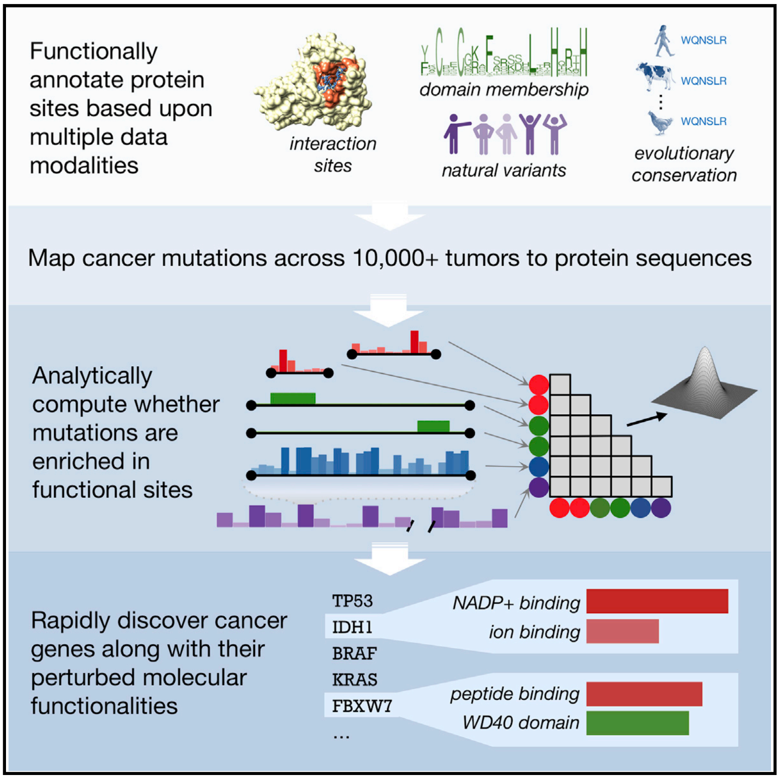
Methods

Cell Systems

PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities

Graphical Abstract


Functionally annotate protein sites based upon multiple data modalities

Map cancer mutations across 10,000+ tumors to protein sequences

Analytically compute whether mutations are enriched in functional sites

Rapidly discover cancer genes along with their perturbed molecular functionalities

Authors

Shilpa Nadimpalli Kobren, Bernard Chazelle, Mona Singh

Correspondence

mona@cs.princeton.edu

In Brief

A fast, analytical framework called PertInInt enables efficient integration of multiple measures of protein site functionality—including interaction, domain, and evolutionary conservation—with gene-level mutation data in order to rapidly detect cancer driver genes along with their disrupted functionalities.

Highlights

- Unified framework identifies cancer genes enriched in mutations in functional sites
- Fast analytical calculations obviate the need for time-prohibitive permutation tests
- Integration of structurally resolved interaction site data for >60% of human genes
- Disrupted molecular functionalities revealed in known or newly predicted cancer genes

Kobren et al., 2020, Cell Systems *11*, 1–12
July 22, 2020 © 2020 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.cels.2020.06.005

Cell Systems

CellPress
OPEN ACCESS

Methods

PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities

Shilpa Nadimpalli Kobren,[1,2,3] Bernard Chazelle,[2] and Mona Singh[2,3,4,*]
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Computer Science, Princeton University, Princeton, NJ, USA
[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
[4]Lead Contact
*Correspondence: mona@cs.princeton.edu

https://doi.org/10.1016/j.cels.2020.06.005

SUMMARY

A major challenge in cancer genomics is to identify genes with functional roles in cancer and uncover their mechanisms of action. We introduce an integrative framework that identifies cancer-relevant genes by pinpointing those whose interaction or other functional sites are enriched in somatic mutations across tumors. We derive analytical calculations that enable us to avoid time-prohibitive permutation-based significance tests, making it computationally feasible to simultaneously consider multiple measures of protein site functionality. Our accompanying software, PertInInt, combines knowledge about sites participating in interactions with DNA, RNA, peptides, ions, or small molecules with domain, evolutionary conservation, and gene-level mutation data. When applied to 10,037 tumor samples, PertInInt uncovers both known and newly predicted cancer genes, while additionally revealing what types of interactions or other functionalities are disrupted. PertInInt's analysis demonstrates that somatic mutations are frequently enriched in interaction sites and domains and implicates interaction perturbation as a pervasive cancer-driving event.

INTRODUCTION

Large-scale, concerted oncogenomic consortia, such as the Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), have sequenced an unprecedented number of tumor genomes from thousands of patients across tens of cancer types (International Cancer Genome Consortium et al., 2010; TCGA Research Network et al., 2013). Computational analyses of these datasets promise a revolution in precision oncology with additional insights into the genetic underpinnings of a staggeringly complex and heterogeneous disease (Chin and Gray, 2008). The recent, successful completion of these efforts, heralded as the "end of the beginning" of cancer genomics, has revealed a critical need for new methods that are able both to detect less frequent cancer-driving mutational events as well as to suggest the mechanistic, molecular impact of these mutations (Bailey et al., 2018; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020). More broadly, the comprehensive detection of cancer-driving mutational events, coupled with an understanding of their biological mechanism of action, has the potential to expand our knowledge of altered cellular processes in tumors, to reveal actionable, genetic similarities between different cancer types, to inform how evolving, heterogeneous populations of tumor cells may impact therapeutic efficacy, and to further translational research and inform downstream clinical treatments (McGranahan and Swanton, 2017; Vogelstein et al., 2013).

Despite considerable efforts, the crucial first step toward these goals—differentiating the small fraction of somatic mutations with functional roles in cancer ("drivers") from the preponderance of neutral "passenger" mutations—still poses a substantial computational obstacle (Porta-Pardo et al., 2017). While initial attempts to uncover cancer drivers at the gene level based on frequency of mutation across tumor samples have been fruitful (Dees et al., 2012; Lawrence et al., 2013), such gene-centric, recurrence-based approaches are inherently unable to detect infrequently mutated driver genes and also cannot distinguish among mutations within the same gene that may lead to distinct tumor phenotypes or clinical responses (Torkamani and Schork, 2008). Indeed, different positions within genes can contribute in varying degrees to different molecular functionalities; mutations falling within different gene positions can therefore have unequal impacts. In order to address the critical need to detect and interpret rare mutational driver events, an emerging class of "subgene" level approaches consider somatic mutations affecting genes within the context of information known about specific sites within their encoded proteins (Porta-Pardo et al., 2017). Existing subgene-level methods have derived such protein site functionality information from analyses of evolutionary conservation (Adzhubei et al., 2010; Ng and Henikoff, 2003; Reva et al., 2011), three-dimensional structure (Gao et al., 2017; Kamburov et al., 2015; Niu et al., 2016; Porta-Pardo et al., 2015; Ryslik et al., 2014a; Tokheim et al., 2016a), domains (Munro et al., 2018; Peterson et al., 2017; Porta-Pardo and Godzik, 2014), or post-translational modification (Reimand and Bader, 2013; Zhao et al., 2017). These methods, however, tend to identify cancer genes by considering whether somatic mutations alter just a single type of functionality (e.g., determining whether mutations are enriched only within protein domains [Porta-Pardo and Godzik, 2014] or only within phosphorylation sites [Reimand and Bader, 2013]), whereas somatic mutations within putative driver genes have been found to disrupt a broad range of protein functionalities. On the other hand, machine learning approaches to classify cancer drivers incorporate multiple types of information, but due to their ''black box'' nature, mechanistic interpretations of their predictions are not possible (Carter et al., 2009; Shihab et al., 2013).

We and others have previously demonstrated that detecting proteins that harbor somatic mutations in their interaction interfaces is a particularly effective approach to pinpoint infrequent driver mutations as well as reason about their molecular impacts and therapeutic sensitivities (Engin et al., 2016; Ghersi and Singh, 2014; Gress et al., 2016; Kamburov et al., 2015; Kar et al., 2009; Nishi et al., 2013; Porta-Pardo et al., 2015; Stehr et al., 2011). Indeed, several cancer driver genes, including *TP53* and *IDH1*, are well known to harbor mutations within their interaction sites (Nishi et al., 2013). While traditionally interaction sites have been identified directly for the small fraction of human genes with actual or modeled cocomplex structures, we have recently developed a domain-based approach that accurately detects residues that interact with DNA, RNA, peptides, ions, or small molecules across 63% of human genes (Kobren and Singh, 2019). A robust computational framework that utilizes this vastly expanded knowledge base about interaction sites and explicitly integrates it with additional lines of evidence regarding subgene functionality would provide a powerful new approach not only to detect but also to interpret a wide range of mutations driving protein dysfunction in cancer.

Here, we introduce a fast, interpretable, and easily extendable framework that enables us to uncover whether somatic mutations within genes are enriched in sites associated with high measures of ''functionality'' as determined by multiple, possibly correlated, lines of evidence. Our implementation PertInInt (pronounced ''pertinent,'' perturbed in interactions) incorporates interaction site information, along with evolutionary conservation and domain membership information, as each of these measures informs which sites are important for protein functioning. We derive analytical calculations that obviate the need to perform time-prohibitive permutation-based significance tests, thereby making it feasible to integrate, in a principled manner, these distinct measures of subgene-level functionality. Further, we extend our framework to consider whole-gene mutation rates, as genes that are recurrently mutated across tumors are often found to be causally implicated in cancers (Forbes et al., 2010). While other approaches have combined the output of multiple programs post hoc (e.g., Bailey et al., 2018), PertInInt integrates multiple alternate sources of subgene resolution data with whole-gene mutational frequency within a single unifying framework in order to detect, evaluate, and infer the molecular impact of patterns of somatic mutations within all human genes.

We apply PertInInt to somatic missense mutation data arising from 10,037 tumor samples across 33 cancer types to identify genes with the most enriched mutational patterns. We find that while each source of information—interaction, domain, evolutionary conservation, and whole-gene mutation frequency—is individually predictive of cancer genes, PertInInt uncovers more comprehensive sets of cancer-relevant genes when considering all sources of information together. We demonstrate that PertInInt is able to identify even those cancer genes with relatively low overall mutation rates, and that PertInInt readily outperforms previous methods while revealing whether and what type of interaction potential is perturbed. PertInInt finds that numerous known oncogenes and tumor suppressors have an enrichment of somatic mutations within their interaction interfaces and, in addition, predicts new cancer-relevant genes along with their altered interaction functionalities. Altogether, PertInInt provides a highly effective integrative framework to analyze large-scale cancer somatic mutation data and further our understanding of the molecular mechanisms driving cancers.

RESULTS

Overview of the PertInInt Framework

PertInInt aggregates somatic mutational data observed across tumor samples and identifies for each gene whether certain types of its functional sites are enriched in somatic mutations and/or whether the gene exhibits a high mutation rate across its length. We briefly overview our approach next (see also Figure 1); more details can be found in the STAR Methods section and Figure S1.

Different measures of protein site functionality are modeled using distinct ''tracks'' where each position within a track has a corresponding 0 to 1 weight that reflects its importance with respect to the functionality being considered (Figure 1A). Though any type of annotated functional region can be incorporated into our framework, here, we consider four specific types of tracks. First, "interaction tracks" model various protein–ligand interaction interfaces, where higher positional weights indicate that those positions are more likely to participate in interactions with a ligand; each interaction track corresponds to the subset of protein positions where we have any knowledge about ligand-binding potential (Figures S1A–S1C). Second, "domain tracks" span the length of the protein and simply identify portions of the protein sequence that correspond to the domain of interest; weights are 1 for amino acid positions within the domain and 0 elsewhere (Figure S1D). Third, the "conservation track" is also the length of the protein sequence, and the weight of each position measures its conservation across vertebrate homologs; higher weights correspond to positions under more evolutionary constraint (Figure S1E). Finally, to determine whether a gene as a whole has more mutations than expected, we extend our framework to incorporate the "natural variation track," which has a single entry per gene that reflects its background mutation rate, as estimated from the number of variants this gene has across healthy populations (Lek et al., 2016; Przytycki and Singh, 2017) (Figure S1F). Approximately 63% and 90% of human genes have per-site information about interactions or domains respectively, while all genes have per-site conservation values and background gene-level mutation rates. A gene may have

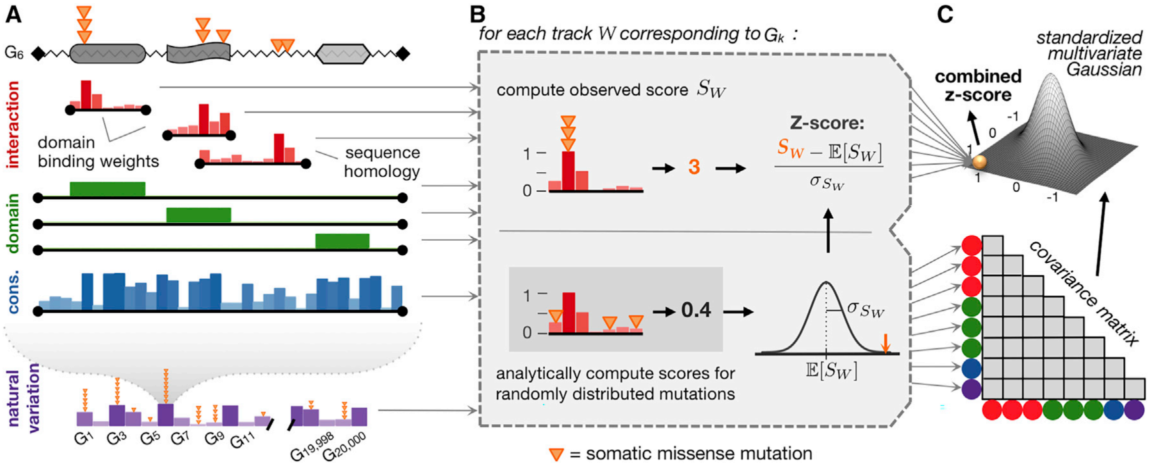

A
G6
interaction
domain binding weights
sequence homology
domain
cons.
natural variation
G1
G3
G5
G7
G9
G11
G19,998
G20,000
B
for each track W corresponding to G_k :
compute observed score S_W
1
0
3
Z-score:
$\frac{S_W - \mathbb{E}[S_W]}{\sigma_{S_W}}$
1
0
0.4
σ_{S_W}
$\mathbb{E}[S_W]$
analytically compute scores for randomly distributed mutations
C
standardized multivariate Gaussian
combined z-score
1
0
-1
1
0
-1
covariance matrix
▽ = somatic missense mutation

Figure 1. PertInInt Uncovers Cancer Driver Genes by Integrating Per-Site Interaction, Domain, and Conservation Information with Whole-Gene Mutation Frequency Data

(A) Somatic mutations (orange triangles) found across sequenced tumors that affect a protein sequence (jagged line) with three domains (gray regions) are evaluated with respect to different measures of functionality, each represented as a "track." In interaction tracks (red), positions that are more likely to participate in ligand interactions have higher weights (vertical bars). Interaction tracks arise from domain-based binding potential calculations (Kobren and Singh, 2019) (top two red tracks, each covering the length of the respective domain) or homology modeling (Ghersi and Singh, 2014) (bottom red track, covering the length of the modeled region). Domain tracks (green) specify which residues within a protein are part of a specific domain by 0/1 positional weights; here we have a track for each domain within the sequence. The conservation track (blue) weights each position by its evolutionary conservation across species. The natural variation track (purple) models how much each gene varies across healthy populations; here the height of the vertical bars indicates the background mutation probability rather than a per-gene weight, which is 1 for the gene being considered and 0 otherwise. Figure S1 gives further intuition about how these track weights are determined.

(B) For each track W, we compute the score S_W of the observed somatic mutations as the sum of the track weights for the positions where they appear (top). To determine whether this score is higher than expected, we consider a model where somatic mutations are shuffled across the positions of the track, and the expected score ($\mathbb{E}[S_W]$) and the standard deviation of the scores (σ_{S_W}) are computed and used to estimate per-track Z scores (bottom); note that in our framework these values are computed analytically instead of relying on the shuffles.

(C) Z scores for all tracks are combined after analytically determining a background covariance model.

numerous interaction and domain tracks (e.g., for different modeled interaction regions and for each of its identified domains) but has only a single conservation and natural variation track.

