TDA. (A) Top left: mutational load distribution in a logarithmic scale across all samples. The dashed red line represents a subsampling threshold. Grid below the distributions summarizes the statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). In most networks the mutational load is significant ($p<0.05$). After subsampling (methods) as per the threshold, the new mutational load distribution is centered around the mutational load median of the non-hypermutated cases. The mutational load is no longer significant in most networks. (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan, artifacts in red. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represents JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort.

Colon adenocarcinoma

Identifying cancer-associated genes in colon adenocarcinoma is a difficult task owing to the bi-modal distribution of the mutational load, separating the cohort (208 patients), into two groups, corresponding to hypermutated and non-hypermutated samples (Figure 15A). Those hypermutated samples have significant mutational load effects (which increases our false discovery rates), as assessed by our connectivity analysis across a series of networks, generated by the Mapper. Subsampling and outlier removal eliminated artifacts in the colon adenocarcinoma dataset (Figure 15A bottom).

We identified in total 20 cancer-associated genes in colon adenocarcinoma, including 10 previously reported in the literature (*42*), and seven that were also reproduced using MutSig2CV (Figure 15C). Of the previously reported genes, particularly interesting is *SOX9* (mutated in 12% of the samples), a developmental gene encoding the SOX-9 protein, an important factor for cell differentiation in intestinal stem cell niche (*42*). *SOX9* is also known to facilitates beta-catenin degradation (*42*), an oncogenic factor in the Wnt signaling pathway (*66*). Loss-of-function mutations in *SOX9*, therefore, could have oncogenic effects. Since *SOX9* mutations have not yet validated experimentally to be a driver of colon adenocarcinoma, to our best knowledge, our finding can facilitate an exploration in this direction.

Two additional genes, *NCOR1,* and *ESRRA* were found to be novel candidates. *ESRRA (*mutated in 11% of the samples), is particularly interesting since it encodes the Estrogen-Related Receptor alpha (ERRα) protein, thus suggests that there may be a subset of steroid hormone responsive tumors. This observation is important since ERRα is a potential biomarker of unfavorable prognosis and is targeted for therapeutic development in breast cancer (*67*).

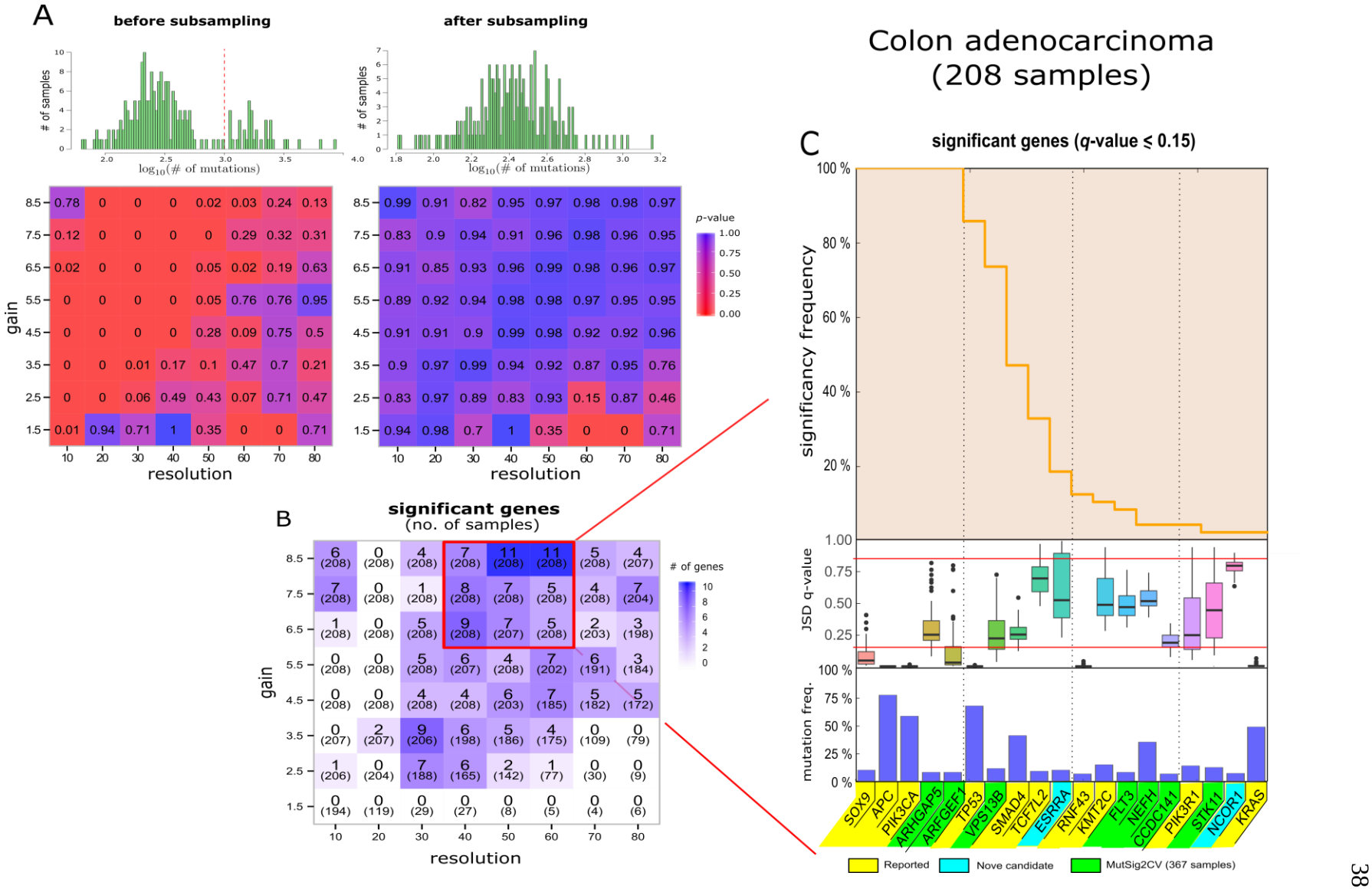

A
before subsampling
after subsampling
# of samples
log10(# of mutations)
gain
resolution
p-value
1.00
0.75
0.50
0.25
0.00
B
significant genes
(no. of samples)
# of genes
C
Colon adenocarcinoma
(208 samples)
significant genes (q-value ≤ 0.15)
significancy frequency
JSD q-value
mutation freq.
SOX9
APC
PIK3CA
ARHGAP5
ARFGEF1
TP53
VPS13B
SMAD4
TCF7L2
ESRRA
RNF43
KMT2C
FLT3
NEFH
CCDC141
PIK3R1
STK11
NCOR1
KRAS
Reported
Nove candidate
MutSig2CV (367 samples)

Figure 15 | Identification of Colon Adenocarcinoma Associated Mutations using TDA. (A) Top left: mutational load distribution in a logarithmic scale across all samples. Before subsampling (Methods) there is an observed bi-modal distribution, separating patients into hypermutated and non-hypermutated cases. The dashed red line represents a subsampling threshold. Grid below the distributions summarizes the statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). In most networks the mutational load is significant ($p<0.05$). After subsampling (Methods) as per the threshold, the new mutational load distribution is centered around the mutational load median of the non-hypermutated case. The mutational load is no longer significant in most networks. (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan, reproduced by MutSig2CV in green. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represent JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort.

Stomach adenocarcinoma

The analysis of the stomach adenocarcinoma cohort (263 samples) involved similar challenges to those we encountered in the analysis of the colon adenocarcinoma dataset. That is, a bimodal distribution of the mutational load which entailed significant mutational load effects in all of the networks, as assessed by our connectivity analysis across a series of networks, generated by the Mapper. The mutational load effects were removed after a subsampling process (Figure 16A).