For each track, we consider the somatic mutations observed across tumor samples that fall within positions of that track and compute a per-track score as the sum of the per-track weights of the positions that each of the mutations fall into (Figure 1B); intuitively, a high score corresponds to the case where a large number of mutations fall within important track positions. To determine whether the score for a track is more than expected by chance, we could shuffle the mutations across the positions of the track, and use the mean and standard deviation computed from these permutations to compute a Z score; however, the mean and standard deviation for each track can be computed analytically. For each protein, we next combine the information from each of its tracks. Because tracks can overlap along the length of the protein sequence, and the somatic mutations that fall in each of them can also overlap, these tracks cannot be treated independently. Instead, for the background model we derive an approach to compute the covariance between tracks analytically and then use this covariance matrix to estimate a combined score (Figure 1C; see STAR Methods for derivations). We find that even when considering just a single track, our analytical formulation leads to >7× speedup over *each* empirical permutation (Figure S2). In practice, numerous shuffles are necessary to compute the mean and variance for a single track, and empirical calculations to estimate the covariance across all tracks is prohibitively slow, highlighting the power and necessity of our analytical formulation.

The final per-gene score output by PertInInt considers whether somatic mutations across samples are enriched (1) in positions with high ligand-binding potential for an interaction track, (2) within domain positions for a domain track, (3) within conserved sites for the conservation track, and (4) within the gene overall.

PertInInt Effectively Identifies Cancer Driver Genes via Integrating Multiple Sources of Information

We run PertInInt on somatic point mutation data aggregated across 10,037 pan-cancer tumor samples and 33 tumor types from TCGA (TCGA Research Network et al., 2013) (Figure S3 and Table S1). PertInInt's analytical formulation enables the simultaneous consideration of multiple types of biological data regarding protein functionality. However, to first uncover to what extent each source of information—per-site interaction, domain, and conservation information as well as overall gene mutational frequency—is independently useful for identifying cancer-relevant genes, we run PertInInt on the pan-cancer dataset when restricted to each of these track types in turn. To validate the method in the absence of a complete gold standard, as we consider an increasing number of output genes, we compute how "enriched" this set is in genes from the Cancer Gene Census (CGC), a curated list of genes implicated in cancer (Futreal et al., 2004). In particular, enrichment is computed as the

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005

ratio between the fraction of CGC genes in the set of top-scoring genes considered (i.e., the precision) and the fraction of CGC genes in the whole set of genes (i.e., the precision you would expect to achieve if genes are randomly ordered).

We find that utilizing subsets of only interaction, only domain, only conservation, or only natural variation tracks in turn can recapitulate known CGC genes to varying degrees, with interaction tracks identifying the largest number of known driver genes while maintaining perfect precision relative to other track subsets (Figure 2A). Our integrative framework that incorporates all track types outperforms every version of our algorithm that considers only subsets of information; indeed, considering any two sources of biological information outperforms versions of PertInInt that utilize only one source, and considering any three sources of data tends to improve performance even further (Figure 2B). Attempts to combine per-track scores without accounting for between-track covariance (e.g., by summing or averaging track *Z* scores) not only are incorrect but also perform considerably worse in detecting cancer-relevant genes (Figure S4). Altogether, these results demonstrate the ability of our approach to effectively leverage the distinct contributions of multiple, complementary data sources regarding protein position and whole-gene functionality in order to uncover cancer driver genes. Furthermore, the enrichment of cancer genes among PertInInt's top predictions remains when considering different gold standards (Figure S5).

We find low overlap between the sets of CGC genes identified when utilizing distinct track types, indicating that mutations within cancer genes tend to target a diverse array of functional elements (Figure 2C). Only a small minority of CGC genes (less than 10%) are identified by all four track types within the top 200 ranked genes. Mutations falling into known tumor suppressor *PTEN*, for instance, tend to hit evolutionarily conserved protein positions but do not alter known inferred interaction interfaces or domain regions more than expected by chance. In contrast, a small molecule-binding pocket in the *IDH2* oncogene is recurrently mutated across cancers, and, thus, it is readily detected using interaction tracks alone but is less significantly ranked when PertInInt is restricted to other functionality data.

Lowly Mutated Genes Harbor Mutations that Preferentially Alter Functional Sites

We next show that PertInInt's integrative approach can highlight genes with preferentially altered functional sites that may be lowly mutated overall; such "long tail" driver genes are easily missed by traditional frequency-based driver gene detection approaches. When run on the pan-cancer dataset utilizing all track types, PertInInt ranks highly several such infrequently mutated genes (Figure 3A). Of the top 35 genes ranked by PertInInt on the pan-cancer dataset, we find that 20 have a missense mutation rate less than one-twentieth of the maximum observed mutation rate (Figure 3B). These high-scoring long tail genes include novel genes with potential roles in cancer as well as known driver genes that cannot have been identified based solely on their relative mutation frequency (e.g., *KMT2D* and *CIC*, Figure 3B). Of the 20 highly ranked infrequently mutated genes, 18 harbor perturbed interaction sites, enabling immediate molecular insights regarding their roles in cancer. For example, among long tail genes that are highly ranked by PertInInt but have not yet been identified as cancer relevant, several have an enrichment of mutations in their DNA or small molecule interaction sites (e.g., *MGA* and *GRIN2D*, Figure 3B), in line with previous observations that many cancer driver genes exhibit these types of protein interaction perturbations (Delgado and León, 2006; Jeggo et al., 2016; Raimondi et al., 2017).

Mutations Are Distributed across Interaction Interfaces

For each protein with a significantly perturbed interaction interface, we next sought to determine whether mutations are found within a small number of interaction sites or across several interaction sites. We consider all sites within the protein with nonzero interaction track weights and use the frequency with which somatic mutations occur within each of them to compute a normalized Shannon entropy (Shannon, 1948). Higher entropies correspond to proteins with mutations spread across many interaction sites, whereas low entropies correspond to mutational patterns that can be uncovered by methods that look for mutation "hotspots" (Chang et al., 2016). As expected, PertInInt highly ranks several oncogenes that have previously been detected by hotspot detection algorithms due to their recurrent mutations in critical interaction positions (e.g., *IDH1*, *BRAF*, and *NRAS*). However, there are also many genes with significantly perturbed interaction interfaces where mutations are spread more widely across their interaction sites (Figure 3C). Known cancer genes *DICER1*, *SMARCA4*, *CREBBP*, and *KMT2D*, for instance, are among the top 35 genes ranked by PertInInt and contain significantly mutated interaction sites (combined score across interaction tracks > 6), each with several interaction sites that together harbor an enriched number of somatic mutations.

Notably, this analysis reveals that the top-ranked genes with significantly perturbed interaction interfaces include both oncogenes and tumor suppressor genes (TSGs), reflecting a dichotomy in the impact of binding interface mutations. Whereas some specific mutations within interaction sites have been linked to oncogenic activity (Stehr et al., 2011), other binding site mutations are known to entirely disrupt critical interactions and overall protein function (Cho et al., 1994). Although we model the interaction sites of similar numbers of oncogenes and TSGs (238 and 246 respectively), we find that among the 50 genes with the highest enrichment of mutations within their interaction sites, the enrichment of oncogenes is 2.36-fold greater than the enrichment of TSGs. Nevertheless, PertInInt uncovers perturbed interaction interfaces in many genes that have been previously identified as drivers due to nonsense, frameshift, or other relatively disruptive mutations typically associated with TSGs (e.g., *RUNX1* and *FOXO1*). Indeed, enriched yet less common interaction-altering missense mutations uncovered by PertInInt may correspond to more subtle knockdown phenotypes or previously underappreciated oncogenic activities of genes traditionally characterized as TSGs.

PertInInt Outperforms Previous Methods in Detecting Cancer Genes

Having demonstrated that PertInInt can identify interaction interfaces enriched in mutations across tumor samples, and that this is highly predictive of cancer genes, we next turn to assessing PertInInt's performance as compared with previously published methods (Chang et al.,2016; Lawrence et al., 2014; Melloni et al.,

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005

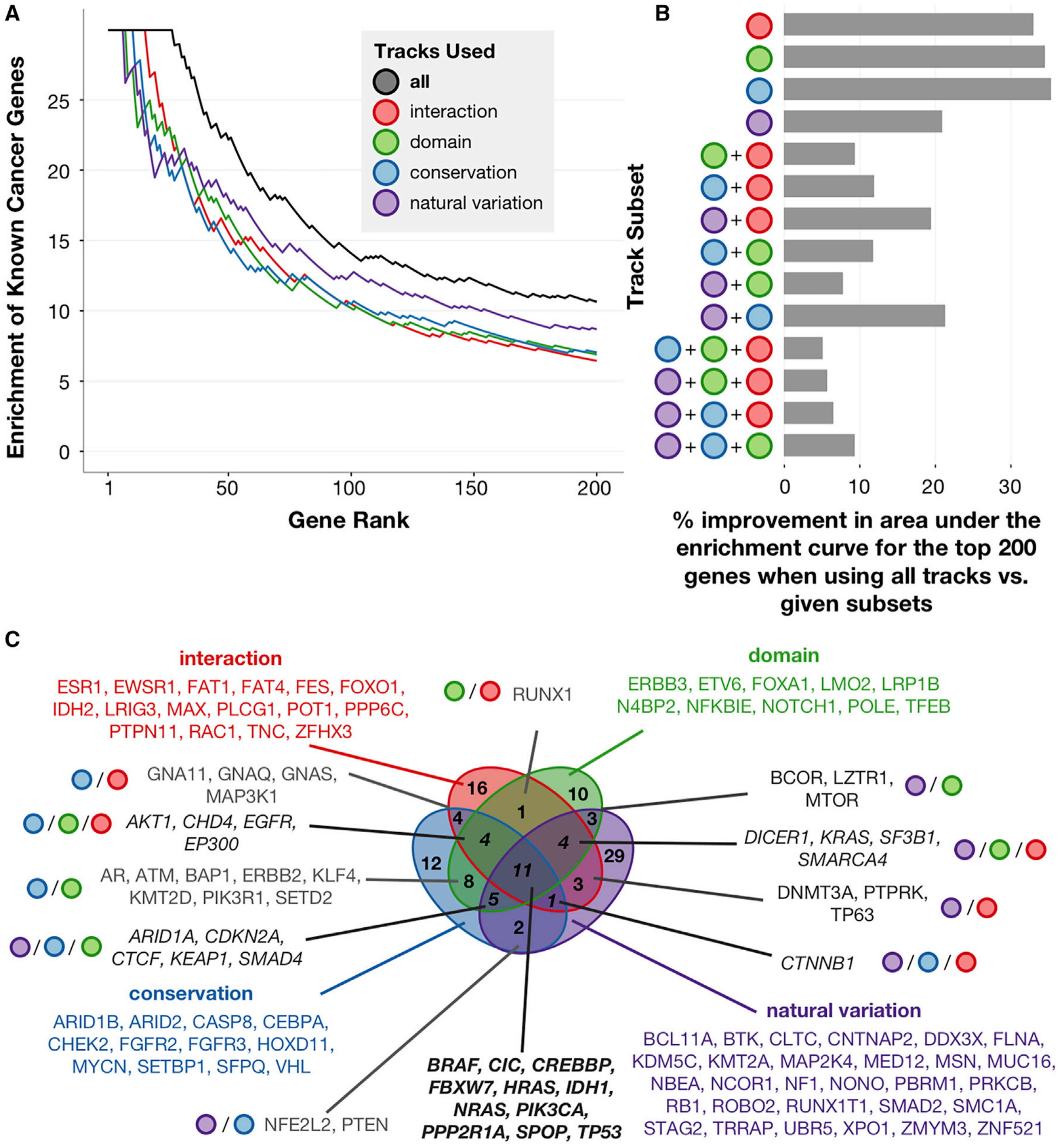


Figure 2. PertInInt Is Highly Effective in Uncovering Cancer Driver Genes Due to Combining Multiple Sources of Information

(A) Enrichment of CGC genes (y axis) within a given number of top-scoring genes (x axis) when run on the pan-cancer dataset using all tracks together (black), only interaction tracks (red), only domain tracks (green), only the conservation track (blue), and only the natural variation track (purple). Enrichment is computed as the ratio between the fraction of CGC genes in the set of top scoring genes considered (i.e., the precision) and the fraction of CGC genes in the whole set of genes (~0.0334). While uncovering genes enriched for somatic mutations within only interaction sites, only domain positions, only conserved sites, or only over their lengths each yields cancer-relevant genes, performance is highest when PertInInt uses all sources of information together.

(B) Percent improvement in the area under the enrichment curve for the top 200 genes when using all track types versus specific subsets of tracks. PertInInt is more effective in uncovering CGC genes when using all sources of information together than when using any other of the possible subsets of information.

(C) Venn diagram showing the overlap of CGC genes detected in the top 200 genes ranked when considering only interaction, only domain, only conservation, or only natural variation tracks. The different sources of information yield distinct yet overlapping sets of cancer genes.

2016; Mularoni et al., 2016; Porta-Pardo et al., 2015; Porta-Pardo and Godzik, 2014; Przytycki and Singh, 2017; Ryslik et al., 2013, 2014b; Tamborero et al., 2013; Ye et al., 2010) for detecting cancer driver genes (see STAR Methods). These methods differ substantially in terms of their statistical models and overall goals and, unlike PertInInt, are largely unable to distinguish among the various types of interaction and other functional perturbations affecting the identified genes.

When applied to tumor samples from the pan-cancer dataset, our method has a greater enrichment for CGC genes than the

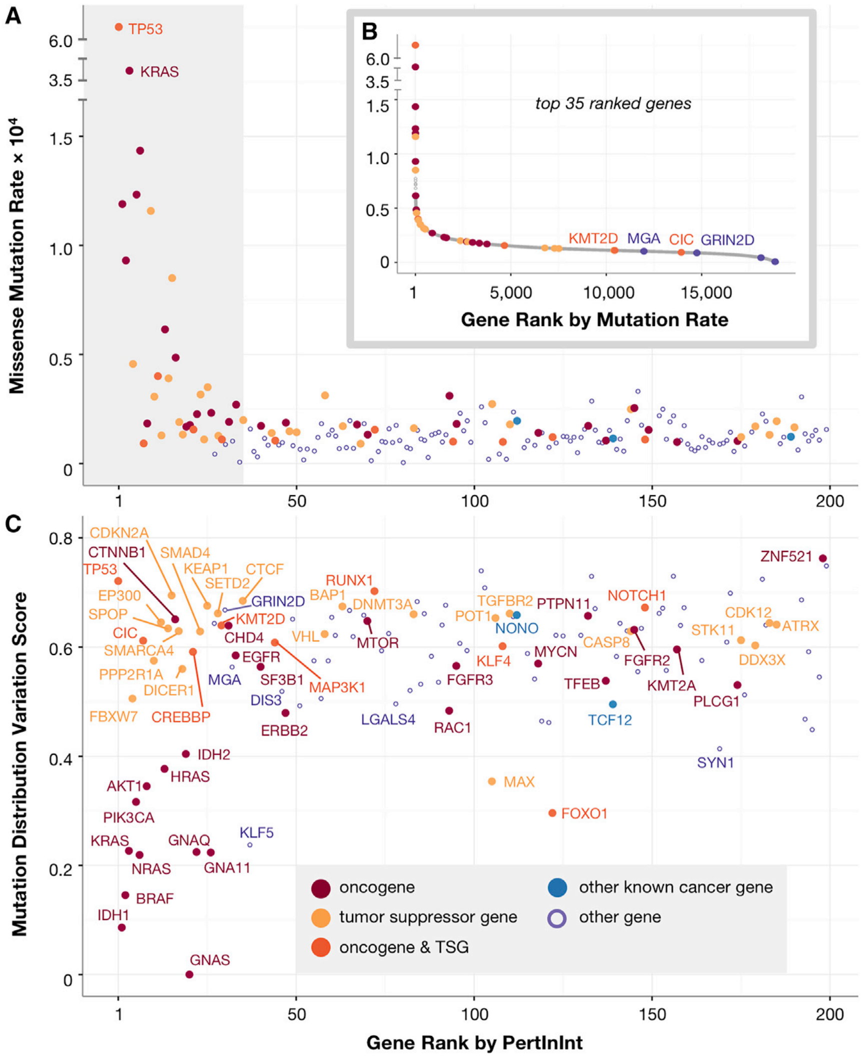


Figure 3. Perturbed Interaction Interfaces across Oncogenes and Tumor Suppressors

(A) Shown are the missense mutation rates (y axis) of the top 200 genes ranked by PertInInt (x axis). Top-ranked genes are both highly and infrequently mutated. Genes are colored as in (C). The shaded gray box highlights the plot to 35 genes, which are featured in the part (B) inset.