Connectivity analysis over a fine range of networks (Figure 16B, red square), revealed in total nine cancer-associated genes (Figure 16C); four of them (*PIK3CA,CDH1,ARID1A,TP53*) were previously reported (*45*), one (*AKAP13*) reproduced by MutSig2CV, and additional four novel candidates. *PEG3* (mutated in 9.8% of the samples), a novel candidate, promotes p53-mediated apoptosis (*68*) and demonstrates tumor suppressor activity in glioma cell lines (*69*), however *PEG3* mutations were never reported associated in stomach adenocarcinoma to our best knowledge. Other novel candidates (*UNC13C, AFF2, PLXNA4*) do not demonstrate a trivial association with cancer, although PLXNA4 encodes the Plexin-A4 protein, a member of the semaphorin receptor family, which is implicated in various oncogenic processes (*70*). *AFF2* is associated with Fragile X syndrome, a genetic condition involving genomic instabilities of the X chromosome (*71*).

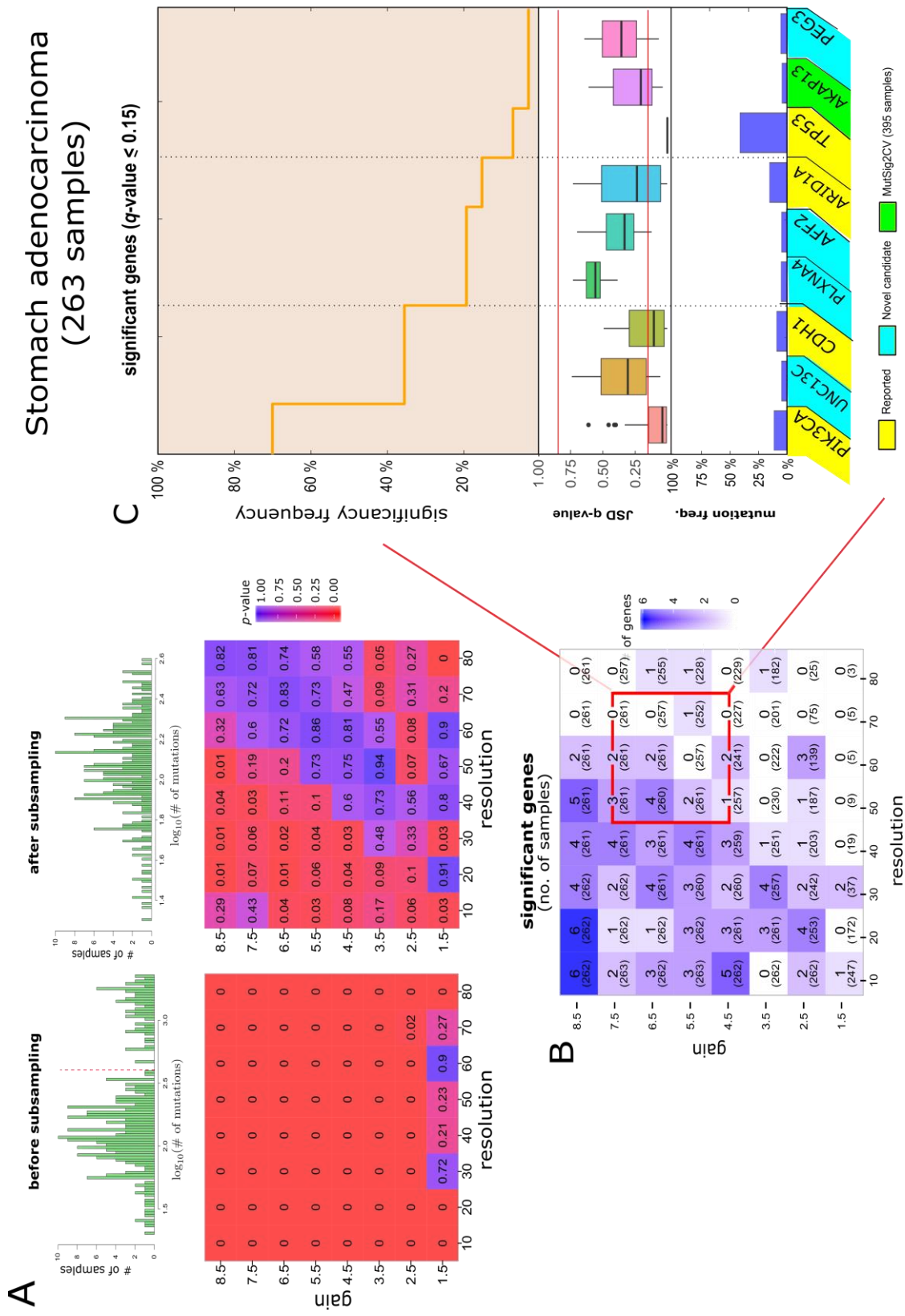

Stomach adenocarcinoma
(263 samples)
A
before subsampling
after subsampling
# of samples
log10(# of mutations)
gain
resolution
p-value
B
significant genes
(no. of samples)
# of genes
C
significant genes (q-value ≤ 0.15)
significance frequency
mutation freq.
JSD q-value
PIK3CA
UNC13C
CDH1
PLXNA4
AFF2
ARID1A
TP53
AKAP13
PEG3
Reported
Novel candidate
MutSig2CV (395 samples)

Figure 16 | Identification of Stomach Adenocarcinoma Associated Mutations using TDA. (A) Top left: mutational load distribution in a logarithmic scale across all samples. Before subsampling (Methods) there is an observed bi-modal distribution, separating patients into hypermutated and non-hypermutated cases. The dashed red line represents a subsampling threshold. Grid below the distributions summarizes the statistical significance of the mutational load, as assessed by our connectivity analysis across a series of networks, generated by Mapper (Methods). In most networks the mutational load is significant ($p<0.05$). After subsampling (Methods) as per the threshold, the new mutational load distribution is centered around the mutational load median of the non-hypermutated case. The mutational load is no longer significant in most networks. (B) Connectivity analysis summary of selected mutations in a coarse grid of networks generated using the Mapper algorithm (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). The red square indicates the selected range of networks used in a second connectivity analysis for same mutated genes (Methods). (C) At the bottom are cancer-associated genes identified by the connectivity analysis. Previously reported genes are depicted in yellow, novel candidates in cyan, reproduced by MutSig2CV in green. Genes are ordered by significance frequency across the networks as presented in the top step function. Boxplots represent JSD q-value of the gene (Methods). Below the red line ($q<0.15$) are genes that show significant positive selection. Above the red line ($q>0.85$) are probable artifacts (Methods). The histogram represents the mutation frequency of the gene in the cohort.

METHODS

Samples collection and preprocessing

We collected global gene expression levels and somatic mutation data of seven tumor types from the public repository TCGA. We only considered patients for which both gene expression and mutation data were available. RNA-seq expression levels were retrieved in RSEM (RNA-Seq by Expectation-Maximization) format (Supplemental Table 3). We expressed RSEM output transcripts per million (TPM) on a logarithmic scale, using the following formula: $r = \log_2(1 + x * 10^6)$ where x is the estimated relative abundance of a particular gene. We then used *r* values to constitute a matrix in which each row represents one patient and each column represents a different gene.