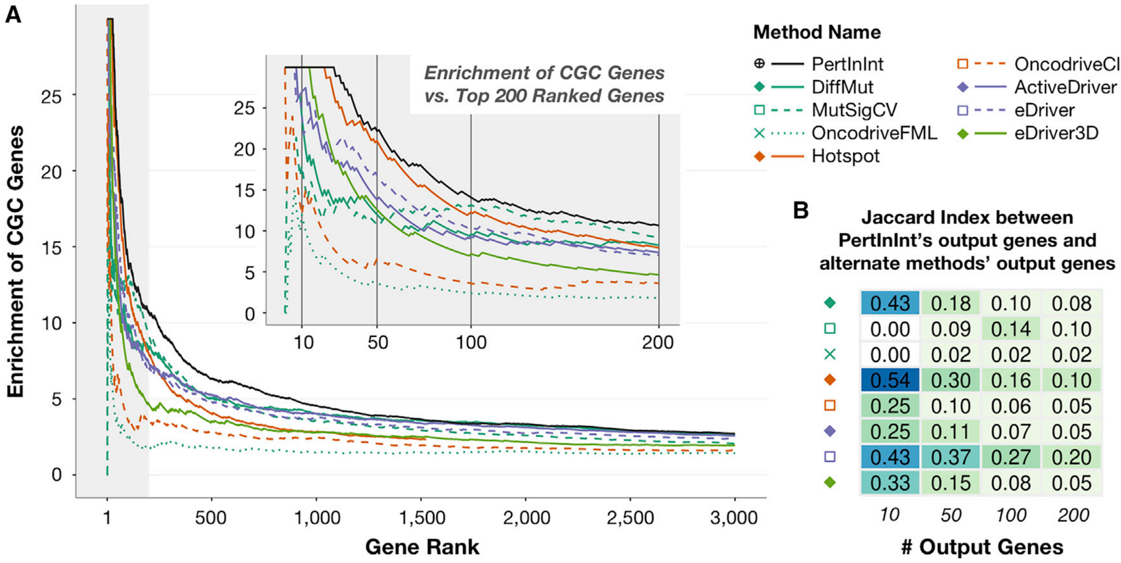
(B) Genes are ordered by their per-tumor-sample missense mutation rate in the pan-cancer dataset (x axis), and their missense mutation rate is given (y axis). PertInInt's top 35-ranked genes are plotted in color and exhibit a wide range of ranks with respect to mutation rate. Of these, only genes with below-median overall mutation rates and a Z score ≥ 1 in at least one interaction track are labeled.

(C) For each of the top 200 genes ranked by PertInInt (x axis), for those with a Z score ≥ 1 in at least one interaction track, we also analyze the distribution of somatic mutations across interaction sites and compute their normalized Shannon entropy (y axis). These genes contain recurrent (low variation) as well as more distributed (high variation) mutations across their binding interfaces.

other tested methods that we were able to run (Figure 4A). PertInInt also outperforms these other methods in terms of enrichment of CGC genes among top-ranked genes even after we exclude tumor samples from the six most highly mutated cancer types with 100+ missense mutations per patient on average, demonstrating that PertInInt's superior pan-cancer performance is not driven by samples from cancer types that contribute large numbers of mutations (Figure S6). PertInInt's greater enrichment as compared with other methods is observed as well for other lists of driver genes, and, furthermore, neither PertInInt nor any of the other tested methods show any enrichment for sets of genes that have been suggested to be unlikely to play roles in cancer (Figure S7; Table S2). Notably, the genes ranked highly by PertInInt differ substantially from those identified by other approaches (Figure 4B). Specifically, the set of genes identified by PertInInt has a consistently low Jaccard index (JI) with sets of genes ranked by alternate methods (JI < 0.5 across all methods for top 25 genes, JI < 0.25 across all methods for top 150 genes). Moreover, due to our analytical formulation, PertInInt can process the pan-cancer mutational data while considering multiple sources of data about protein functionality in 10 min on a single core of standard desktop; alternate methods each consider a limited set of mutational patterns and range in runtime from minutes to days (Table S3).

We also repeat our analysis on datasets restricted to samples from one cancer type, as many alternate methods that failed to run on the pan-cancer dataset are able to run on these substantially smaller subsets of tumor genomes. We find that in general across individual cancer datasets, PertInInt tends to achieve a higher area under the enrichment curve than other methods, including whole-gene methods, and a version of PertInInt that includes only subgene resolution tracks also outperforms other subgene methods (Figure S8). We note that since individual cancer types obviously have smaller total numbers of somatic mutations as compared with the combined pan-cancer dataset, the Z scores computed by PertInInt when run on individual cancers tend to be smaller than PertInInt's pan-cancer Z scores; this trend is especially notable for cancer types with fewer samples and/or lower mutation rates. Similarly, fewer tracks can be evaluated for significance in the per-cancer analysis; this has a larger effect on interaction tracks in particular as they tend to involve fewer protein positions. Indeed, we find that the relative proportions of track types with positive Z scores for each cancer type are not notably different from each other, with the exception that the highly mutated cancer types (i.e., colorectal, lung, and stomach adenocarcinomas; skin cutaneous melanoma; and uterine corpus endometrial carcinoma, see Figure S3) each have a greater number of positively scoring domain and interaction tracks relative to other cancer types. Despite these differences when run on different cancer types, PertInInt is able to readily recover "cancer-specific" drivers (e.g., *EIF1AX* in uveal melanoma), and our

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005



	10	50	100	200
◆	0.43	0.18	0.10	0.08
□	0.00	0.09	0.14	0.10
×	0.00	0.02	0.02	0.02
◆	0.54	0.30	0.16	0.10
□	0.25	0.10	0.06	0.05
◆	0.25	0.11	0.07	0.05
□	0.43	0.37	0.27	0.20
◆	0.33	0.15	0.08	0.05

Figure 4. Detection of Known Cancer Genes from a Pan-Cancer Dataset by PertInInt and Alternate Methods

Each driver gene detection method was run on the pan-cancer set of missense mutations.

(A) Curves indicate the enrichment of CGC genes (y axis) as we consider an increasing number of output genes (x axis) for each driver gene detection method. Enrichment is computed as the ratio between the fraction of CGC genes in the set of top-scoring genes considered (i.e., the precision) and the fraction of CGC genes in the whole set of mutated genes (~0.0334). All methods scored at least 3,000 genes except for Hotspot (orange solid line), which only returned 1,530 genes and whose curve ends at that point. The gray shaded area highlights the plot to 200 genes, a closeup of which is shown in the inset. Vertical lines at 10, 50, 100, and 200 ranked genes in the inset correspond to gene set sizes featured in part (B).

(B) JIs are calculated between the top 10, 50, 100, and 200 genes output by PertInInt and the corresponding top 10, 50, 100, and 200 genes output by each other method. Lighter colors indicate lower JIs and less overlap between the gene sets.

analysis also reveals that genes that play dominant roles in certain cancer types may be important for smaller proportions of tumor samples in alternate cancer types as well (e.g., the same gene in thyroid carcinoma, Figure S9). Overall, our results show that PertInInt is a powerful method for evaluating mutational patterns across tumors of the same cancer type as well as across a pan-cancer dataset covering over 10,000 tumor genomes.

Distinct Perturbed Molecular Mechanisms Uncovered across Genes

Having shown that PertInInt is highly effective in identifying cancer genes, we next demonstrate its additional power to pinpoint which specific functional regions and mechanisms are perturbed by analyzing each track separately and determining which have positive Z scores (Table S4). Altogether, we find that 665 CGC genes have at least one subgene functionality track with a Z score $\geq$ 0.5, representing functional coverage of 93% of all CGC genes (Figure 5). Specifically, we find that DNA, RNA, peptide, ion, and small molecule interaction sites are enriched in mutations in 16%, 5%, 19%, 14%, and 22% of CGC genes, respectively; these numbers go up to 23%, 5%, 27%, and 24% of CGC genes if including those that are more broadly enriched in mutations across, respectively, DNA-binding, RNA-binding, peptide-binding, or metabolite-binding domains (as categorized in Pfam2Go; Mitchell et al., 2015). Up to 77% of CGC genes are enriched in mutations across at least one domain or interaction interface. We note that the perturbed nucleic acid and small molecule domains or binding sites found across 45% of cancer genes would not be readily identified by analyses that focus exclusively on protein–protein interaction alterations (Porta-Pardo et al., 2015).

We now highlight a few of these genes that, though not present in the CGC, were uncovered by PertInInt as having significantly mutated interaction interfaces. For instance, transcription factors *MGA* and *KLF5* harbor mutations within their basic helix-loop-helix and C2H2-ZF domains, respectively, that alter their DNA base-binding positions (Figure 6A), suggesting cancer-specific changes to normal DNA-binding and downstream regulatory activity. Indeed, *KLF5*'s E419Q mutation has recently been experimentally shown to change wild-type binding preferences and increase the expression of tumor progression genes *in vivo* (Zhang et al., 2018). Similarly, *MGA* normally subdues the activity of well-known oncogene *MYC*; its frequent deletion, truncation, or mutated binding properties across cancers further indicates its role as a tumor suppressor (Schaub et al., 2018). We also find that two RNA-binding genes *DIS3* and *SF1* exhibit significant mutations in their putative RNA-binding sites, with recurrent mutations in *DIS3* altering multiple distinct RNA-contacting positions (Figure 6B). In support of our predictions, *DIS3* is recurrently mutated in blood and skin cancers and has been identified as a candidate oncogene in colorectal cancer (de Groen et al., 2014). *SF1* is recurrently mutated across cancers in a mutually exclusive fashion (i.e., indicating its analogous functionality) to *RBM10*, a gene found to drive aberrant splicing events in cancer (Seiler et al., 2018).

PertInInt also newly implicates a number of genes—present neither in the CGC nor on other lists of known cancer genes (Bailey et al., 2018; Kandoth et al., 2013; Lawrence et al., 2014;

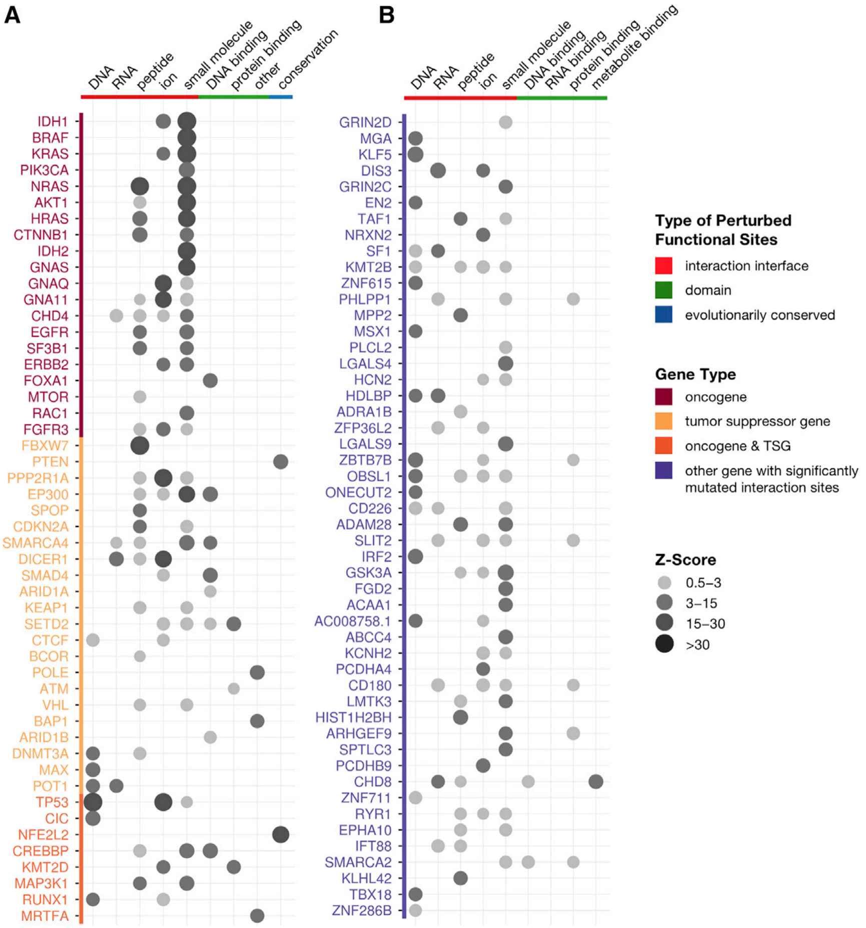


Figure 5. Perturbed Mechanisms across Oncogenes, Tumor Suppressor Genes, and Putative Cancer Genes

Gene names are colored by driver status; genes that are not yet known to be cancer drivers but have a Z score $\geq$ 0.5 in one or more interaction tracks are in lavender. For each gene, the circles indicate the Z scores for enrichment of mutations in particular types of tracks, with interaction tracks in red, domain tracks in green, and the conservation track in blue. Z scores for mutational enrichments in domain tracks are shown only if the Z scores for the corresponding interaction tracks are < 0.5. Z scores for the conservation track are shown only if Z scores for all other track types are < 0.5.

(A and B) (A) PertInInt's top 50 ranked known cancer driver genes and (B) top 50 ranked putative cancer driver genes with a significantly mutated interaction track exhibit a wide range of perturbed functionalities.

Vogelstein et al., 2013)—with mutations that appear to alter critical small molecule-binding positions (Figure 6C). The highly conserved kinase *GSK3A* for instance harbors a significant enrichment of mutations altering its ATP-binding positions. Supporting our prediction, suppression of this gene is associated with impaired growth and induction of apoptosis and it has recently been proposed as a potential therapeutic target in acute myeloid leukemia (Banerji et al., 2012; McCubrey et al., 2014). We also find that the S-type lectin *LGALS4* has an enrichment of mutations altering the β-galactoside sugar-binding positions in its galectin domains; indeed, *LGALS4* has been linked to the regulation of the cancer-relevant Wnt signaling pathway and has been experimentally implicated as a tumor suppressor in colorectal cancer cells *in vitro* (Satelli et al., 2011).

To more broadly characterize PertInInt's novel cancer gene predictions, we use gene set enrichment analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005) with Reactome pathways (Jassal et al., 2020) on the ranked list of non-CGC genes output by PertInInt with combined score $\geq$ 1 (Table S5). GSEA uncovers 12 enriched Reactome pathways at false discovery rate (FDR) corrected p value (or q-value) < 0.2, including known cancer pathways such as *Ras* (q-value = 0.102), and signaling pathways mediated by known cancer genes, such as *CREB1* (q-value = 0.042), *NOTCH3* (q-value = 0.071), and *PDGF* (q-value = 0.195). In addition, a pathway related to cell adhesion, which is known to be disrupted especially in metastatic tumors, is also enriched (e.g., L1-ankyrin interactions, q-value = 0.078). Intriguingly, two enriched pathways suggest that nervous system-related functionalities may play roles across tumors (q-value < 0.2), an observation that has only very recently been explored (Zeng et al., 2019).

DISCUSSION

In this work, we have introduced a fast, integrative framework to detect cancer driver genes by uncovering whether somatic mutations across tumors are enriched in sites of different types of functionalities. Our method utilizing this framework, PertInInt, integrates knowledge from the largest set of protein–ligand interaction sites to date (Ghersi and Singh, 2014; Kobren and Singh, 2019) with additional biological data regarding subgene functionality and whole gene mutability (Figures 1 and S1). When applied to over 10,000 tumor samples from 33 cancer types, PertInInt reveals a broad range of perturbed functionalities in several known driver genes as well as in relatively rarely mutated genes with predicted tumorigenic roles (Figures 3A, 3B, and 5). Notably, PertInInt finds that mutations within many known driver genes are enriched in protein interaction interfaces (Figure 2A) and more broadly implicates interaction perturbation as a frequent phenomenon in cancer cells (Figure 5).