Somatic mutation data were retrieved from Broad's institute firehose pipeline (*47*), which aggregates and curates somatic mutation data from various mutation calling centers (Supplemental Table 3). Somatic mutations were identified by the original center by comparing tumor to normal samples of each patient. Further details on the specific thresholds used by each center can be found in the original TCGA publications for each tumor (*8*–*10*, *42*–*45*)). Common variants were filtered out by removing mutations in sites that were present in dbSNP (*72*), a repository of known common variants. The final MAF files from the various centers were further curated by Broad's Oncotator software (*73*), which annotates reads based on the Human Genome Reference Consortium build 37 (GRCh37) and incorporates additional common variant information from 1000Genomes (*74*). Since the intersection of available mutation data and gene expression data is often small, we preferred to use an expanded somatic mutation dataset when available (marked as Oncotator_RAW).

From the MAF file, we extracted patient's ID, mutated genes, and the mutation type (non-synonymous or synonymous) and generated a non-synonymous binary matrix, where rows represent patients and columns correspond to individual genes. The binary

entries (0 or 1) in the matrix describe whether the gene appeared non-synonymously mutated at least once in the patient's genome. We later calculated for each gene the ratio between non-synonymous mutations and a total number of mutations of that gene. We used this ratio as a scoring system to filter out genes in our downstream connectivity analysis.

Finally, we intersected the expression matrix and the non-synonymous binary matrix based on common samples and concatenated them together as input to the Mapper algorithm. To avoid discrepancies between gene expression and annotations, we converted all annotations to comply with NCBI's Entrez ID database as of July-07-2015.

Topological representations

We mapped the multidimensional gene expression data to a two-dimensional network using the Mapper algorithm as implemented in Ayasdi Cure software (*31*). The use of Mapper is particularly suited to complex biological data sets with a continuous structure such as global gene expression levels. The resulting topological representation is a two-dimensional network comprised of nodes and edges that preserve the continuous global expression patterns in the original multidimensional space. Each node in the network contains data points (samples) that are similar to one another in terms of gene expression. Edges connect samples that share at least one data point. Nodes that are not part of the largest connected component of the network were considered as outliers and removed from the downstream analysis (the percentage of nodes removed varies with each generated network, for absolute numbers see Results section).

To construct the network, we used Pearson's correlation between the 4,500 genes with the highest variance as a metric, and two nearest neighborhood lenses as our filtering functions. Correlation is a standard measure of similarity between gene expression patterns, being less sensitive to normalization effects in comparison to other similarity

measures such as Euclidean distance. The choice of k-nearest neighbor lenses was based on their performance in capturing known biological features, such as separation between normal and tumor samples and agreement with known disease subtypes.

Connectivity analysis

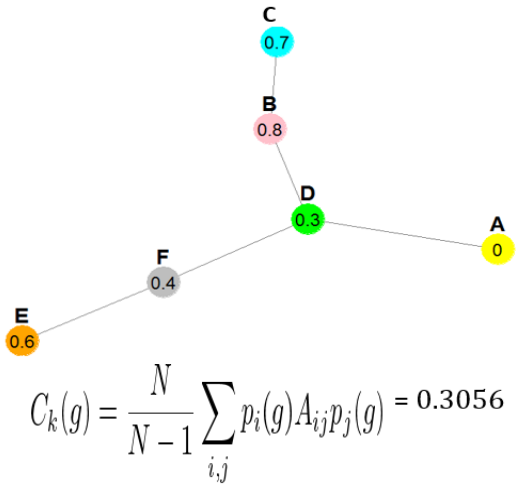
We nominated a feature (such as a given mutated gene) as associated with global gene expression and therefore with the disease if it is localized or connected in the corresponding simplicial complex more than random. These features can be any function with support on the simplicial complex, e.g. frequency of a somatic mutation within each node, mutational load or batch effects. We quantified their localization using the following equation:

$$C_k(g) = \frac{N}{N-1} \sum_{i,j} p_i(g) A_{ij} p_j(g)$$

where, for every feature of interest (g) we measured the connectivity value (C) by summing over all nodes pairs (i, j) the normalized feature value (p) times the adjacency matrix (A) of the simplicial complex. In this expression, $p_i = \frac{e_i}{\sum e_i}$ where e_i is the average value of the feature in that node. We also normalized this quantity using the network size, measured as the total number of nodes (N) in the network. In this way, connectivity can only take values between 0 and 1. An example is provided in Figure 17.

To assess the statistical significance of the connectivity, we generated a null distribution of connectivity values for each tested feature by permuting the labels (samples) across the network ($10^{\wedge}4$ times when testing for cancer-associated genes and $2.5x10^{\wedge}3$ when testing for mutational load and batch effects). The p-value is simply the ratio between the number of connectivity values larger than the true connectivity value

and the total number of permutations. Finally, we adjusted for multiple testing using Benjamini-Hochberg's false discovery rate (FDR). An example of significantly connected mutations in lower-grade glioma patients is provided in Supplemental Figure 1.



Normalized values $p_i = \frac{e_i}{\sum e_i}$

A	B	C	D	E	F
0	0.3	0.25	0.11	0.21	0.14

Adjacency matrix (A)

	A	B	C	D	E	F
A	0	0	0	1	0	0
B	0	0	1	1	0	0
C	0	1	0	0	0	0
D	1	1	0	0	0	1
E	0	0	0	0	0	1
F	0	0	0	1	1	0

$$C_k(g) = \frac{N}{N-1}\sum_{i,j} p_i(g)A_{ij}p_j(g) = 0.3056$$

Figure 17 | Connectivity analysis. Connectivity is calculated for a feature over a network of six nodes (A-F) and five edges. The values in the nodes represent the feature average value (e_i) within the node (In the context of this work, the value represents the fraction of patients within the node that harbor a specific mutation). The normalized values for each node and the adjacency matrix of the network are used to calculate the connectivity of the feature in the network.

Genes filtering for connectivity analysis

Not all genes were selected for connectivity analysis, to avoid corrections that were too large due to multiple testing adjustments. Accordingly, we set two types of filtering thresholds which together allowed a maximum of 350 genes (This number was chosen based on our observations that it is large enough to allow identification of non-trivial genes, yet it is small enough to avoid large correction due to multiple testing adjustments). The first filter sets a threshold based on recurrence in the cohort, ranging from 4 to 6 percent of the samples in most cases while allowing a lower threshold in some large cohorts. The second filter ranks genes based on their non-synonymous ratio, defined as the number of non-synonymous mutations of the gene over its total number of mutations in the cohort. We kept the 350 top ranked genes.

Subsampling of hypermutated samples and batch effects

Our method is sensitive to situations where there is an association between differential mutation rates and global gene expression patterns. This could happen in cases where a mutation in the DNA mismatch repair pathway (such as MSH3) render cells hypermutated and simultaneously leads to a similar transcriptional program in the samples with the mutated version of the gene. In order to circumvent these effects, we statistically assessed the connectivity of the mutational load function in the simplicial complex, defined as the average number of the mutations in each node. If the mutational load was significantly connected in the networks ($p<0.05$), then we subsampled mutations from the hypermutated samples and tested for mutational load connectivity again.

The exact steps we took are as follows.

1. We examined the mutational load histogram (c.f. example in Figure 16A) and set a threshold that separates the samples into two groups: hypermutated and non-hypermutated.
2. We defined a subsampling factor $s = (a/b)$ where

$$a = median\ of\ mutational\ load\ of\ hypermutated\ samples.$$

$$b = median\ of\ mutational\ load\ of\ not\ hypermutated\ samples.$$

3. We subsampled mutations from the hypermutated samples by keeping randomly 1 of every s mutations, so that both groups have the same median after subsampling.
4. We tested for mutational load effects again. If still significant, we chose another threshold and repeated the process from step 1.

Batch effects are also potential confounding factors. In the Colon Adenocarcinoma dataset, for instance, somatic mutation data is aggregated from different mutation calling centers (Supplemental Table 3). Differences in technologies lead to a certain bias. We

marked samples based on their center of origin and tested for connectivity on the simplicial complex of each batch. We removed the contribution of results originating from networks that demonstrated statistically significant batch effects.

Parameter scan and statistical power

To increase statistical power and robustness of the results, we generated for each cancer type a series of topological representations by scanning over the Mapper algorithm *gain* and *resolution* parameters. We scanned over the resolution range of 10-80, in coarse intervals of 10, and the gain range 1.5-8.5, with a coarse interval of 1. This procedure yields a series of 49 simplicial complexes or networks. On each network, we used the analysis described above to test for localization of various features, including batch effects, mutational load, and somatic mutations. We then chose a finer range of networks for each cancer type, based on the following criteria:

1. Relatively a high number of significant genes ($q<0.15$) as determined by our connectivity analysis. This increases our statistical power or sensitivity (Figure 18).
2. Relatively a high number of samples in the first connected component of the networks. Since we considered only the first connected component in our downstream analysis, networks with as many samples as possible were preferred.
3. Not significant mutational load and batch effects.

We generated a finer set of networks with resolution and gain intervals of 5 and 0.5, respectively, for the selected regions of the parameter space. We tested for localization of somatic mutations across the fine range of networks and summarized the results.

$$statistical\ \text{power}\ = sensitivity = \frac{TP}{TP + FN}$$

$$FDR\ = \frac{FP}{TP + FP} \qquad statistical\ \text{power} = \frac{k(1 - FDR)}{k(1 - FDR) + FN}$$

$$k = (TP + FP) = no.of\ significant\ genes$$

$$statistical\ \text{power} = \frac{1 - FDR}{1 - FDR + \frac{FN}{k}}$$

$$k * FDR = FP$$

$$k(1 - FDR) = TP \qquad k \to \infty : \quad \boldsymbol{statistical\ power} \to \mathbf{1}$$

Figure 18 | Statistical power. The relationship between the number of significant genes and the statistical power.

Positive selection control

Studies in cancer (*17*) have shown the existence of anti-correlation between gene expression levels and neutral mutation rates of that gene (Figure 3). This is partially explained by the transcription-coupled repair mechanism, a process in which transcribed regions are more accessible to repair processes such as nucleotide excision repair enzymes (*21*). Accordingly, we reasoned that if the mutation rate of a gene deviates from the expected anti-correlation with the gene's expression levels, this implies some positive selection of this mutation. Conversely, an agreement between low mutation rates and high expression levels could be explained by the neutral model and, therefore, is an indication a false positive.

For every statistically significant mutated gene according to the above analysis, we measured the similarity between its mutation rate and gene expression distributions across the simplicial complex using Jensen-Shannon Distance (JSD). Boxplot diagrams (Figure 9C, boxplot) represent the distribution of JSD's q-value (statistical significance after adjusting for multiple testing using Benjamini-Hochberg procedure) for that gene across a series of simplicial complexes, as determined by a permutation test of 2,000 permutations. Mutation rates for genes below the red line represent a significant ($q < 0.15$)

deviation from the expected model and are probably subject to positive selection. Indeed, in many cases we found well-established cancer-associated genes below the red line (Figure 9C, boxplot). Genes above the red line are in agreement with the neutral model ($q > 0.85$) and are probably false positives. Mutation rates for genes in between the red lines are not correlated nor anti-correlated with expression levels, hence, this test does not add or remove confidence in the genes' association with the disease.

DISCUSSION

The application of recent mathematical development in topological data analysis, specifically the Mapper algorithm, has been demonstrated, in this work, to be highly effective in analyzing complex genomic datasets involved in cancer research. Mapper's ability to reduce the dimensionality of the data, while reliably capturing the continuous structure of the global gene expression space, allowed us to implement an analysis pipeline involving an original statistical test (connectivity analysis) in order to extract novel as well as previously reported cancer associated genes. This robust and scalable framework performed well across seven tumor types, including challenging cases such as the hypermutated landscape of colon and stomach adenocarcinoma. Importantly, reproducing previous reports of cancer-associated genes set confidence in our novel method, and it highlights the power of topological data analysis, and specifically the Mapper algorithm, in the analysis of complex and multidimensional datasets that we often encounter in cancer research and other fields of science.

Our method brings a new perspective from a different angle into previous cancer studies. Using our orthogonal approach which does not rely on complex modeling of the mutational landscape, but rather on a topological representation of the global gene expression levels (the disease phenotype) is very informative when re-analyzing heterogeneous and complex conditions such as cancer, where complex modelling of the mutational landscape is involved and introduces systematic errors to the analysis. Our robust results are, therefore, instrumental in further solidifying previous reports, and also provides insight into novel cancer-associated genes, which, possibly, open the door to subsequent cancer studies, both computational and experimental.

Admittedly, deeper analysis of the results, along the lines of association with known subtypes, pathway analysis, correlation with epigenomic markers and copy number alterations is required prior to assembling a viable biological story that can be further validated experimentally. Repeating similar analysis on tumor subsets such as non-hypermutated samples in colon adenocarcinoma or distinct clinical subsets of breast cancer would be instrumental in depicting the landscape of the disease in even finer detail. Nevertheless, a clear picture emerges: our method is capable of identifying cancer-associated genes, even in complex hypermutated tumors such as colon and stomach adenocarcinoma, recapitulating well reported ones and augment the mutational landscape of other cancers with novel candidates such as *FMN2* mutations in lung and bladder cancer, *PTPRD* in bladder cancer, and adding further support to recently reported findings in gliomas such as *NIPBL* mutations and *SOX9* mutations in colon adenocarcinoma. Some of our novel candidates, such as the group of ion-channel genes in lung adenocarcinoma (*ANO4, SLC8A1, ANK2,SCN2A and CACNA2D)* are particularly encouraging since ion-channels have already been explored as therapeutics targets (*53*). Moreover, since our method is tuned to identify cancer-associated genes based on their association with the disease phenotype, namely, global gene expression levels, we are able to detect rarely mutated genes as well as genes that are potentially dismissed due to low recurrence or irrelevant genomic characterizations such as replication time and gene length, parameters taken into account by standard methods.

Moving forward, we will further solidify our results, with the aim to experimentally validate some of the novel candidates. Additionally, we have already started expanding the project to include all other tumor types in TCGA for which sufficient data exists, namely a minimum of 150 patients. A few tumor cohorts apply and include lung, cervical, and head and neck squamous cell carcinoma, as well as liver and thyroid carcinoma. Pan-glioma

and colorectal studies, as well as analysis of clinical subset in breast cancer, are also possible future projects, in consonance with previous research (42)(*44*) (58).

Before we conclude, it's worth mentioning some of the limitations of the method that we have developed. Since we are identifying genes based on association with gene expression levels, we might miss mutations that are related to the disease but have a mild effect on expression levels. Similarly, our method is sensitive to the association between differential mutation rates and differential expression levels (Methods) and, therefore, requires testing for mutational load effects which are sometimes hard to remove (for this reason we neglected analysis of skin cutaneous melanoma).

Additionally, due to a large correction for multiple testing, our method is limited to 350 potential genes (sweet spot number of genes that allows identification of non-trivial genes yet does not incur too large correction from multiple testing), as determined by our filtering parameters (Methods), inevitably missing potential candidates in the process. In terms of statistical power, our method does not seem to be effective with cohorts smaller than 140-150 patients, which was just about enough to include the GBM cohort (142 samples) in our analysis. Lastly, since our method relies on many parameters related to the Mapper algorithm (i.e. choice of filter and metric functions, and range of gain and resolution parameters), we have to scan over a wide range of gain and resolution parameters, which make it a bit cumbersome, nevertheless robust.