Analyses of predicted cancer-relevant coding mutations often involve—whenever possible—assessing their putative effect with respect to protein structure (Bailey et al., 2018; Chang et al., 2016; Kamburov et al., 2015; Niknafs et al., 2013; Raimondi et al., 2017). Although using structure directly to identify relevant mutations is rarely scalable in terms of runtime and coverage (Ryslik et al., 2013), PertInInt's use of structurally predefined regions mediating protein interactions makes large-scale analyses in the context of protein structure feasible. Moreover, since cancer-driving genetic aberrations do not always involve mutation of protein–ligand interaction interfaces, a critical additional feature of PertInInt—that extends its coverage to all human genes—is

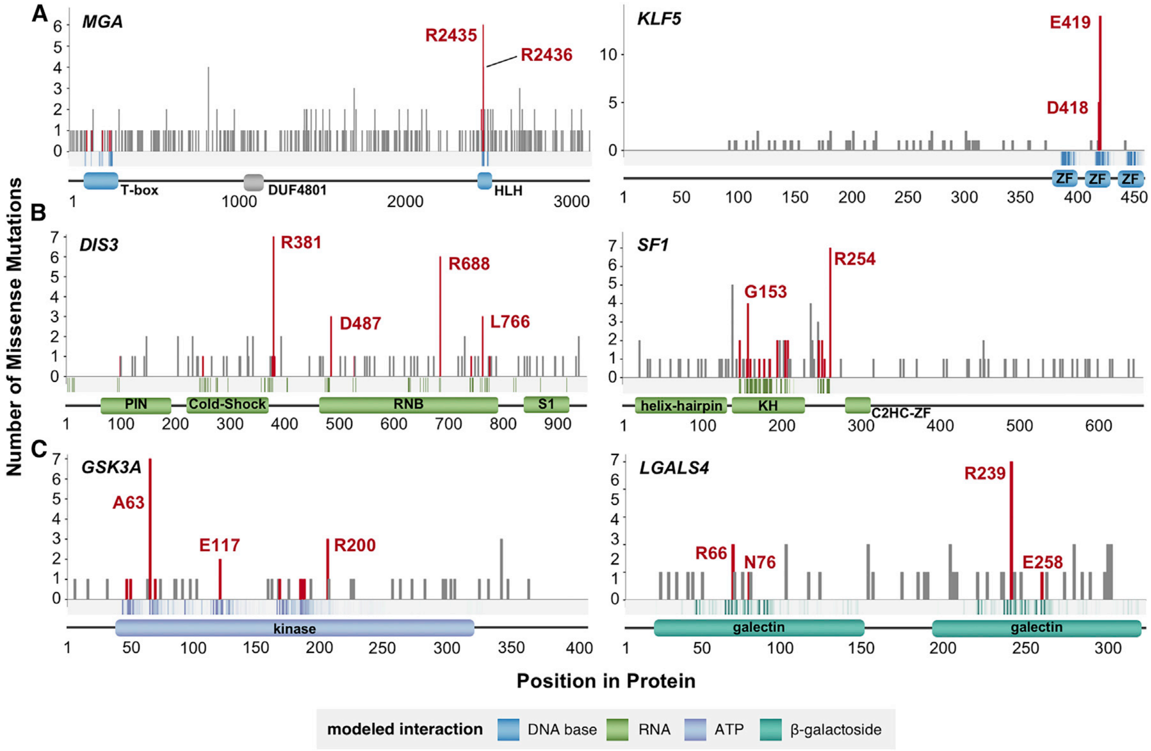


Figure 6. Examples of Genes Ranked Highly by PertInInt that Are Not Known to Be Drivers

Across the length of each gene (x axis), the number of missense mutations at each protein position is given (y axis). Vertical bars corresponding to mutations affecting binding sites are colored red. The band along the x axis depicts the likelihoods with which residues at each protein position are expected to interact with the specified ligand, with darker bars corresponding to higher (≥ 0.25) binding likelihoods. Domain locations and names are shown below.

(A) Putative cancer genes *MGA* and *KLF5* are enriched for mutations in DNA base-binding positions.

(B) Putative cancer genes *DIS3* and *SF1* are enriched for mutations in RNA-binding positions.

(C) Putative cancer genes *GSK3A* and *LGALS4* are enriched for mutations in small molecule- (ATP and β-galactoside sugar, respectively) binding positions.

that it seamlessly incorporates additional lines of evidence regarding protein site functionality. While here we have demonstrated that PertInInt effectively utilizes per-site evolutionary conservation and domain knowledge, we anticipate that encoding more sources of functional information within our framework (e.g., known phosphorylation sites or intrinsically disordered regions) will unearth other driver mutations and alternate mechanisms of action. Incorporating structurally resolved information from protein–protein interaction networks will also be a valuable direction for future work.

Genes that are frequently mutated across their lengths tend not to overlap genes that exhibit nonrandom patterns of mutations across individual protein positions, a pattern that has previously been leveraged to distinguish TSGs from oncogenes (Tokheim et al., 2016b; Vogelstein et al., 2013). By incorporating whole gene mutability information into our existing framework, we are able to uncover and profile a much more comprehensive set of both oncogenes and TSGs (Figures 2B and 3C). Although previous methods have also considered the frequency and spatial patterning of mutations within genes together (Korthauer and Kendziorski, 2015; Lawrence et al., 2013; Tokheim et al., 2016b), we also simultaneously infer specific perturbed molecular mechanisms within uncovered genes. We note that while mutation deleteriousness predictors—developed both in the context of cancer (Carter et al., 2009; Shihab et al., 2013) and otherwise (Adzhubei et al., 2010)—can evaluate the impact of somatic mutations, they tend to integrate multiple sources of protein site functionality information via complex statistical or machine learning approaches, where the contribution of each data source and thus subsequent mechanistic interpretations are obscured. In contrast, by determining mutational enrichments in specific types of functional sites, PertInInt is able not only to identify cancer-relevant genes but also to begin to explicitly reason about the biomolecular impacts of mutations. Indeed, uncovering the mechanisms of action for cancer-driving mutational events has been a major bottleneck in the critical step of translating this knowledge to improve patient care and outcomes (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium et al., 2020).

Given the success of large-scale cancer genome sequencing consortia projects in expanding our knowledge of basic cancer biology (Bailey et al., 2018; Ding et al., 2018; Hoadley et al., 2018; Sanchez-Vega et al., 2018), coupled with the decreased cost of genome sequencing, it is clear that sequencing tumor genomes will be routine practice in both basic science and clinical settings, thereby rapidly increasing the number of sequenced tumors available for analysis. Importantly, PertInInt's analytical framework enables it to efficiently process increasing numbers of tumor genomes; further, this speed is accompanied by better identification of cancer-relevant genes when run on larger numbers of tumor samples (Figure S10). Since PertInInt's underlying analytical framework is general, we anticipate that it will also be effective in other settings. For example, because very few non-coding somatic mutations in cancer tend to be recurrent (Khurana et al., 2016), it may be especially powerful for identifying regulatory regions with an enrichment of mutations within sites associated with different measures of functionality (e.g., binding sites for different proteins).

In the future, one of the most tantalizing prospects of cancer genomics is its potential in transforming clinical practice. While identifying and linking cancer mutations to personalized treatments remains a daunting challenge, PertInInt dramatically accelerates the detection of rare mutational driver events from sequenced tumors while providing important information about their mechanisms of action, a key step in developing and customizing targeted therapeutic regimens.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead Contact
  - ○ Materials Availability
  - ○ Data and Code Availability
- METHOD DETAILS
  - ○ Protein Site-Based Functional Tracks
  - ○ Per-track Functional Mutation Scores
  - ○ Per-track Analytical *Z* score Calculations
  - ○ Per-site Background Mutational Model
  - ○ Between-Track Analytical Covariance Calculation
  - ○ Whole Gene Mutability Tracks
  - ○ Combining Per-track *Z* scores for Each Protein
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Cancer Mutation Data Preparation
  - ○ Runtime Analysis
  - ○ Testing Related Driver Detection Methods

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2020.06.005.

ACKNOWLEDGMENTS

Thanks to members of the Singh lab for their insight and comments. Thanks especially to Dario Ghersi, Anton Persikov, and Pawel Przytycki for helpful discussions regarding PDB files and TCGA mutation data. The results published here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. This work was funded in part by NIH grant R01-CA208148 (to M.S.) and NSF GRFP grant DGE-1148900 (to S.N.K.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

S.N.K., B.C., and M.S. designed the study. S.N.K. performed the analysis and developed the software. S.N.K. and M.S. wrote the manuscript. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 2, 2019
Revised: February 23, 2020
Accepted: June 5, 2020
Published: July 14, 2020

REFERENCES

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. Cell *173*, 371–385.e18.

Banerji, V., Frumm, S.M., Ross, K.N., Li, L.S., Schinzel, A.C., Hahn, C.K., Kakoza, R.M., Chow, K.T., Ross, L., Alexe, G., et al. (2012). The intersection of genetic and chemical genomic screens identifies GSK-3α as a target in human acute myeloid leukemia. J. Clin. Invest. *122*, 935–947.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875–1882.

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. *69*, 6660–6667.

Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163.

Chin, L., and Gray, J.W. (2008). Translating insights from the cancer genome into clinical practice. Nature *452*, 553–563.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265*, 346–355.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.

de Groen, F.L., Krijgsman, O., Tijssen, M., Vriend, L.E., Ylstra, B., Hooijberg, E., Meijer, G.A., Steenbergen, R.D., and Carvalho, B. (2014). Gene-dosage dependent overexpression at the 13q amplicon identifies DIS3 as candidate oncogene in colorectal cancer progression. Genes Chromosomes Cancer *53*, 339–348.

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., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Delgado, M.D., and León, J. (2006). Gene expression regulation and cancer. Clin. Transl. Oncol. *8*, 780–787.

Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.L., Tokheim, C., et al. (2018). Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell *173*, 305–320.e10.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195.

Engin, H.B., Kreisberg, J.F., and Carter, H. (2016). Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. PLoS One *11*, e0152929.

Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A., and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. *17*, 178.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Forbes, S.A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., Kok, C.Y., Jia, M., Ewing, R., Menzies, A., et al. (2010). Cosmic (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. *38*, D652–D657.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gao, J., Chang, M.T., Johnsen, H.C., Gao, S.P., Sylvester, B.E., Sumer, S.O., Zhang, H., Solit, D.B., Taylor, B.S., Schultz, N., and Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. Genome Med. *9*, 4.

Ghersi, D., and Singh, M. (2014). Interaction-based discovery of functionally important genes in cancers. Nucleic Acids Res. *42*, e18.

Gress, A., Ramensky, V., Büch, J., Keller, A., and Kalinina, O.V. (2016). StructMAn: annotation of single-nucleotide polymorphisms in the structural context. Nucleic Acids Res. *44*, W463–W468.

Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M. (2016). Toward a shared vision for cancer genomic data. N. Engl. J. Med. *375*, 1109–1112.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell *173*, 291–304.e6.

International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature *464*, 993–998.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway KnowledgeBase. Nucleic Acids Res. *48*, D498–D503.

Jeggo, P.A., Pearl, L.H., and Carr, A.M. (2016). DNA repair, genome stability and cancer: a historical perspective. Nat. Rev. Cancer *16*, 35–42.

Kamburov, A., Lawrence, M.S., Polak, P., Leshchiner, I., Lage, K., Golub, T.R., Lander, E.S., and Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc. Natl. Acad. Sci. U S A *112*, E5486–E5495.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Kar, G., Gursoy, A., and Keskin, O. (2009). Human cancer protein-protein interaction network: a structural perspective. PLoS Comput. Biol. *5*, e1000601.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat. Rev. Genet. *17*, 93–108.

Kobren, S.N., and Singh, M. (2019). Systematic domain-based aggregation of protein structures highlights DNA-, RNA- and other ligand-binding positions. Nucleic Acids Res. *47*, 582–593.

Korthauer, K.D., and Kendziorski, C. (2015). MADGiC: a model-based approach for identifying driver genes in cancer. Bioinformatics *31*, 1526–1535.

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

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., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

McCubrey, J.A., Steelman, L.S., Bertrand, F.E., Davis, N.M., Sokolosky, M., Abrams, S.L., Montalto, G., D'Assoro, A.B., Libra, M., Nicoletti, F., et al. (2014). GSK-3 as potential target for therapeutic intervention in cancer. Oncotarget *5*, 2881–2911.

McGranahan, N., Favero, F., de Bruin, E.C., Birkbak, N.J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci. Transl. Med. *7*, 283ra54.

McGranahan, N., and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. Cell *168*, 613–628.

Melloni, G.E.M., de Pretis, S., Riva, L., Pelizzola, M., Céol, A., Costanza, J., Müller, H., and Zammataro, L. (2016). LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. BMC Bioinformatics *17*, 80.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., et al. (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. *43*, D213–D221.

Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. *34*, 267–273.

Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A., and López-Bigas, N. (2016). OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol. *17*, 128.

Munro, D., Ghersi, D., and Singh, M. (2018). Two critical positions in zinc finger domains are heavily mutated in three human cancer types. PLoS Comput. Biol. *14*, e1006290.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Niknafs, N., Kim, D., Kim, R.G., Diekhans, M., Ryan, M., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum. Genet. *132*, 1235–1243.

Nishi, H., Tyagi, M., Teng, S., Shoemaker, B.A., Hashimoto, K., Alexov, E., Wuchty, S., and Panchenko, A.R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. PLoS One *8*, e66273.

Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.W., Zhang, Q., McLellan, M.D., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. *48*, 827–837.

Peterson, T.A., Gauran, I.I.M., Park, J., Park, D., and Kann, M.G. (2017). Oncodomains: a protein domain-centric framework for analyzing rare variants in tumor samples. PLoS Comput. Biol. *13*, e1005428.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., and Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. Methods *74*, 83–89.

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., and Godzik, A. (2015). A pan-cancer catalogue of cancer driver protein interaction interfaces. PLOS Comput. Biol. *11*, e1004518.

Porta-Pardo, E., and Godzik, A. (2014). e-driver: a novel method to identify protein regions driving cancer. Bioinformatics *30*, 3109–3114.

Porta-Pardo, E., Kamburov, A., Tamborero, D., Pons, T., Grases, D., Valencia, A., Lopez-Bigas, N., Getz, G., and Godzik, A. (2017). Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat. Methods *14*, 782–788.

Przytycki, P.F., and Singh, M. (2017). Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes. Genome Med. *9*, 79.

Raimondi, F., Singh, G., Betts, M.J., Apic, G., Vukotic, R., Andreone, P., Stein, L., and Russell, R.B. (2017). Insights into cancer severity from biomolecular interaction mechanisms. Sci. Rep. *6*, 34490.

Reimand, J., and Bader, G.D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol. Syst. Biol. *9*, 637.

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118.

Ryslik, G.A., Cheng, Y., Cheung, K.H., Modis, Y., and Zhao, H. (2013). Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *14*, 190.

Ryslik, G.A., Cheng, Y., Cheung, K.H., Bjornson, R.D., Zelterman, D., Modis, Y., and Zhao, H. (2014a). A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. BMC Bioinformatics *15*, 231.

Ryslik, G.A., Cheng, Y., Cheung, K.H., Modis, Y., and Zhao, H. (2014b). A graph theoretic approach to utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics *15*, 86.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in The Cancer Genome Atlas. Cell *173*, 321–337.

Satelli, A., Rao, P.S., Thirumala, S., and Rao, U.S. (2011). Galectin-4 functions as a tumor suppressor of human colorectal cancer. Int. J. Cancer *129*, 799–809.

Schaub, F.X., Dhankani, V., Berger, A.C., Trivedi, M., Richardson, A.B., Shaw, R., Zhao, W., Zhang, X., Ventura, A., Liu, Y., et al. (2018). Pan-cancer alterations of the MYC oncogene and its proximal network across the Cancer Genome Atlas. Cell Syst. *6*, 282–300.e2.

Seiler, M., Peng, S., Agrawal, A.A., Palacino, J., Teng, T., Zhu, P., Smith, P.G., Cancer Genome Atlas Research Network, Buonamici, S., and Yu, L. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. Cell Rep. *23*, 282–296.e4.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics *29*, 1504–1510.

Silverbush, D., Cristea, S., Yanovich-Arad, G., Geiger, T., Beerenwinkel, N., and Sharan, R. (2019). Simultaneous integration of multi-omics data improves the identification of cancer driver modules. Cell Syst *8*, 456–466.e5.

Stehr, H., Jang, S.H., Duarte, J.M., Wierling, C., Lehrach, H., Lappe, M., and Lange, B.M. (2011). The structural impact of cancer-associated missense mutations in oncogenes and tumor suppressors. Mol. Cancer *10*, 54.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.

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

TCGA Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas pan-cancer analysis project. Nat. Genet. *45*, 1113–1120.

The 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res *47*, D506–D515.

Tokheim, C., Bhattacharya, R., Niknafs, N., Gygax, D.M., Kim, R., Ryan, M., Masica, D.L., and Karchin, R. (2016a). Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. Cancer Res. *76*, 3719–3731.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., and Karchin, R. (2016b). Evaluating the evaluation of cancer driver genes. Proc. Natl. Acad. Sci. U S A *113*, 14330–14335.

Torkamani, A., and Schork, N.J. (2008). Prediction of cancer driver mutations in protein kinases. Cancer Res. *68*, 1675–1682.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Ye, J., Pavlicek, A., Lunney, E.A., Rejto, P.A., and Teng, C.H. (2010). Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics *11*, 11.