Conclusively, our method emphasizes the robustness and viability of topological data analysis in general and specifically highlights the advantages of using the Mapper algorithm, in analyzing multidimensional datasets that we often find in cancer research and in other fields of science. In this era of data, as such datasets are becoming more readily available, it is safe to assume an increasing number of open data-driven questions, which invites more development and application of novel analytical methods of this kind. Finally, the method that we have developed and demonstrated here proved powerful for

identifying cancer-associated genes based on gene expression levels, which is a naively more suitable approach to detect cancer-associated genes than other methods relying on recurrence or complex modeling, which do not account for the disease phenotype. However, this is not to argue in any way that our method is superior to other methods. Instead, we offer this framework as a complementary method to currently employed techniques in cross-sectional cancer studies, and this is just the beginning of an effort to provide a new perspective on the identification of genes that may play a role in cancer.

REFERENCES

1. R. Weinberg, *The biology of cancer* (Garland science, 2014).
2. B. Stewart, C. P. Wild, World cancer report 2014. *World* (2015).
3. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature*. **458**, 719–724 (2009).
4. D. Hanahan, R. A. Weinberg, Review Hallmarks of Cancer : The Next Generation. *Cell*. **144**, 646–674 (2011).
5. C. G. A. R. N. Weinstein, John N., Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
6. A. Hudson, T.J., Anderson, W., Aretz, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S. and Guttmacher, International network of cancer genome projects. *Nature*. **464**, 993–8 (2010).
7. N. Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M.S., Getz, G., Bader, G.D., Ding, L. and Lopez-Bigas, Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
8. G. Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P. and Alexe, Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. **17**, 98–110 (2010).
9. Cancer Genome Atlas Research Network, Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. **511**, 543–550 (2014).

10. Cancer Genome Atlas Research Network, Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*. **507**, 315–322 (2014).

11. J. C. Singh, K. Jhaveri, F. J. Esteva, HER2-positive advanced breast cancer: optimizing patient outcomes and opportunities for drug development. *Br. J. Cancer*. **111**, 1888–98 (2014).

12. C. Gravalos, A. Jimeno, HER2 in gastric cancer: A new prognostic factor and a novel therapeutic target. *Ann. Oncol.* **19**, 1523–1529 (2008).

13. Y. J. Bang, E. Van Cutsem, G. Bang, Y.J., Van Cutsem, E., Feyereislova, A., Chung, H.C., Shen, L., Sawaki, A., Lordick, F., Ohtsu, A., Omuro, Y., Satoh, T. and Aprile, Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): A phase 3, open-label, randomised controlled trial. *Lancet*. **376**, 687–697 (2010).

14. G. Valabrega, F. Montemurro, M. Aglietta, Trastuzumab: Mechanism of action, resistance and future perspectives in HER2-overexpressing breast cancer. *Ann. Oncol.* **18**, 977–984 (2007).

15. B. J. Raphael, J. R. Dobson, L. Oesper, F. Vandin, Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* **6**, 5 (2014).

16. B. J. Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C. and Raphael, Pathway and network analysis of cancer genomes. *Nat. Methods*. **12**, 615–621 (2015).

17. A. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. and Kiezun, Mutational heterogeneity in cancer and the search for new cancer-associated

genes. *Nature*. **499**, 214–8 (2013).

18. R. K. Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R. and Wilson, MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).

19. K. C. Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M. and Anderson, Initial genome sequencing and analysis of multiple myeloma. *Nature*. **471**, 467–72 (2011).

20. E. Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A. and Shefler, The mutational landscape of head and neck squamous cell carcinoma. *Science (80-.).* **333**, 1157–1160 (2012).

21. M. Fousteri, L. H. F. Mullenders, Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. *Cell Res.* **18**, 73–84 (2008).

22. K. D. Makova, R. C. Hardison, The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **advance on**, 213–223 (2015).

23. A. Gonzalez-Perez, N. Lopez-Bigas, Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, 1–10 (2012).

24. V. Trifonov, L. Pasqualucci, R. Dalla Favera, R. Rabadan, MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. *BMC Syst. Biol.* **7**, 25 (2013).

25. R. Ghrist, *Elementary applied topology* (Createspace, 2014).

26. G. L. Alexanderson, About the Cover: Euler and Königsberg's Bridges: a Historical View. *Bull. New. Ser. Am. Math. Soc.* **43**, 567–573 (2006).

27. P. E. C. Compeau, P. A. Pevzner, G. Tesler, How to apply de Bruijn graphs to genome assembly. *Nat Biotech*. **29**, 987–991 (2011).

28. A. Hatcher, Algebraic topology. 2002. *Cambridge UP, Cambridge*. **606**.

29. D. S. Richeson, *Euler's Gem: The polyhedron formula and the birth of topology* (Princeton University Press, 2012).

30. M. Nakahara, Series Editor : GEOMETRY , TOPOLOGY AND PHYSICS. *Text*. **822**, 173–204 (2003).

31. G. Carlsson, *Topology and Data* (2009), vol. 46.

32. M. Nicolau, A. J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7265–70 (2011).

33. J. T. Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P. and Dudley, Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 1–16 (2015).

34. R. Hinks, T.S., Brown, T., Lau, L.C., Rupani, H., Barber, C., Elliott, S., Ward, J.A., Ono, J., Ohta, S., Izuhara, K. and Djukanović, Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3–like protein 1. *J. Allergy Clin. Immunol.*, 1–15 (2016).

35. J. Tierny, J.-P. Vandeborre, M. Daoudi, 3D Mesh Skeleton Extraction Using Topological and Geometrical Analyses. *14th Pacific Conf. Comput. Graph. Appl. (Pacific Graph. 2006)* (2006), p. s1poster.

36. J. M. Chan, G. Carlsson, R. Rabadan, Topology of viral evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18566–71 (2013).

37. G. Carlsson, Topological pattern recognition for point cloud data. *Acta Numer.* **23**, 289–368 (2014).

38. G. Carlsson, T. Ishkhanov, V. De Silva, A. Zomorodian, On the local behavior of spaces of natural images. *Int. J. Comput. Vis.* **76**, 1–12 (2008).

39. R. W. Ghrist, Barcodes: the Persistent Topology of Data. *Bull. (New Ser. THEAMERICAN Math. Soc.* **45**, 15 (2008).

40. G. Singh, F. Mémoli, G. Carlsson, Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *Methods*, 91–100 (2007).

41. S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding. *Science*. **290**, 2323–6 (2000).

42. Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. **487**, 330–337 (2012).

43. L. Gliomas, Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N. Engl. J. Med.*, 2481–2498 (2015).

44. Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours. *Nature*. **490**, 61–70 (2012).

45. Cancer Genome Atlas Network, Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*. **513**, 202–9 (2014).

46. C. Frattini, V., Trifonov, V., Chan, J.M., Castano, A., Lia, M., Abate, F., Keir, S.T., Ji, A.X., Zoppoli, P., Niola, F. and Danussi, The integrated landscape of driver genomic alterations in glioblastoma. *Nat. Genet.* **45**, 1141–9 (2013).

47. M. J. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, GenePattern 2.0. *Nat Genet*. **38**, 500–501 (2006).

48. B. M. Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P. and Bot, The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).

49. R. S. and G. C. Rudd, M.F., Webb, E.L., Matakidou, A., Sellick, G.S., Williams,

R.D., Bridle, H., Eisen, T., Houlston, Variants in the GH-IGF axis confer susceptibility to lung cancer. *Genome Res.* **16**, 693–701 (2006).

50. K. Yamada, M. Ono, N. D. Perkins, S. Rocha, A. I. Lamond, Identification and Functional Characterization of FMN2, a Regulator of the Cyclin-Dependent Kinase Inhibitor p21. *Mol. Cell*. **49**, 922–933 (2013).

51. A. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. **14**, 128 (2013).

52. N. Prevarskaya, R. Skryma, Y. Shuba, Ion channels and the hallmarks of cancer. *Trends Mol. Med.* **16**, 107–121 (2010).

53. A. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R. and Ma'ayan, Ion channel gene expression in lung adenocarcinoma: potential role in prognosis and diagnosis. *PLoS One*. **9**, e86569 (2014).

54. S. Cal, J. M. Argüelles, P. L. Fernández, C. López-Otín, Identification, Characterization, and Intracellular Processing of ADAM-TS12, a Novel Human Disintegrin with a Complex Structural Organization Involving Multiple Thrombospondin-1 Repeats. *J. Biol. Chem.* **276**, 17932–17940 (2001).

55. Q. Wu, T. Maniatis, A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*. **97**, 779–790 (1999).

56. A. Ostman, C. Hellberg, F. D. Böhmer, Protein-tyrosine phosphatases and cancer. *Nat. Rev. Cancer*. **6**, 307–320 (2006).

57. J. Chen, C.L., Cen, L., Kohout, J., Hutzen, B., Chan, C., Hsieh, F.C., Loy, A., Huang, V., Cheng, G. and Lin, Signal transducer and activator of transcription 3 activation is associated with bladder cancer cell growth and survival. *Mol. Cancer*. **7**, 78 (2008).

58. S. Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray,

B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M. and Anjum, Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*. **164**, 550–563 (2016).

59. G. Wang, X., Lu, G., Li, L., Yi, J., Yan, K., Wang, Y., Zhu, B., Kuang, J., Lin, M., Zhang, S. and Shao, HUWE1 interacts with BRCA1 and promotes its degradation in the ubiquitin-proteasome pathway. *Biochem. Biophys. Res. Commun.* **444**, 549–554 (2014).

60. L. O'Neill, C.F., Urs, S., Cinelli, C., Lincoln, A., Nadeau, R.J., León, R., Toher, J., Mouta-Bellum, C., Friesel, R.E. and Liaw, Notch2 signaling induces apoptosis and inhibits human MDA-MB-231 xenograft growth. *Am. J. Pathol.* **171**, 1023–1036 (2007).

61. I. Espinoza, L. Miele, Notch inhibitors for cancer treatment. *Pharmacol. Ther.* **139**, 95–110 (2013).

62. L. O'Neill, C.F., Urs, S., Cinelli, C., Lincoln, A., Nadeau, R.J., León, R., Toher, J., Mouta-Bellum, C., Friesel, R.E. and Liaw, Tumor cell-specific inhibition of MYC function using small molecule inhibitors of the HUWE1 ubiquitin ligase. *EMBO Mol. Med.* **6**, 1–17 (2014).

63. W. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., ... & Spooner, Ensembl Comparative Genomics Resources. *Database*, 1–17 (2016).

64. S. T. Hirai, Y., Utsugi, K., Takeshima, N., Kawamata, Y., Furuta, R., Kitagawa, T., ... & Noda, Putative gene loci associated with carcinogenesis and metastasis of endocervical adenocarcinomas of uterus determined by conventional and array-based CGH. *Am. J. Obstet. Gynecol.* **191**, 1173–1182 (2004).

65. T. A. Lagace, PCSK9 and LDLR degradation: regulatory mechanisms in circulation and in cells. *Curr. Opin. Lipidol.* **25**, 387–93 (2014).

66. B. T. MacDonald, K. Tamai, X. He, Wnt/??-Catenin Signaling: Components,

Mechanisms, and Diseases. *Dev. Cell*. **17**, 9–26 (2009).

67. E. a. Ariazi, G. M. Clark, J. E. Mertz, Estrogen-related receptor alpha and estrogen-related receptor gamma associate with unfavorable and favorable biomarkers, respectively, in human breast cancer. *Cancer Res.* **62**, 6510–6518 (2002).

68. Y. Deng, X. Wu, inducing Bax translocation from cytosol to mitochondria (2000).

69. T. Kohda, T., Asai, A., Kuroiwa, Y., Kobayashi, S., Aisaka, K., Nagashima, G., & Kaneko- Ishino, Tumour suppressor activity of human imprinted gene PEG3 in a glioma cell line. *Genes Cells*. **6**, 237–247 (2001).

70. S. Rizzolio, L. Tamagnone, Semaphorin signals on the road to cancer invasion and metastasis. *Cell Adh. Migr.* **1**, 62–68 (2007).

71. H. A. Lubs, R. E. Stevenson, C. E. Schwartz, Fragile X and X-linked intellectual disability: Four decades of discovery. *Am. J. Hum. Genet.* **90**, 579–590 (2012).

72. K. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–11 (2001).

73. G. Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., Getz, Oncotator: Cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).

74. 1000 Genomes Project Consortium., A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

APPENDIX

Supplemental Table 1 | Extended results list. The columns' labels represent abbreviated tumor name provided by TCGA and are sorted alphabetically. Previously reported genes are depicted in yellow, reproduced by MutSig2CV in green, novel candidates in blue, artifacts in red. Superscript digit next to the artifacts indexes the reason for exclusion: 1 - low expression levels (TPM<2), 2 - false positive (JSD q-value>0.85), 3 - batch effects, 4 - pseudogene.