Zaykin, D.V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. J. Evol. Biol. *24*, 1836–1841.

Zeng, Q., Michael, I.P., Zhang, P., Saghafinia, S., Knott, G., Jiao, W., McCabe, B.D., Galván, J.A., Robinson, H.P.C., Zlobec, I., et al. (2019). Synaptic proximity enables NMDAR signalling to promote brain metastasis. Nature *573*, 526–531.

Zhang, X., Choi, P.S., Francis, J.M., Gao, G.F., Campbell, J.D., Ramachandran, A., Mitsuishi, Y., Ha, G., Shih, J., Vazquez, F., et al. (2018). Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. Cancer Discov. *8*, 108–125.

Zhao, J., Cheng, F., and Zhao, Z. (2017). Tissue-specific signaling networks rewired by major somatic mutations in human cancer revealed by proteome-wide discovery. Cancer Res. *77*, 2810–2821.

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
InteracDome	Kobren and Singh, 2019	https://interacdome.princeton.edu
Masked somatic exome mutations and RNA-Seq expression data from The Cancer Genome Atlas (TCGA)	Grossman et al., 2016	https://portal.gdc.cancer.gov
Cancer Gene Census	Futreal et al., 2004	https://cancer.sanger.ac.uk/census
UCSC Genome Browser 100-vertebrate alignment	Meyer et al., 2013	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/multiz100way/
Pfam	Finn et al., 2014	https://pfam.xfam.org
1000 Genomes Project	The 1000 Genomes Project Consortium, 2012	https://www.internationalgenome.org/data
Kandoth cancer driver genes	Kandoth et al., 2013, Table S4	https://media.nature.com/original/nature-assets/nature/journal/v502/n7471/extref/nature12634-s1.zip
Lawrence cancer driver genes	Lawrence et al., 2014, Table S2	https://media.nature.com/original/nature-assets/nature/journal/v505/n7484/extref/nature12912-s3.xlsx
Bailey, Tokheim cancer driver genes	Bailey et al., 2018, Table S1	https://ars.els-cdn.com/content/image/1-s2.0-S009286741830237X-mmc1.xlsx
Vogelstein cancer driver genes	Vogelstein et al., 2013, Tables S2A, S2B, S3A–S3C, and S4	http://science.sciencemag.org/highwire/filestream/594203/field_highwire_adjunct_files/1/1235122TablesS1-4.xlsx
Davoli "negative" driver genes	Davoli et al., 2013, Table S2A	https://ars.els-cdn.com/content/image/1-s2.0-S0092867413012877-mmc2.xlsx
DISEASES	Pletscher-Frankild et al., 2015	http://download.jensenlab.org/human_disease_textmining_full.tsv
UniProtKB	The UniProt Consortium, 2019	https://www.uniprot.org/uniprot/
NegAgoFull and NegAgoClean "negative" driver genes	Silverbush et al., 2019; Tables S1D and S1C	https://ars.els-cdn.com/content/image/1-s2.0-S2405471219301474-mmc2.xlsx
Protein Data Bank	Berman et al., 2000	https://www.rcsb.org
Pfam2Go	Mitchell et al., 2015	http://current.geneontology.org/ontology/external2go/pfam2go
Reactome pathways	Jassal et al., 2020	https://reactome.org
Software and Algorithms		
PertInInt	This paper	https://github.com/Singh-Lab/PertInInt
HMMER	Eddy, 2011	http://hmmer.org
DiffMut	Przytycki and Singh, 2017	https://diffmut.princeton.edu
CanBind	Ghersi and Singh, 2014	http://canbind.princeton.edu
Hotspot	Chang et al., 2016	https://github.com/taylor-lab/hotspots
MutSigCV	Lawrence et al., 2014	https://software.broadinstitute.org/cancer/cga/mutsig_run
OncodriveFML	Melloni et al., 2016	https://bitbucket.org/bbglab/oncodrivefml/src/master/
OncodriveClust	Tamborero et al., 2013	http://bg.upf.edu/group/projects/oncodrive-clust.php
NMC	Ye et al., 2010	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2822753/bin/1471-2105-11-11-S1.DOC

(*Continued on next page*)

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ActiveDriver	Reimand and Bader, 2013	https://github.com/reimandlab/ActiveDriver
eDriver	Porta-Pardo and Godzik, 2014	https://github.com/eduardporta/e-Driver.git
LowMACA	Melloni et al., 2016	https://bioconductor.org/packages/LowMACA
GraphPAC	Ryslik et al., 2014a	https://bioconductor.org/packages/GraphPAC/
iPAC	Ryslik et al., 2013	https://www.bioconductor.org/packages/iPAC/
SpacePAC	Ryslik et al., 2014a	https://www.bioconductor.org/packages/SpacePAC/
eDriver3D	Porta-Pardo et al., 2015	https://github.com/eduardporta/e-Driver.git
Broad's GSEA, version 4.0.3	Mootha et al., 2003; Subramanian et al., 2005	https://www.gsea-msigdb.org/gsea/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mona Singh (mona@cs.princeton.edu).

Materials Availability

This study did not generate new unique reagents or materials.

Data and Code Availability

All original code generated during this study are available at http://github.com/Singh-Lab/PertInInt.

METHOD DETAILS

Protein Site-Based Functional Tracks

Any pre-defined functional region of a protein can be encoded as a track in the PertInInt framework. Currently, we consider three types of per-site functional annotations—interaction, domain, and conservation—the former two of which may yield multiple subgene resolution tracks per protein. Each type of track is described in more detail below.

Interaction Tracks

Interaction tracks correspond to portions of a protein that are inferred to interact with ligands. These tracks arise in two ways.

First, we leverage sequence homology directly to transfer information from co-complex structures to human protein sequences as previously described (Ghersi and Singh, 2014). For proteins with one or more regions whose structure in complex with a ligand could be homology-modeled, we introduce a track for each contiguous homology-matched region. Per-position weights reflect the observed residue-to-ligand proximities, computed as the fraction of atoms in the amino acid residue found within 4.0Å of the ligand.

Second, we utilize the set of "confident" domain–ligand interactions from the InteracDome database (v0.3) (Kobren and Singh, 2019) to identify putative ligand-binding positions. We subset this collection to the 9,142 domain–ligand interactions across 1,850 domains that were characterized by InteracDome using at least five structural instances. Each position within an InteracDome domain is associated with a "binding frequency" between 0 and 1 that corresponds to the fraction of the time residues in this position were found to be in contact with the ligand of interest when analyzing co-complex structures. For each human protein, we identify instances of InteracDome domains using HMMER (v2.3.2 and v3.1b2), and require complete, high-scoring domain instances as previously described (Eddy, 2011; Finn et al., 2014; Kobren and Singh, 2019). Within a protein, there is a separate track for each domain–ligand instance within it; this track consists of the residues comprising the match states of the domain, and the weights of these residues are the binding frequencies for the ligand in the corresponding domain positions.

Finally, we note that some domain interactions are mediated not by individual domain instances but by repeating instances of the same domain family. To capture these interfaces, we also consider additional tracks encoding multiple instances of the same domain family in a protein; these "aggregate" tracks span noncontiguous intervals that correspond to the locations of individual domain instances, with track positions weighted according to the binding frequencies at corresponding domain match states as described above. Interaction domain tracks corresponding to domain families with 40+ instances in the same protein are replaced by their aggregate tracks.

Please cite this article in press as: Kobren et al., PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities, Cell Systems (2020), https://doi.org/10.1016/j.cels.2020.06.005

Domain Tracks

For each Pfam-A (v31.0) domain instance within a protein sequence, there is a domain track that specifies which amino acids comprise the domain (Finn et al., 2014). Domain tracks span the length of the protein, and positions within and outside of the domain instance are respectively assigned weights of 1 and 0. We again also encode aggregate domain tracks as before to model functional regions mediated by repetitive domain families.

Conservation Tracks

Each protein has a single conservation track. We obtain the 100-vertebrate cross-species protein multiple sequence alignment from the UCSC Genome Browser (Meyer et al., 2013) and compute per-protein-position conservation-based functionality weights by multiplying the fraction of non-gap residues in the column by the Jensen-Shannon divergence (JSD) between those non-gap residues and a Blosum 62 background amino acid distribution (Capra and Singh, 2007).

Per-track Functional Mutation Scores

Suppose we have a protein sequence of length L spanning positions $P = \{p_1,\ldots,p_L\}$. This protein is associated with multiple "tracks" W, each defined as $W \subseteq P$, where each position $p_i \in W$ is associated with a real-valued weight $w_i \in [0,1]$ reflecting its functionality with respect to the track. Suppose there are n cancer somatic missense mutations across a cohort of tumor samples that fall in positions included in track W. For each mutation i, let $z_i \in \{z_1,\ldots,z_n\}$ be the weight in track W of the position where that mutation lies. We further consider the case where each mutation i is associated with a value $f_i \in (0,1]$; here, each f_i is set to the proportion of sequencing reads that contain the mutation (i.e., its subclonal fraction), which has previously been shown to be associated with a mutation's relevance in cancer (McGranahan et al., 2015). The score of the somatic mutations with respect to track W is then defined as:

$$S_W = \sum_{i=1}^{n} f_i z_i. \tag{Equation 1}$$

Intuitively, this score reflects the extent to which somatic mutations are falling into functionally important positions within a track.

Per-track Analytical *Z* score Calculations

For a given score S_W for a track, we next want to determine if this score is higher than we would expect by chance. One approach would be to repeatedly randomize the mutations within the positions of the track and use the distribution of resulting scores to compute an empirical p value. Here we show that we can determine the significance of these scores analytically, obviating the need for empirical mutation shuffles and dramatically improving runtime (Figure S2). Note that in the absence of any selective pressure, the values $z_1,\ldots,z_n$ are independent and identically distributed (i.i.d.) random variables. We leverage this observation to directly compute the significance of S_W. First, we model all mutation locations z_i as being drawn from the same background mutation model $\lambda_1, \ldots, \lambda_L$, where λ_i is the probability that a mutation affects position i. If every position i within a protein of length L is equally likely to harbor a missense mutation, $\lambda_i = 1/L$. Here, we incorporate codon-specific missense mutation probabilities as well as cancer-specific C/G-mutation biases into our background mutation model (see "Per-site background mutational model" STAR Methods section below). We linearly scale these values with respect to each track W such that $\sum_{j\in W}\lambda_j^W = 1$. The expected weight of the position in which mutation i lies ($\mathbb{E}[z_i]$) and its variance ($\sigma_{z_i}^2$) with respect to this null distribution are computed as

$$\mathbb{E}[z_i] = \sum_{j\in W}\lambda_j^W w_j$$

and

$$\sigma_{z_i}^2 = \mathbb{E}\left[z_i^2\right] - (\mathbb{E}[z_i])^2 = \sum_{j\in W}\lambda_j^W w_j^2 - \left(\sum_{j\in W}\lambda_j^W w_j\right)^2.$$

Because the total score of the set of mutations affecting track W (i.e., S_W) is a sum of independent random variables (Equation 1), the expectation and variance of S_W can also be calculated directly as

$$\begin{aligned}\mathbb{E}[S_W] &= \sum_{i=1}^{n}\mathbb{E}[f_i z_i] \\ &= \sum_{i=1}^{n} f_i \bullet \mathbb{E}[z_i] \\ &= \sum_{i=1}^{n} f_i\left(\sum_{j\in W}\lambda_j^W w_j\right)\end{aligned} \tag{Equation 2}$$

and

$$\begin{aligned}\sigma^2_{S_W} &= \sum_{i=1}^{n} \sigma^2_{f_i z_i} \\ &= \sum_{i=1}^{n} f_i^2 \sigma^2_{z_i} \\ &= \sum_{i=1}^{n} f_i^2 \left(\sum_{j \in W} \lambda_j^W w_j^2 - \left(\sum_{j \in W} \lambda_j^W w_j \right)^2 \right).\end{aligned} \qquad \text{(Equation 3)}$$

Finally, to determine the significance of the actual score S_W, which indicates the propensity of somatic mutations to fall into highly weighted positions in a track, since the sum of independent random variables tends towards a normal distribution, we compute the mutational enrichment Z score for each track W as

$$Z_W = \frac{S_W - \mathbb{E}[S_W]}{\sigma_{S_W}}. \qquad \text{(Equation 4)}$$

We note that if we (1) restricted each weight within a track to be 0/1 rather than real-valued, (2) restricted mutations to have equal f_i values of 1, and (3) restricted the λ_i to be uniform across the track, we could determine per-track significance analytically using the binomial distribution. Note that with these restrictions, however, we would not be able to incorporate real-valued functionality weights from conservation or interaction tracks, subclonal mutation fractions, or mutational signatures.

Per-site Background Mutational Model

We model the likelihoods of protein positions $p_1, \ldots, p_L$ to harbor a missense mutation as $\lambda_1, \ldots, \lambda_L$ such that

$$\lambda_j = \sum_{d \in \{1,2,3\}} \left(B_{jd} \bullet \sum_{u \in \{A,T,C,G\}} M_{jdu} \right) \qquad \text{(Equation 5)}$$

where

$$B_{jd} = \begin{cases} 1 & \text{if the } d\text{th nucleotide in the codon at position } p_j \text{ is A or T} \\ b & \text{otherwise, where } b \text{ is the relative frequency of a C/G mutation in the pan-cancer dataset as compared} \\ & \text{to a A/T mutation (i.e., 3.063)} \end{cases}$$

and

$$M_{jdu} = \begin{cases} 1 & \text{if changing the } d\text{th nucleotide in the codon at position } p_j \text{ to } u \text{ results in a missense mutation} \\ 0 & \text{otherwise} \end{cases}$$

Between-Track Analytical Covariance Calculation

In our framework, a single protein may be associated with *multiple* tracks, each representing a distinct aspect of protein functioning. Since tracks can share positions, the track scores with respect to a set of somatic mutations are not independent of each other, and thus we need to determine their covariance.

Suppose we consider two tracks $V \subseteq P$ and $W \subseteq P$, where each position $p_i \in V$ is associated with a weight v_i and each position $p_i \in W$ is associated with a weight w_i. Suppose there are m mutations (with associated values $f'_1, \ldots, f'_m$) that involve positions within track V, and n mutations (as before with associated values $f_1, \ldots, f_n$) that involve positions within track W. Let $y_1, y_2, \ldots, y_m$ be the weights of the positions that the m mutations in track V fall into, and let $z_1, z_2, \ldots, z_n$ be the weights of the positions that the n mutations in track W fall into. Scores are thus calculated as before for tracks V and W as

$$S_V = \sum_{i=1}^{m} f'_i y_i$$

and

$$S_W = \sum_{i=1}^{n} f_i z_i.$$

Let $X = V \cap W$. If the two tracks do not overlap (i.e., $X = \varnothing$), then the covariance between S_V and S_W is 0. Otherwise, note that $S_V = S'_V + S^X_V$, where $S'_V = \sum_{j \notin X} f'_j y_j$ and $S^X_V = \sum_{j \in X} f'_j y_j$. Similarly, $S_W = S'_W + S^X_W$. Therefore, we can write covariance as

$$\text{cov}\left[S_V, S_W\right] = \text{cov}\left[\left(S'_V + S^X_V\right), \left(S'_W + S^X_W\right)\right].$$

Because the covariance is bilinear, we can now expand this equation as

$$\mathrm{cov}\left[\left(S'_V + S^X_V\right), \left(S'_W + S^X_W\right)\right] = \mathrm{cov}[S'_V, S'_W] + \mathrm{cov}[S'_V, S^X_W] + \mathrm{cov}[S^X_V, S'_W] + \mathrm{cov}\left[S^X_V, S^X_W\right].$$

Finally, because mutations landing in track V outside of the overlap region X have no bearing on S_W and vice versa, the first three covariance terms in the equation above will be evaluated as 0, leaving us with

$$\mathrm{cov}[S_V, S_W] = \mathrm{cov}\left[S^X_V, S^X_W\right].$$

In our framework, we compute covariance conditional on the q mutations observed to fall on positions shared by tracks V and W. Let $F = f^2_{i_1} + \cdots + f^2_{i_q}$, with the f_i associated with the q mutations in X. With the number of mutations q fixed, we have

$$\mathrm{cov}\left[S^X_V, S^X_W\right] = \mathrm{cov}\left[\sum_{j=1}^{q} f_{i_j} y_{i_j}, \sum_{j=1}^{q} f_{i_j} z_{i_j}\right] = \sum_{j=1}^{q} \sum_{k=1}^{q} \mathrm{cov}\left[f_{i_j} y_{i_j}, f_{i_k} z_{i_k}\right].$$

Note that the *same* mutations from tracks S^X_V and S^X_W land on the *same* position in the overlap region and simultaneously impact the S_V and S_W scores, whereas any other pair of mutations $j \neq k$ are independent. Hence,

$$\begin{aligned}
\mathrm{cov}\left[S^X_V, S^X_W\right] &= \sum_{j=1}^{q} \mathrm{cov}\left[f_{i_j} y_{i_j}, f_{i_j} z_{i_j}\right] \\
&= \sum_{j=1}^{q} f^2_{i_j} \cdot \mathrm{cov}\left[y_{i_j}, z_{i_j}\right] \\
&= F \cdot \mathrm{cov}\left[y_{i_1}, z_{i_1}\right] \\
&= F\left(\sum_{j \in X} \lambda^X_j v_j w_j - \left(\sum_{j \in X} \lambda^X_j v_j\right)\left(\sum_{j \in X} \lambda^X_j w_j\right)\right)
\end{aligned} \qquad \text{(Equation 6)}$$

Analytical Formulation Enables Precomputation

Note that the per-track expectation, variance and covariance calculations (Equations 2, 3, and 6) can each be rewritten as $C \bullet \sum_i f_i$ or $C \bullet \sum_i f^2_i$, where C is fixed per track. We therefore precompute the per-track expectations, variances, and cross-track covariances assuming a single mutation of value 1, and scale these precomputed values at runtime by the mutations observed to fall into each track; this allows PertInInt to achieve an additional 16–18× speedup at runtime (Figure S11).