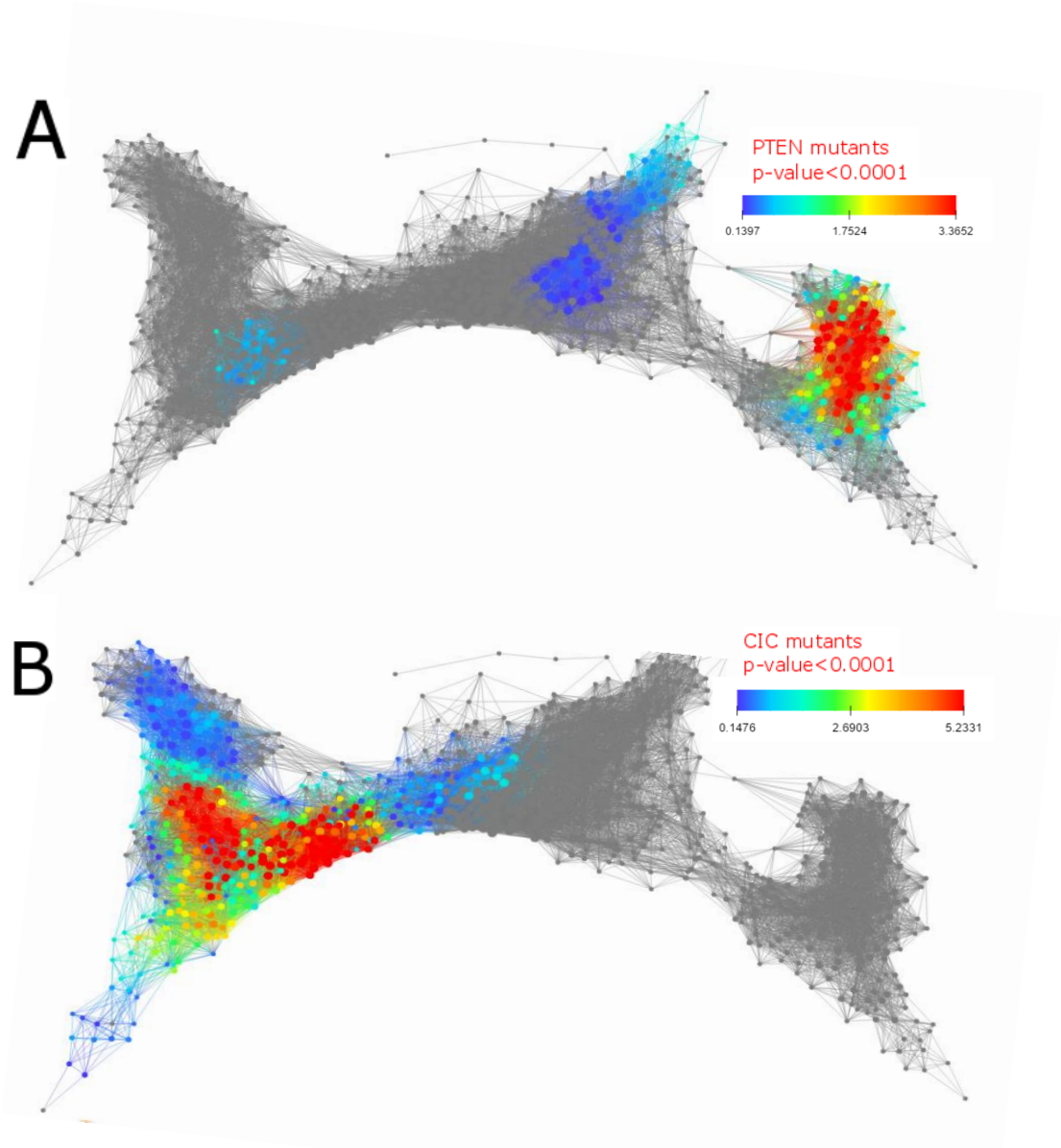
BLCA	BRCA	COAD	GBM	LGG	LUAD	STAD
FGFR3	*TP53*	*SOX9*	*IDH1*	*IDH1*	*STK11*	*PIK3CA*
RB1	*PIK3CA*	*APC*	*ATRX*	*NOTCH1*	*EGFR*	*CDH1*
ELF3	*CDH1*	*PIK3CA*	*EGFR*	*PTEN*	*KEAP1*	*ARID1A*
TP53	*MAP3K1*	*TP53*	*NF1*	*TP53*	*AKAP9*	*TP53*
PTPRD	*CBFB*	*SMAD4*	*COL6A3*	*CIC*	*KRAS*	*AKAP13*
MED13	*GATA3*	*TCF7L2*	*CALN1*	*FUBP1*	*ATM*	*UNC13C*
FMN2	*MAP2K4*	*RNF43*	*SLCO6A1* [1]	*ATRX*	*TP53*	*PLXNA4*
HSPG2	*RUNX1*	*KMT2C*		*EGFR*	*SMARCA4*	*AFF2*
MUC17 [2]	*CTCF*	*PIK3R1*		*NF1*	*SLC8A1*	*PEG3*
HERC2P2 [4]	*NOTCH2*	*KRAS*		*TCF12*	*SCN2A*	
	FAT3	*ARHGAP5*		*NIPBL*	*SLITRK4*	
	LRP2	*ARFGEF1*		*ZBTB20*	*SATB2*	
	HUWE1	*VPS13B*		*SMARCA4*	*CPED1*	
	USH2A [1]	*FLT3*		*ZNF292*	*CCDC129*	
		NEFH		*IDH2*	*FMN2*	
		CCDC141		*COL6A3*	*PCDHB4*	
		STK11		*SYNE1*	*POLQ*	
		ESRRA		*BAGE2* [1]	*GPR158*	
		NCOR1		*POTEC* [1]	*DGKB*	
		NLRP13 [1]		*FAM47C* [1]	*CACNA2D1*	
		CDH8 [1]			*CHD5*	
		MYH3 [3]			*MYOM2*	
					ADAMTS12	
					EPHB1	
					CDH12	
					RP1L1	
					ANO4	
					TSSC2	
					ANK2	
					CROCCP2	
					PTPRC	
					KLHL4 [2]	
					SI [1]	
					PSG8 [2]	
					FAM47C [1]	

Supplemental Table 2 | Connectivity analysis parameters. Connectivity analysis parameters. This table summarizes the parameters we used to analyze each tumor using our pipeline as described in the Methods section. In the "Subsampling threshold" column, NA means that subsampling is not required. The values in the "Samples threshold" column represent the fraction of samples harboring the mutation as our first filter (For example, 0.06 means 6%).

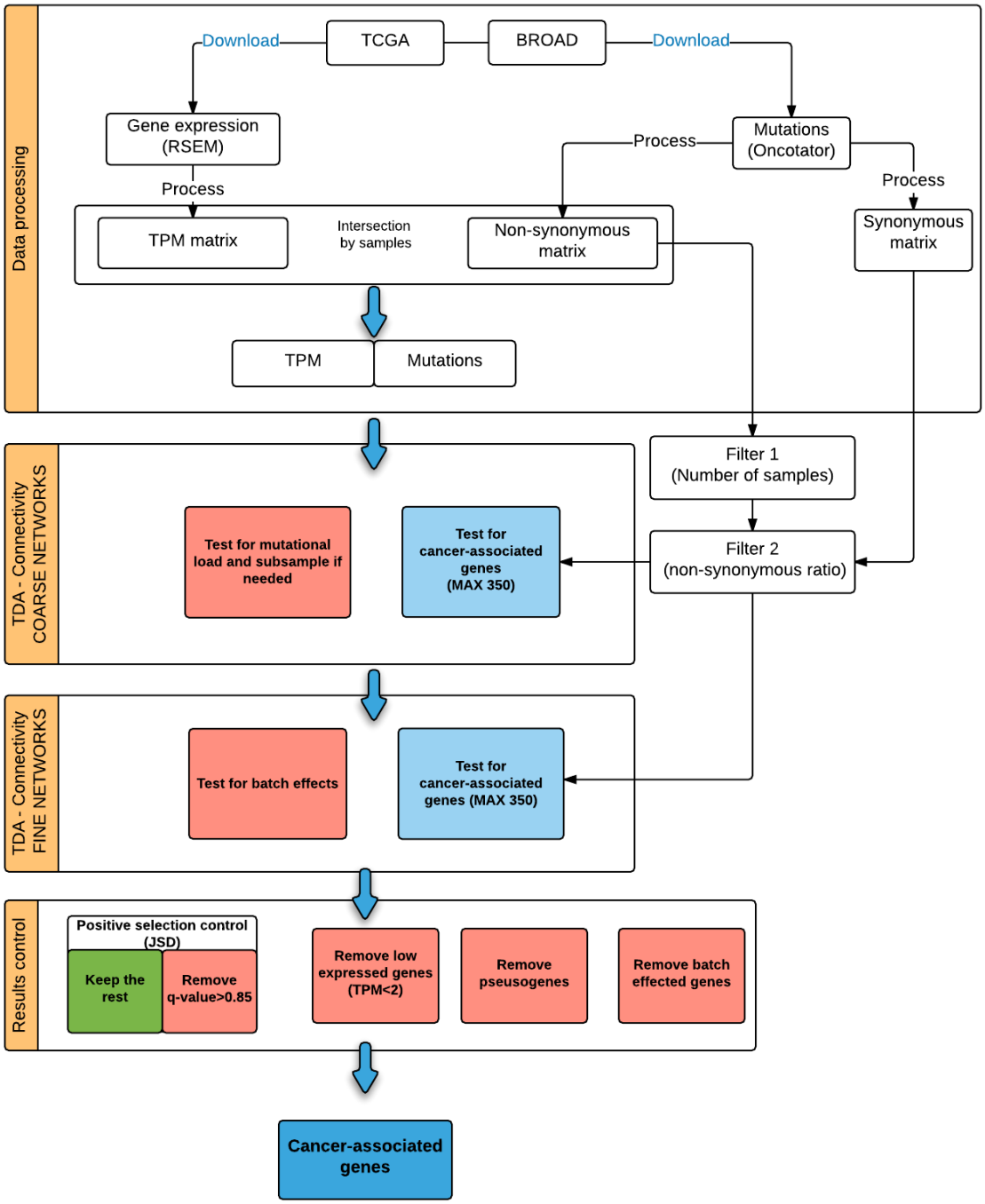
Tumor type	Subsampling threshold	Sample threshold	Non-synonymous ratio threshold	genes permutations	Mutational load permutations
BLCA	NA	0.06	350	10000	2500
BRCA	2.3	0.015	350	10000	2500
COAD	3	0.06	350	10000	2500
GBM	NA	0.04	350	10000	2500
LGG	2.5	0.02	350	10000	2500
LUAD	NA	0.06	350	10000	2500
STAD	2.6	0.04	350	10000	2500

Supplemental Table 3 | Raw data. This table documents RSEM (gene expression levels) version used for each tumor analysis, Oncotator version corresponds to the MAF (mutations data) file collected from the mutation calling center.

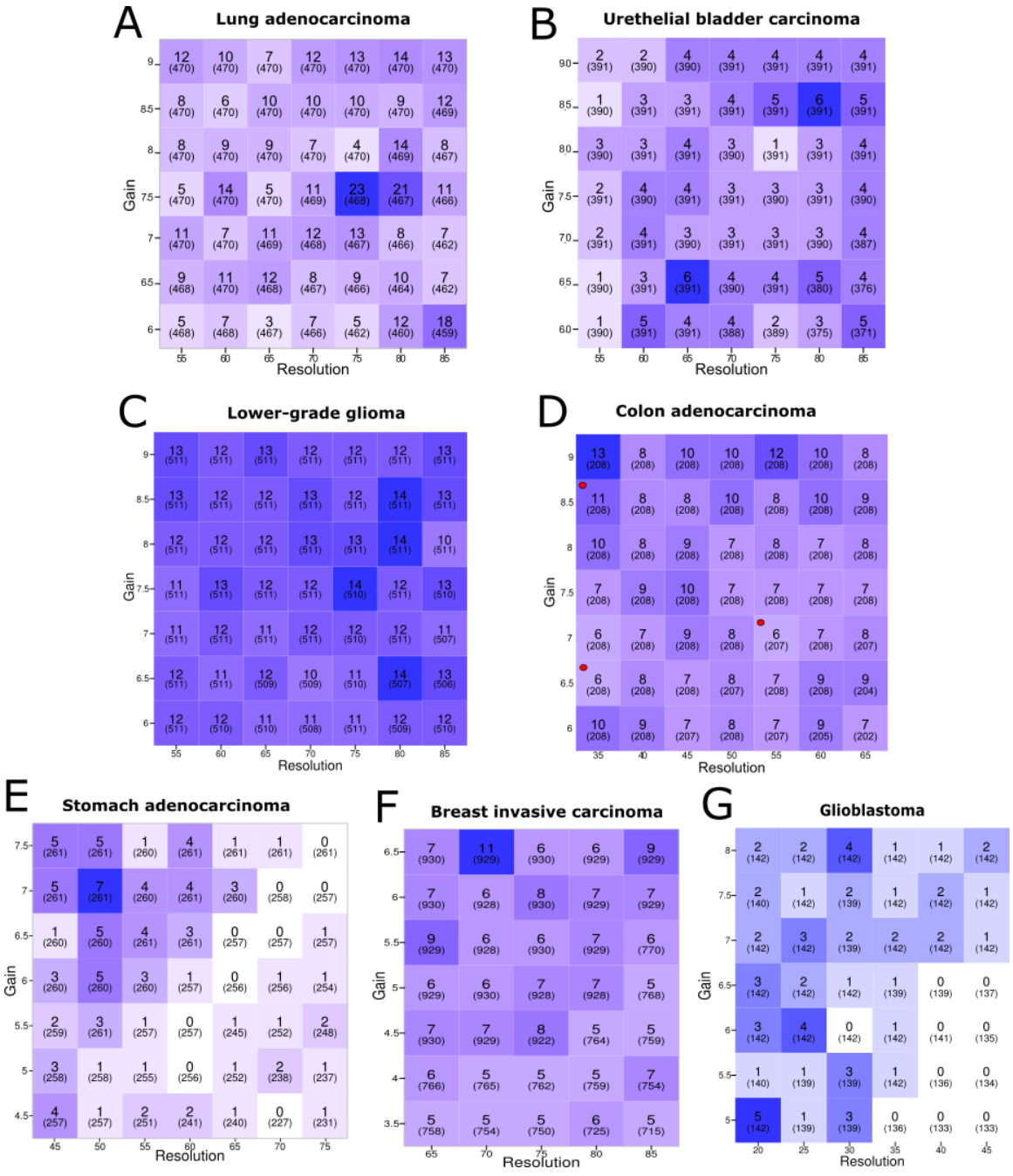
Tumor type	RSEM version (TCGA)	Oncotator MAF version (BROAD)	Mutation calling center
BLCA	3.1.18.0	Raw_Level_3.2015082100	Broad
BRCA	3.1.11.0	Level_3.2015110100	WUSTL
COAD	3.1.12.0	Raw_Level_3.2015082100	Broad, BCM
GBM	3.1.2.0	Raw_Level_3.2015082100	Broad
LGG	3.1.13.0	Raw_Level_3.2015082100	Broad
LUAD	3.1.14	Raw_Level_3.2015082100	Broad
STAD	3.1.0.0	Level_3.2015082100	Broad



Supplemental Figure 1 | *PTEN* and *CIC* mutations in lower-grade glioma. The simplicial complex was generated using the Mapper algorithm on the gene expression levels of 513 lower grade glioma tumors. Each node contains samples clustered together based on similar gene expression levels. Edges connect nodes that share at least one sample. Nodes are colored based on mutation frequency across the samples in each node. Grey means zero mutations (A) Significant localization ($p<0.0001$) of *PTEN* mutations on the right side of the network. (B) *CIC* mutations are significantly localized ($p<0.0001$) on the left part of the network.



Supplemental Figure 2 | Pipeline flow diagram. Our pipeline is separated into four main parts, and is described in detail in the Methods section: 1 - Data processing, which include retrieval of gene expression and mutation data from TCGA and BROAD institute respectively and filtering genes for downstream connectivity analysis. 2 – Topological data analysis followed by connectivity analysis over a coarse range of networks in order to assess mutational load effects and preliminary identification of cancer-associated genes. 3 – Connectivity analysis over a finer range of networks to identify a final list of cancer-associated genes. We also test for batch effects at this stage. 4 – Identifying positively selected genes and controlling for various kinds of artifacts.



Supplemental Figure 3 | Fine networks summary. Fine networks summary. Connectivity analysis summary, across the seven tumors (A-F), of selected mutations in a fine grid of networks generated using the Mapper algorithm (Methods). The fine range was determined after considering various parameters (Methods). Numbers in the tiles represent the total number of significant mutations found in that network. In parenthesis are the number of samples in the first connected component of the network (Methods). Red dots (applies to colon adenocarcinoma only) represent networks in which we identified significant batch effects (Methods).