Whole Gene Mutability Tracks

Using the same analytical formulation described above, we can also compute a Z score per gene reflecting whether the gene is more mutated overall than we might expect. We define a natural variation track of length L = 19,460 for each gene, where the entry corresponding to the gene of interest has a functionality weight of 1, and all other entries have weights of 0 (i.e., one-hot gene encodings). We then compute a corresponding background mutability probability distribution $\lambda_1, \ldots \lambda_L$ based on how much each gene varies naturally across healthy human populations. Specifically, for each of 2,504 individuals included in the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2012), we first min-rank all protein-coding genes by their variant count, linearly scale these ranks to fall between 0 and 1, then round each normalized rank down to its nearest hundredth, which we refer to as its bin. We compute the expected bin value (across individuals) for each gene, and finally to derive the values of $\lambda_1, \ldots, \lambda_L$ linearly scale these per-gene expected bin values such that they sum to 1 across all genes. For each track, we use this background mutation model and the n mutations observed to fall across all 19,460 genes to analytically compute *per-gene* expectations, variances, and Z scores as before. Note that PertInInt models 23,278 genes—of which 20,356 are on chromosomes 1–22, X or Y—but only 19,460 genes were profiled in the 1000 Genomes Project, and thus only this many genes have associated natural variation tracks. The covariance between the natural variation track and all subgene tracks is set to 0.

Since the whole-gene track W for gene G_j is a one-hot encoding, we can simplify Equation 1 as $S_W = \sum_{i \in G_j} f_i$, Equation 2 as $\mathbb{E}[S_W] = \lambda^W_j \sum_{i=1}^{n} f_i$ and Equation 3 as $\sigma^2_{S_W} = \lambda^W_j (1 - \lambda^W_j) \sum_{i=1}^{n} f^2_i$.

Because the number of mutations affecting all genes is often substantially larger than the number of mutations to affect any single gene, the whole-gene Z scores can be much larger than for the other tracks. We thus effectively subsample the total number of mutations by a factor s—set to $\frac{1}{\sqrt{n}}$ in our implementation—to compute the whole-gene Z scores using the values below before combining them with other subgene Z scores:

$$S_W = s \sum_{i \in G_j} f_i,$$

$$\mathbb{E}[S_W] = s \lambda^W_j \sum_{i=1}^{n} f_i,$$

$$\sigma^2_{S_W} = s\lambda_j^W\left(1 - s\lambda_j^W\right)\sum_{i=1}^{n} f_i^2.$$

Combining Per-track *Z* scores for Each Protein

We evaluate the significance of the scores for all tracks simultaneously using a multivariate normal distribution. Recall that our per-track somatic mutation functional scores (S_W, Equation 1) and their analytically-derived Z scores (Equation 4) computed for random assignments of mutations are normally distributed when the number of mutations (n) is sufficiently large (i.e., by the Central Limit Theorem).

For each track W, we empirically determine this minimum n by randomly assigning up to 500 mutations to the track 1,000 times in accordance with the corresponding background mutation model (i.e., the λ_i's) and recomputing S_W each time. At each mutation count, we ask whether we can reject the null hypothesis that the mutation functional scores are normally distributed via the Shapiro-Wilk test with p value < 5e-5. We keep track of the minimum number of mutations per track where we could no longer confidently reject the normality assumption. Only scores derived from mutated tracks with the corresponding required minimum mutation count are modeled together in our multivariate Gaussian. We pre-compute this minimum mutation count value for each track (i.e., before evaluating any cancer somatic mutation data).

For each mutated protein, we compute a single combined score using a weighted Z-transform test with correlation correction (Zaykin, 2011) as

$$Z = \frac{\sum_{i=1}^{k} c_i Z_i}{\sqrt{\sum_{i=1}^{k} c_i^2 + 2\sum_{i<j} c_i c_j r_{ij}}} \qquad \text{(Equation 7)}$$

where k is the number of tracks with their required minimum mutation count and positive Z scores, Z_i corresponds to the Z score associated with track i, c_i is a weight indicating the "confidence" of track i, and r_{ij} is the correlation between tracks i and j (i.e., $r_{ij} = \mathrm{cov}(S_i, S_j)/\sigma_i\sigma_j$). In order to consider each type of functionality data equally, we assign per-track confidences c_i such that the four functional track groups (i.e., interaction, domain, conservation, and natural variation; Figures 1A and S1) each contribute a quarter of the overall confidence. Within the interaction and domain track groups, where there may be 2+ tracks per group, confidence weights are assigned proportionally to $\sqrt{m}$, where m is the total number of mutations to fall into positively-weighted positions in the track. Finally, we assign a single score per gene by taking the maximum combined score achieved by any of its corresponding protein isoforms.

QUANTIFICATION AND STATISTICAL ANALYSIS

Cancer Mutation Data Preparation

We downloaded all open-access TCGA somatic exome mutation data and RNA-seq expression data from NCI's Genomic Data Commons on July 15, 2017 (Fan et al., 2016; Grossman et al., 2016). We convert gene expression values (FPKM) to transcripts per million (TPM) and exclude mutations from genes that were expressed at <0.1 TPM in the corresponding tumor sample. For the 765 samples with missing expression data, we exclude mutations from genes that were expressed at <0.1 TPM on average across other tumor samples of the same tissue type. These steps resulted in a filtered set of 1,141,609 missense, 442,070 silent, and 94,813 nonsense mutations across 18,613 genes from 33 cancer types (Figure S3); note that we consider the unfiltered (by expression) set of 1,473,729 missense, 578,407 silent and 118,921 nonsense mutations across 19,550 genes when running alternate methods and when running PertInInt to compare to alternate methods. We combine COAD and READ cancer types into the COAD-READ group, and GBM and LGG cancer types into the GBMLGG group for per-cancer performance testing (Figures S8 and S9).

Runtime Analysis

PertInInt, as well as all algorithm variants of PertInInt and all alternate methods, are run as sole processes on single CPUs, each with a 2.4–2.7 Ghz processor and 30GB of RAM. Methods are timed using Python's time package, and the real (i.e., "wall clock") elapsed time is reported.

Testing Related Driver Detection Methods

We classify alternate cancer driver detection methods based on the mutational patterns they detect; these include whole gene enrichment, *de novo* linear clustering, enrichment in linear externally defined regions, *de novo* three-dimensional (3D) clustering, or enrichment in 3D externally defined regions (as in Porta-Pardo et al., 2017). We include methods from each of these five groups that require only mutational and/or structural input from the user and have open-source implementations that run locally on a 64-bit Linux machine using sample input. We test the whole gene methods DiffMut (Przytycki and Singh, 2017), MutSigCV (Lawrence et al., 2014), and OncodriveFML (Mularoni et al., 2016); the linear clustering methods Hotspot (Chang et al., 2016), OncodriveClust (Tamborero et al., 2013), and NMC (Ye et al., 2010); the linear externally defined regions methods ActiveDriver (Reimand and Bader, 2013), eDriver (Porta-Pardo and Godzik, 2014), and LowMACA (Melloni et al., 2016); the 3D clustering methods GraphPAC (Ryslik et al., 2014b), iPAC (Ryslik et al., 2013), and SpacePAC (Ryslik et al., 2014a); and the 3D externally defined regions method eDriver3D

(Porta-Pardo et al., 2015). We note that in addition to overall mutation frequency, MutSigCV also considers linear clustering of mutations within genes and the functional impact of mutations based on evolutionary conservation.

All methods including PertInInt are run on the same mutation datasets before our filtering step of limiting to mutations from expressed genes. Additional data files required for individual methods are obtained from their most recent online repositories or otherwise from their original publications. For 3D clustering methods, we select a single structural template for each human protein wherever possible as suggested (i.e., preferring native over bound form, longer length, higher sequence identity, higher resolution, and smaller R-value; Berman et al., 2000). We note that because these 3D clustering methods only run on proteins with corresponding structural information, their results may be biased toward known cancer genes (i.e., 55.3% of cancer genes have structural templates whereas 28.1% of all genes have structural templates). Methods are run with default parameters, except GraphPAC and SpacePAC, where the significance threshold (α) is set to 1.0 to maximize the number of scored genes returned.

For each method, enrichment for known driver genes on increasingly larger sets of predictions is computed; enrichment is computed at each gene rank on pan-cancer results and at every tenth gene on per-cancer results to reduce the impact of minor reorderings of the relatively small number of known driver genes detected across these datasets. Specifically, we calculate enrichment as the fraction of known driver genes in the gene set (i.e., the precision) divided by the fraction of known driver genes in the whole gene set considered; unmutated driver genes with respect to each mutation dataset are excluded entirely. For this evaluation, we consider the CGC set of known driver genes as a gold standard, as well as alternate positive and negative sets of cancer genes in turn.

Note that LowMACA, NMC, and all three 3D clustering methods did not finish running without error within 30 days on the pan-cancer dataset. We were also unable to run and obtain results from NMC on an additional four individual cancer types (UCEC, SKCM, COADREAD, and LUSC), and thus exclude this method from our evaluation.

Cell Systems, Volume *11*

Supplemental Information

PertInInt: An Integrative, Analytical Approach to Rapidly Uncover Cancer Driver Genes with Perturbed Interactions and Functionalities

Shilpa Nadimpalli Kobren, Bernard Chazelle, and Mona Singh

Supplemental Figures

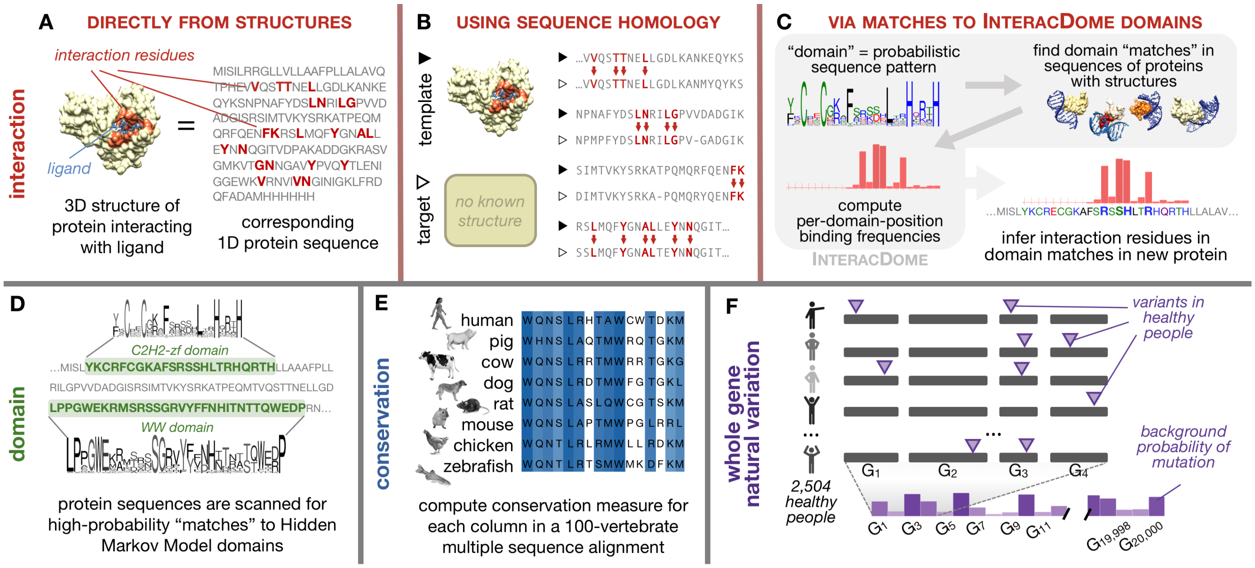

A
DIRECTLY FROM STRUCTURES
interaction
interaction residues
ligand
=
MISILRRGLLVLLAAFPLLALAVQ
TPHEVVQSTTNELLGDLKANKE
QYKSNPNAFYDSLNRILGPVVD
ADGISRSIMTVKYSRKATPEQM
QRFQENFKRSLMQFYGNALL
EYNNQGITVDPAKADDGKRASV
GMKVTGNNGAVYPVQYTLENI
GGEWKVRNVIVNGINIGKLFRD
QFADAMHHHHHH
3D structure of protein interacting with ligand
corresponding 1D protein sequence
B
USING SEQUENCE HOMOLOGY
template ▶
target ▷
no known structure
▶ …VVQSTTNELLGDLKANKEQYKS
▷ …VVQSTTNELLGDLKANMYQYKS
▶ NPNAFYDSLNRILGPVVDADGIK
▷ NPMPFYDSLNRILGPV-GADGIK
▶ SIMTVKYSRKATPQMQRFQENFK
▷ DIMTVKYSRKA-PQMQRYQENFK
▶ RSLMQFYGNALLEYNNQGIT…
▷ SSLMQFYGNALTEYNNQGIT…
C
VIA MATCHES TO INTERACDOME DOMAINS
"domain" = probabilistic sequence pattern
find domain "matches" in sequences of proteins with structures
compute per-domain-position binding frequencies
INTERACDOME
…MISLYKCRECGKAFSRSSHLTRHQRTHLLALAV…
infer interaction residues in domain matches in new protein
D
domain
C2H2-zf domain
…MISLYKCRFCGKAFSRSSHLTRHQRTHLLAAAFPLL
RILGPVVDADGISRSIMTVKYSRKATPEQMTVQSTTNELLGD
LPPGWEKRMSRSSGRVYFFNHITNTTQWEDPRN…
WW domain
protein sequences are scanned for high-probability "matches" to Hidden Markov Model domains
E
conservation
human WQNSLRHTAWCWTDKM
pig WHNSLAQTMWRQTGKM
cow WQNSLRRTMWRRTGKG
dog WQNSLRDTMWFGTGKL
rat WQNSLASLQWCGTSKM
mouse WQNSLAPTMWPGLRRL
chicken WQNTLRLRMWLLRDKM
zebrafish WQNTLRTSMWMKDFKM
compute conservation measure for each column in a 100-vertebrate multiple sequence alignment
F
whole gene natural variation
variants in healthy people
…
…
2,504 healthy people
G_1
G_2
G_3
G_4
background probability of mutation
G_1 G_3 G_5 G_7 G_9 G_{11}
$G_{19,998}$ $G_{20,000}$

Figure S1: Intuition and data sources behind PertInInt's four track types. ***Related to Figure 1.*** Graphical description of the sources of (A–C) interaction tracks, (D) domain tracks, (E) conservation tracks, and (F) natural variation tracks used by PertInInt. **(A)** Residues within human proteins that contact ligands can be directly determined from a 3-dimensional structure of that protein in complex with a ligand, if such a co-complex structure exists (left). These positions are then marked as "interaction residues" in the corresponding protein sequence (right). **(B)** Interaction residue information from a template protein (depicted by ▶) with a solved co-complex structure can be transferred to a homologous target protein (depicted by ▷) in regions with high sequence similarity, as previously described (Ghersi and Singh, 2014). **(C)** Steps in the shaded box summarize how the previously published InteracDome database (Kobren and Singh, 2019) was generated: matches to protein domains—represented as probabilistic sequence patterns in the form of Hidden Markov Models in Pfam (Finn et al., 2014)—are found in the sequences of proteins that have solved co-complex structures, and then each position within the domain is assigned a "binding frequency" value that corresponds to the fraction of times a residue at that position is found to be in contact with a ligand across co-complex structures. Human protein sequences are scanned for matches to InteracDome domains, and binding frequency information is transferred from the InteracDome domain pattern to the human protein sequence at the site of the domain match (bottom right). **(D)** Human protein sequences are scanned for matches to any Pfam domain. Each domain match generates a new domain track, where protein positions within the domain match region get a score of 1 and protein positions outside get a score of 0. **(E)** For each human protein, a per-position score reflects that position's conservation, computed as previously described (Capra and Singh, 2007) from the corresponding column in a 100-vertebrate multiple sequence alignment. **(F)** Human genes are ranked by the number of variants observed to affect them across a population of healthy individuals and then converted to a background probability of mutation, as previously described (Przytycki and Singh, 2017), to comprise the natural variation tracks.

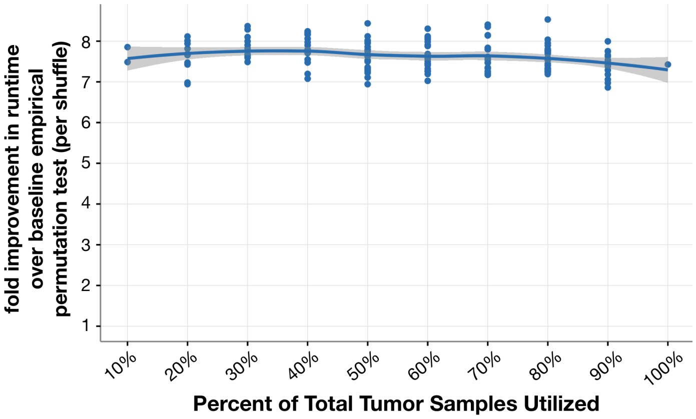


Figure S2: PertInInt's analytical approach results in $>7\times$ speedup over baseline empirical permutation approach. ***Related to Figure 1; STAR Methods.*** As a function of the percent (10–100%) of all tumor samples randomly selected from the pan-cancer dataset (*x*-axis), PertInInt's runtime is compared to a baseline version that uses 1,000 empirical permutations of mutations to estimate *Z*-scores for each track. Shown on the *y*-axis is the fold speedup in runtime for ten random selections of tumor samples of each size. The speedup shown is *per permutation* (i.e., divided by 1,000—the total number of permutations performed across each track). The solid blue line represents the local polynomial regression line, with the gray shading showing standard error. Due to the relatively large runtime of the empirical shuffling procedure, these runtime comparisons use only a single track per protein, conservation.

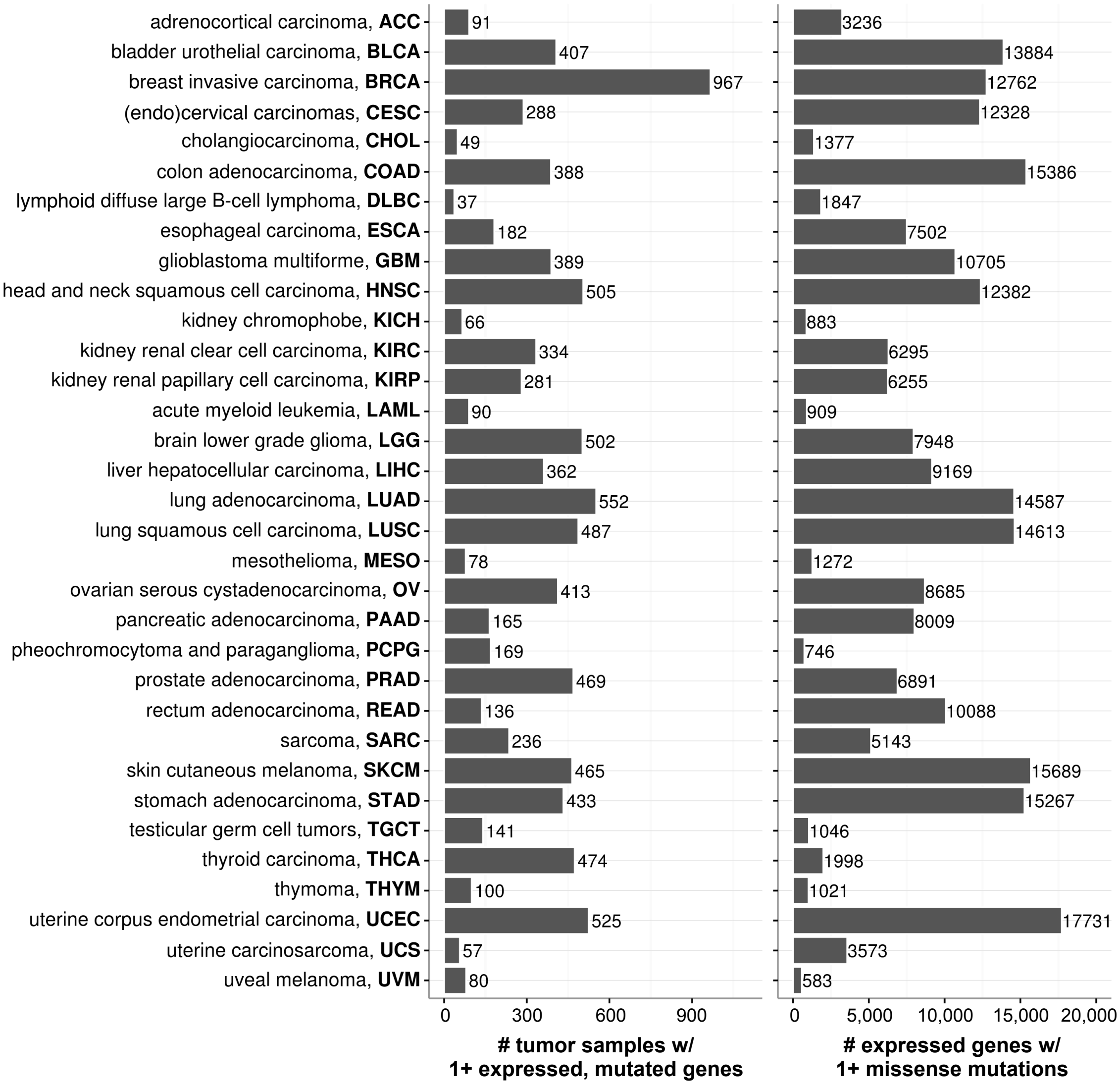


Figure S3: Summary of somatic mutation data. ***Related to Figure 2; Figure 3.*** Somatic mutation data obtained from NCI's Genomic Data Commons Data Portal for 33 cancer types (Fan et al., 2016). The number of tumor samples with 1+ expressed (TPM $\geq$ 0.1) genes with at least one missense mutation is shown in the left plot. The number of genes that are expressed in 1+ tumor samples and have at least one missense mutation is shown in the right plot.

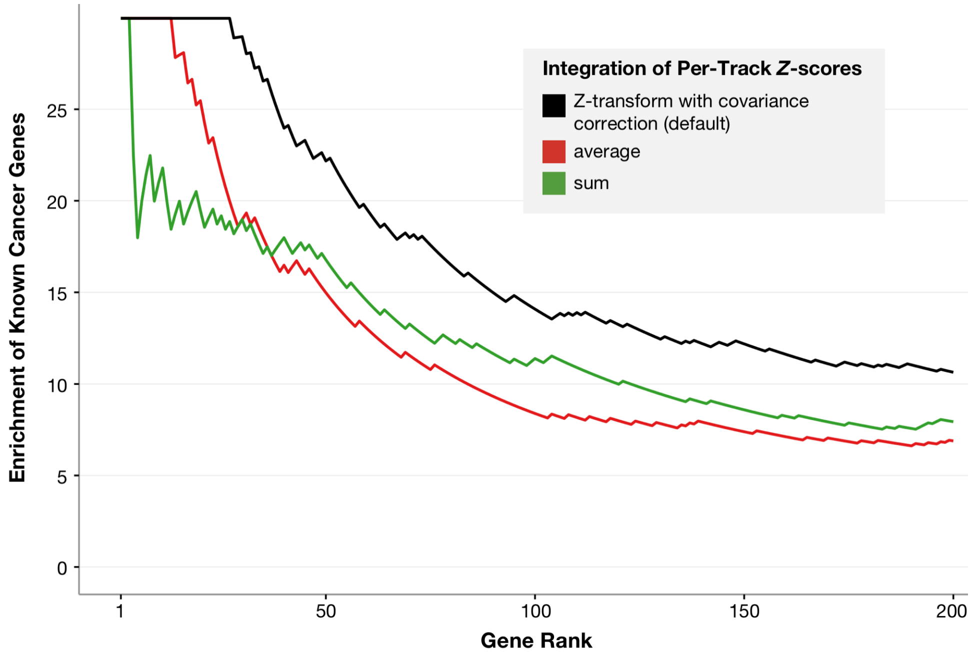


Figure S4: Covariance-based track integration outperforms naïve track integration. ***Related to STAR Methods.*** The default version of PertInInt (black line) combines per-track Z-scores using an analytically-computed covariance matrix to account for between-track dependencies. We implemented versions of PertInInt where per-track Z-scores are combined using mean (red line) and summation (green line) to generate two new ranked lists of genes on the pan-cancer dataset. Note that these two naïve track integrations are incorrect because they do not account for the dependencies across tracks. For each ranked list of genes, we compute enrichment as the ratio between the fraction of gold standard CGC genes in the top ranked genes (i.e., the precision) and the fraction of CGC genes in the whole set of genes (i.e., the expected precision given a random ordering of genes). All curves converge to an enrichment of 1 by the end of the ranked list of genes (not shown).

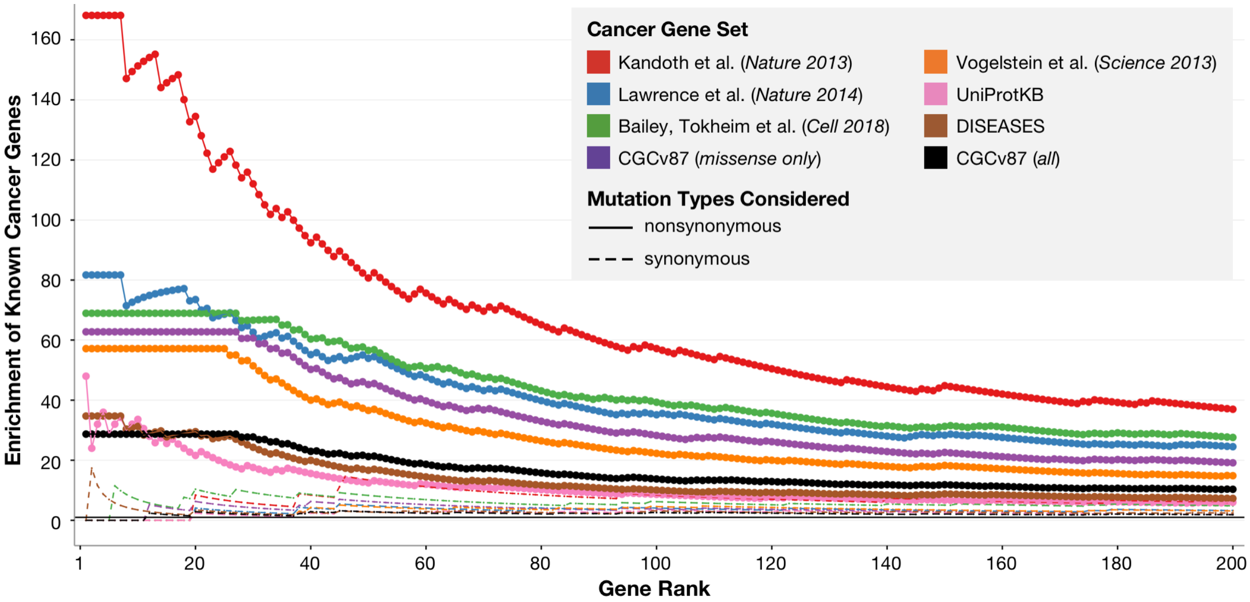


Figure S5: Highly ranked genes are enriched in cancer genes. ***Related to Figure 4; Table S2.*** Gold standard driver gene sets include: 123 genes listed in Kandoth et al., 2013, Table S4 (red), 249 genes listed in Lawrence et al., 2014, Table S2 (blue), 295 genes listed in Bailey et al., 2018, Table S1 (green), all 358 oncogenes and TSGs listed in Vogelstein et al., 2013, Tables S2A-B, S3A-C, S4 (orange), 428 genes from UniProtKB (The UniProt Consortium, 2018) annotated with keywords “oncogene” (KW-0553), “proto-oncogene” (KW-0656) or “tumor suppressor” (KW-0043) (pink), 590 genes from the DISEASES database (Pletscher-Frankild et al., 2015) with confident (i.e., edge weight $>$ 2.75, where the maximum possible edge weight is 5) literature-mined associations with “cancer” (DOID:162) (brown), 713 genes listed in the CGC, version 87 (black), and 324 genes in the CGC with driver statuses due to missense mutations (purple). Ranked gene lists are obtained by applying PertInInt to pan-cancer nonsynonymous mutations (shown as solid lines) and to pan-cancer synonymous mutations (shown as dashed lines). Enrichment for each gold standard set is computed as the ratio between the fraction of gold standard genes in PertInInt’s top ranked genes (i.e., the precision) and the fraction of gold standard genes in the whole set of genes (i.e., the expected precision given a random ordering of genes). All curves converge to an enrichment of 1 by the end of the ranked list of genes (not shown).

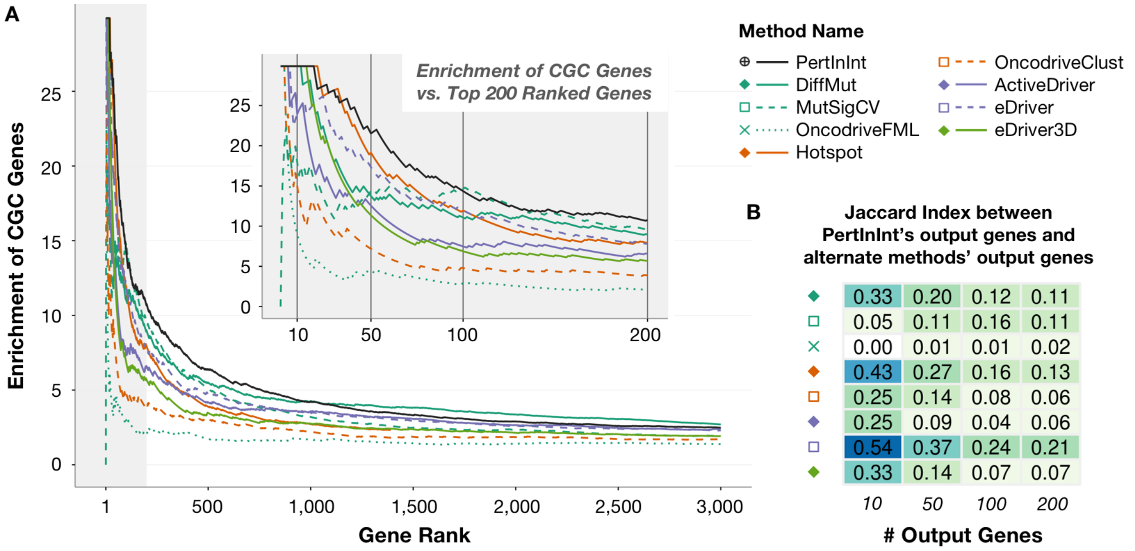


Figure S6: Detection of CGC genes from a pan-cancer dataset excluding highly mutated cancers by PertInInt and alternate methods. ***Related to Figure 4.*** Each driver gene detection method was run on the pan-cancer set of mutations with tumor samples from highly-mutated BLCA, STAD, SKCM, LUAD, LUSC, and ESCA cancers—where there are more than 100 mutations per tumor sample on average—excluded. **(A)** Curves indicate the enrichment for genes in the CGC as we consider an increasing number of output genes for each driver gene detection method. All methods scored at least 3,000 genes except for Hotspot (orange solid line), which only returned 1,397 genes and whose curve ends at that point. The gray shaded area highlights the plot to 200 genes, a closeup of which is shown in the inset. Vertical lines at 10, 50, 100, and 200 ranked genes in the inset correspond to gene set sizes featured in part (B). **(B)** Jaccard Indices (JIs) are calculated between the top 10, 50, 100, and 200 genes output by PertInInt and the corresponding top 10, 50, 100, and 200 genes output by each other method. Lighter colors indicate lower JIs and less overlap between the gene sets.

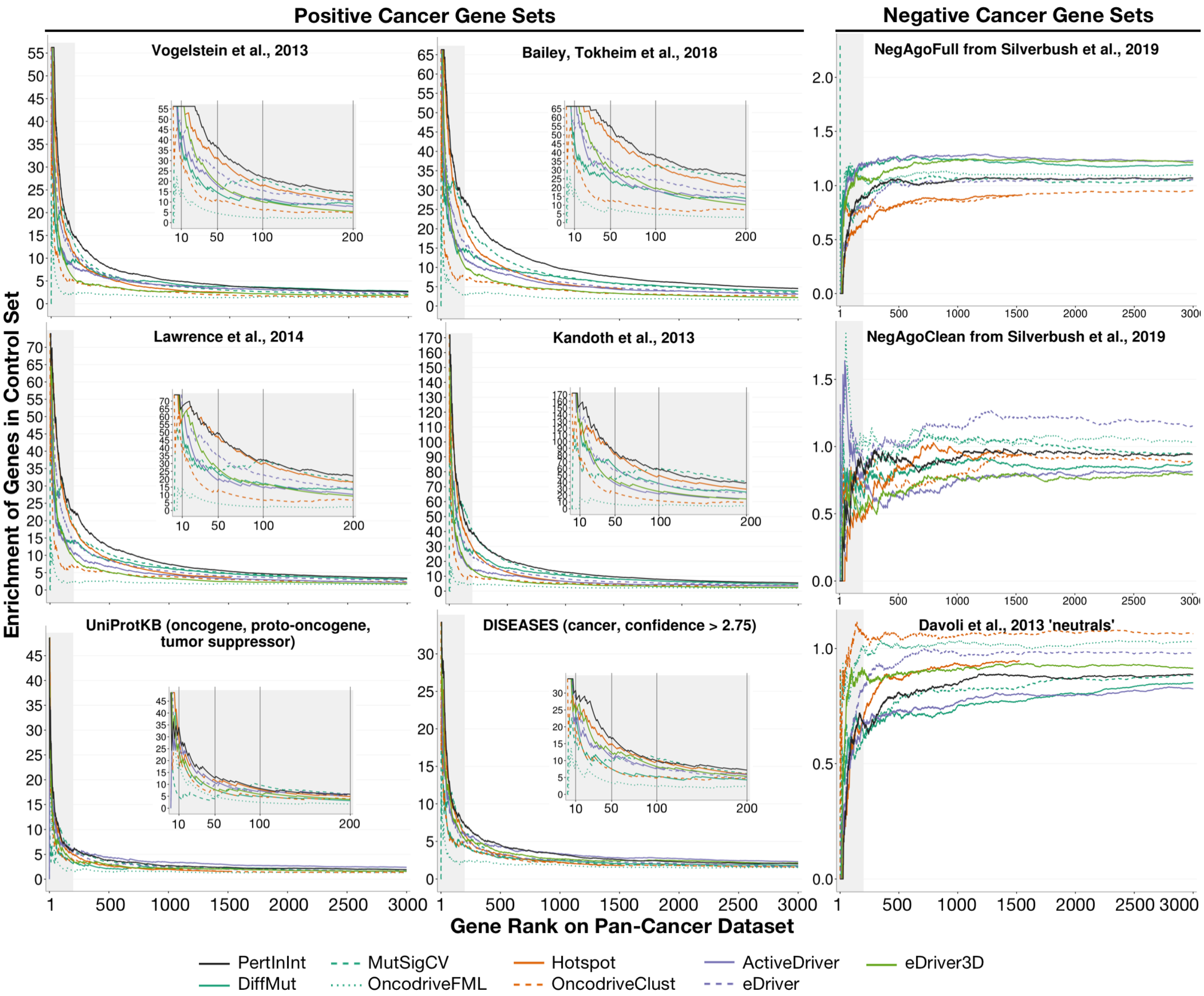


Figure S7: Detection of positive and negative driver genes by PertInInt and alternate methods. ***Related to Figure 4; Table S2; Figure S5.*** Each method was run on the pan-cancer set of mutations as described in STAR Methods. Curves indicate the enrichment for genes in selected positive or negative cancer driver gene sets as we consider an increasing number of output genes for each driver gene detection method. The gray shaded areas highlight each plot to 200 genes, closeups of which are shown in the insets. Positive driver gene sets are described in the caption for Figure S5. Negative driver gene sets include: 8,893 genes that have been proposed to be unlikely to be implicated in cancer and a filtered set of 2,839 of these genes listed in Silverbush et al., 2019, Tables S1D and S1C and 10,303 “neutral” non-driver genes listed in Davoli et al., 2013, Table S2A.

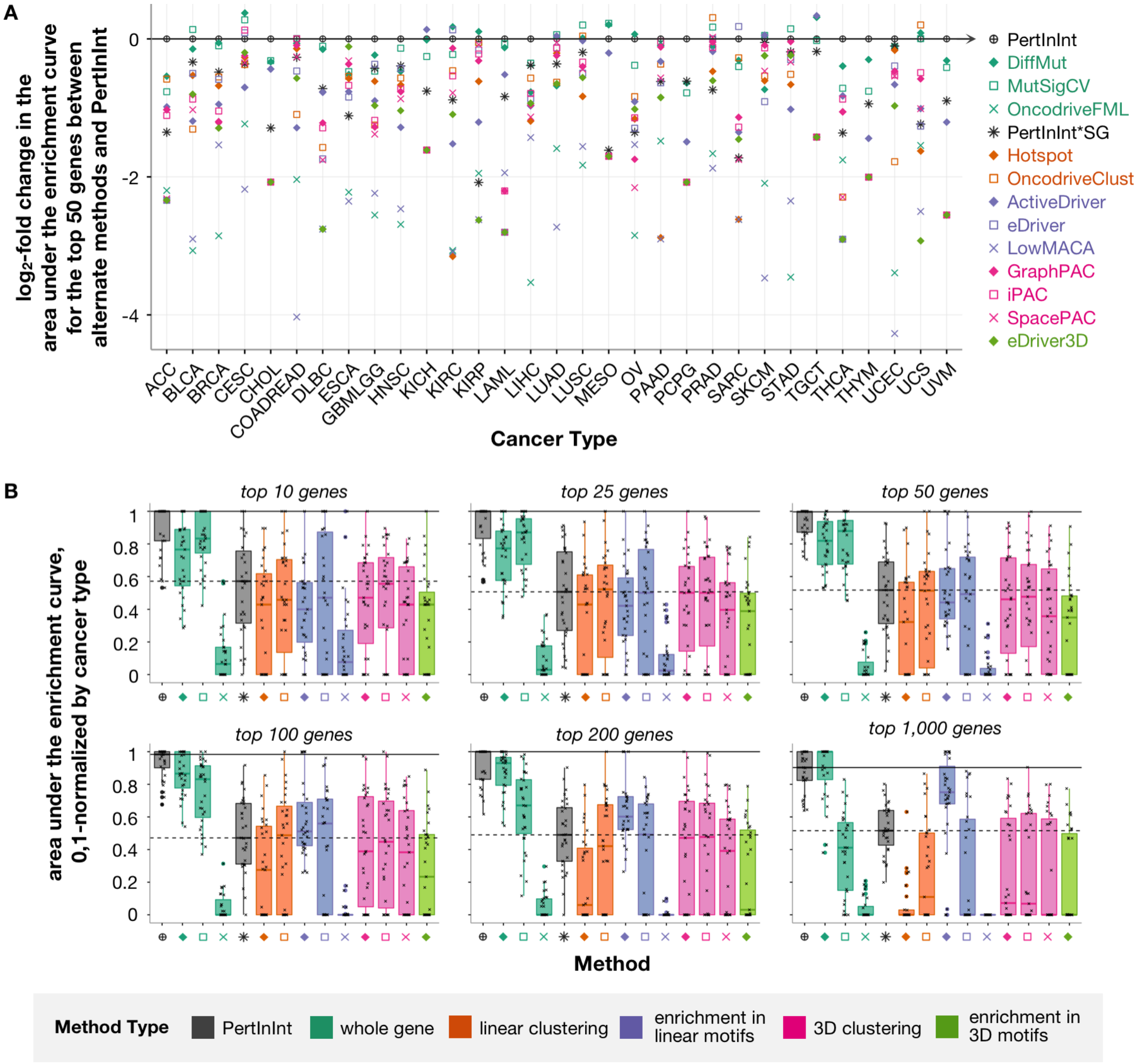


Figure S8: Relative detection of known cancer genes from individual cancer datasets. ***Related to Figure 4.*** **(A)** Log_2-fold change between the area under the enrichment curves for the top 50 genes scored by alternate methods and the top 50 genes scored by PertInInt across individual cancer types. “PertInInt*SG” refers to a version of PertInInt where only subgene resolution tracks are included. PertInInt tends to perform better than the alternate methods, as most of these values are below 0. **(B)** For each cancer type, the areas under the enrichment curves computed for the top 10 (or 25, 50, 100, 200, or 1,000) genes ranked by each driver gene detection method are linearly scaled to fall between 0 and 1. For example, when looking at the top 50 genes ranked by each method when run on SARC mutations, Hotspot has the relatively smallest area under the enrichment curve and thus gets a scaled value of 0, whereas PertInInt has the relatively largest area under the enrichment curve and thus gets a scaled value of 1. Then for each computational method, a box plot of their corresponding values across cancer types is shown. Jittered data points representing different cancer types are overlaid on boxplots. Horizontal solid and dashed lines are drawn at the median relative area under the enrichment curve for PertInInt and PertInInt*SG respectively in each plot. Methods are labeled as in (A).

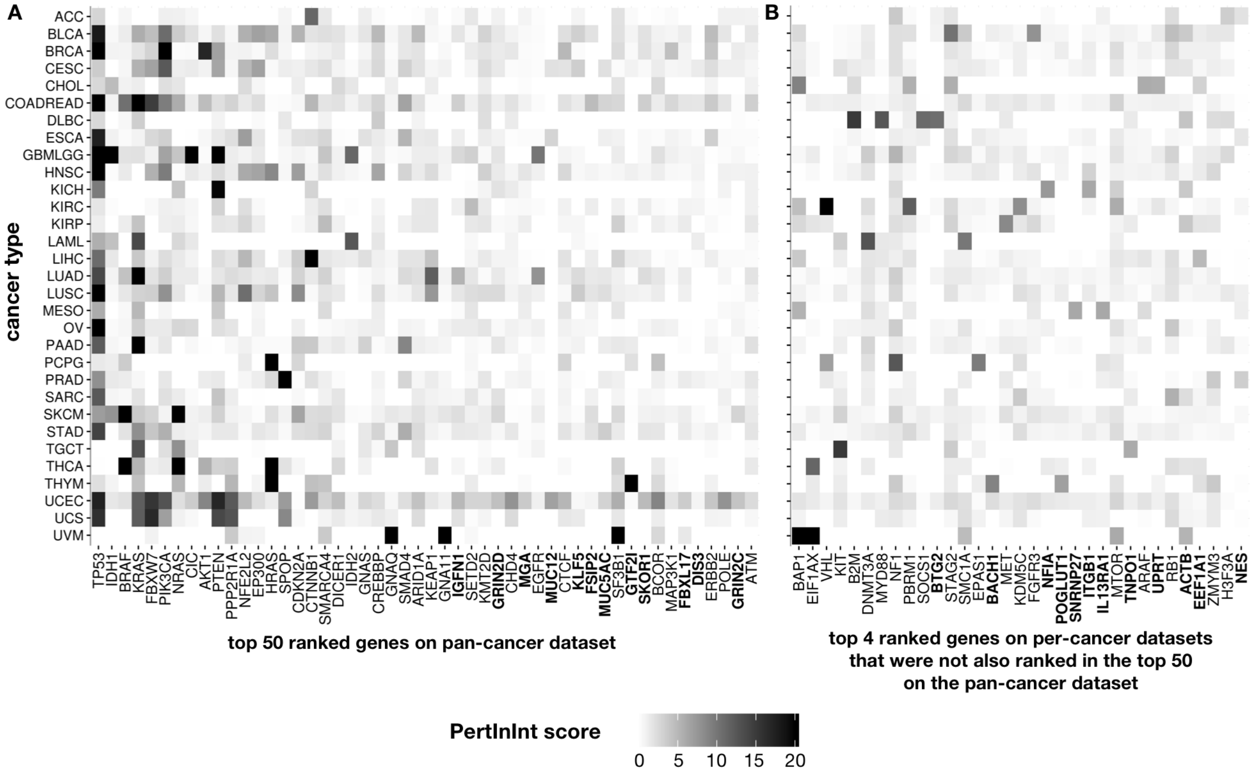


Figure S9: Distinct cancer-relevant genes are highly ranked in individual cancer datasets. ***Related to Figure S8.*** Each entry corresponds to a gene–cancer pair and is colored by the PertInInt score of that gene (genes listed along the x-axis) when run on data from the corresponding cancer type individually (cancer types listed along the y-axis). All PertInInt scores $\geq$20 are recorded as 20 for visualization purposes. Genes that are *not* in the CGC are bolded in the *x*-axis. **(A)** Top 50 genes ranked by PertInInt when run on the pan-cancer dataset. **(B)** Genes that are ranked within the top four by PertInInt when run on individual per-cancer datasets, but are not found in the top 50 genes when PertInInt is run on the pan-cancer dataset.

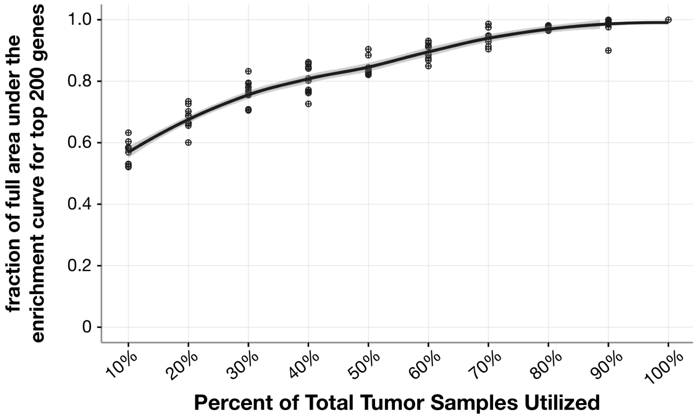


Figure S10: PertInInt's power increases with more tumor samples. ***Related to STAR Methods.*** As a function of the percent (10–100%) of all tumor samples randomly selected from the pan-cancer dataset (x-axis), we show the area under the enrichment curve for the top 200 genes scored by PertInInt when run on each tumor sample subset, normalized by the area under the enrichment curve for PertInInt's top 200 predictions when using all tumor samples (y-axis). Ten random selections of samples are analyzed at each sample size. The solid black line represents the local polynomial regression line of these normalized areas under the enrichment curve with respect to the sample size. PertInInt's ability to recapitulate cancer genes increases with sample size.

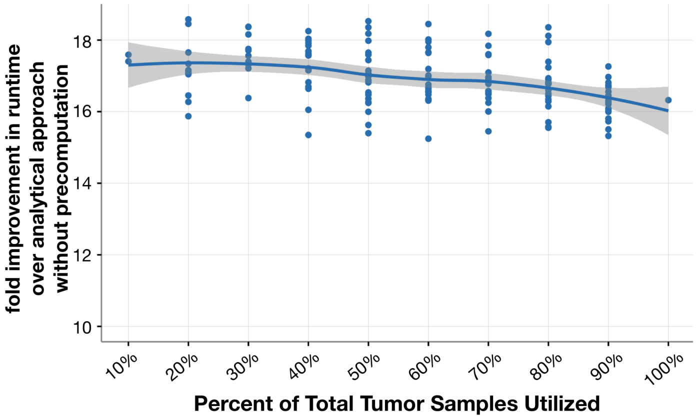


Figure S11: Precomputation enables >16× speedup over basic analytical approach. ***Related to Figure 1; Figure S2; STAR Methods.*** As a function of the percent (10–100%) of all tumor samples randomly selected from the pan-cancer dataset (x-axis), PertInInt's runtime is compared to a baseline version that does not use precomputed expectation and variance estimates to compute Z-scores for each track. Shown on the y-axis is the fold speedup in runtime for ten random selections of samples of each size. The solid blue line represents the local polynomial regression line, with the grey shading showing standard error. These runtime comparisons use only a single track per protein, conservation, as in Figure S2.