ARTICLE

Pathway and network analysis of more than 2500 whole cancer genomes

Matthew A. Reyna [1,2], David Haan[3], Marta Paczkowska[4], Lieven P.C. Verbeke [5,6], Miguel Vazquez [7,8], Abdullah Kahraman [9,10], Sergio Pulido-Tamayo[5,6], Jonathan Barenboim[4], Lina Wadi[4], Priyanka Dhingra[11], Raunak Shrestha [12], Gad Getz [13,14,15,16], Michael S. Lawrence[13,14], Jakob Skou Pedersen [17,18], Mark A. Rubin [11], David A. Wheeler[19], Søren Brunak[20,21], Jose M.G. Izarzugaza[20,21], Ekta Khurana [11], Kathleen Marchal [5,6], Christian von Mering [9], S. Cenk Sahinalp[12,22], Alfonso Valencia[7,23], PCAWG Drivers and Functional Interpretation Working Group, Jüri Reimand [4,24]*, Joshua M. Stuart [3]*, Benjamin J. Raphael [1]* & PCAWG Consortium

The catalog of cancer driver mutations in protein-coding genes has greatly expanded in the past decade. However, non-coding cancer driver mutations are less well-characterized and only a handful of recurrent non-coding mutations, most notably *TERT* promoter mutations, have been reported. Here, as part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which aggregated whole genome sequencing data from 2658 cancer across 38 tumor types, we perform multi-faceted pathway and network analyses of non-coding mutations across 2583 whole cancer genomes from 27 tumor types compiled by the ICGC/TCGA PCAWG project that was motivated by the success of pathway and network analyses in prioritizing rare mutations in protein-coding genes. While few non-coding genomic elements are recurrently mutated in this cohort, we identify 93 genes harboring non-coding mutations that cluster into several modules of interacting proteins. Among these are promoter mutations associated with reduced mRNA expression in *TP53*, *TLE4*, and *TCF4*. We find that biological processes had variable proportions of coding and non-coding mutations, with chromatin remodeling and proliferation pathways altered primarily by coding mutations, while developmental pathways, including Wnt and Notch, altered by both coding and non-coding mutations. RNA splicing is primarily altered by non-coding mutations in this cohort, and samples containing non-coding mutations in well-known RNA splicing factors exhibit similar gene expression signatures as samples with coding mutations in these genes. These analyses contribute a new repertoire of possible cancer genes and mechanisms that are altered by non-coding mutations and offer insights into additional cancer vulnerabilities that can be investigated for potential therapeutic treatments.

A list of author affiliations appears at the end of the paper.

Over the past decade, cancer genome sequencing efforts such as The Cancer Genome Atlas (TCGA) have identified millions of somatic aberrations; however, the annotation and interpretation of these aberrations remain a major challenge[1]. Specifically, while some somatic aberrations occur frequently in specific cancer types, there is a "long tail" of rare aberrations that are difficult to distinguish from random passenger aberrations in modestly sized patient cohorts[2,3]. In many cancers, a significant proportion of patients do not have known driver mutations in protein-coding regions[4], suggesting that additional driver mutations remain undiscovered. The vast majority of known driver mutations affect protein-coding regions. Only a few recurrent non-coding driver mutations, most notably mutations in the *TERT* promoter[5,6], have been identified. In other studies, a genome-wide analysis has identified recurrent mutations in several regulatory elements, and expression quantitative trait loci (eQTLs) analysis has identified non-coding somatic mutations that correlate with gene expression changes in some cancer types[7].

Cancer driver mutations unlock oncogenic properties of cells by altering the activity of hallmark pathways[8]. Accordingly, cancer genes have been shown to cluster in a small number of cellular pathways and interacting subnetworks[3,9]. Consequently, pathway and network analysis has proven useful for implicating infrequently mutated genes as cancer genes based on their pathway membership and physical or regulatory interactions with recurrently mutated genes[10–14]. However, the interactions between coding and non-coding driver mutations in known or novel pathways have not yet been systematically explored.

As part of the Pan-Cancer Analysis of Whole Genomes (PCAWG) project of the International Cancer Genome Consortium (ICGC)[15], we performed pathway and network analysis of coding and non-coding somatic mutations from 2583 tumors from 27 tumor types. The PCAWG consortium curated whole-genome sequencing data from a total of 2658 cancers across 38 tumor types. In the marker paper[15], this work provided the largest collection of uniformly processed cancer whole genomes to date with germline and somatic variants from reanalyzed sequencing data aligned to the human genome (reference build hs37d5) using standardized and highly accurate pipelines. Recent work from the PCAWG project of the ICGC reveals few recurrent non-coding drivers in analyses of individual genes and regulatory regions[16]. Here, we employ seven distinct pathways and network analysis methods and derive consensus sets of *pathway-implicated driver* (PID) genes from the predictions of these methods. Specifically, we identify a consensus set of 93 high-confidence *pathway-implicated driver genes using non-coding variants* (PID-N) and a consensus set of 87 *pathway-implicated driver genes using coding variants* (PID-C). Both sets of PID genes, particularly the PID-N set, contain rarely mutated genes that interact with well-known cancer genes, but were not identified as significantly mutated by single gene tests[16]. In total, 121 novel PID-N and PID-C genes are revealed as promising candidates, expanding the landscape of driver mutations in cancer.

We examined the relative contributions of coding and non-coding mutations in altering biological processes, finding that while chromatin remodeling and some well-known signaling and proliferation pathways are altered primarily by coding mutations, other important cancer pathways, including developmental pathways, such as Wnt and Notch, are altered by both coding and non-coding mutations in PID genes. Intriguingly, we find many non-coding mutations in PID-N genes with roles in RNA splicing, and samples with these non-coding mutations exhibit similar gene expression signatures as samples with well-known coding mutations in RNA splicing factors. Our analysis demonstrates that somatic non-coding mutations in untranslated and in cis regulatory regions constitute a complementary set of genetic perturbations with respect to coding mutations, affect several biological pathways and molecular interaction networks, and should be further investigated for their role in the onset and progression of cancer.

Results

The long tail of coding and non-coding driver mutations. We analyzed the genes targeted by single-nucleotide variants (SNVs) and short insertions and deletions (indels) identified by whole-genome sequencing in the 2538 ICGC PCAWG tumor samples from 27 tumor types. Our pathway and network analyses focused on a subset of 2252 tumors that excluded melanomas and lymphomas due to their atypical distributions of mutations in regulatory regions[17]. We utilized the pan-cancer driver *p*-values of single protein-coding and non-coding elements from the PCAWG Drivers and Functional Interpretation Working Group analysis[16], including exons, promoters, untranslated regions (5′ UTR and 3′ UTR), and enhancers. This analysis integrates predictions from 16 driver discovery methods, resulting in consensus driver *p*-values for coding and non-coding elements; see ref. [16] for further details. The *p*-values of individual genes and non-coding elements indicate their statistical significance as drivers, according to diverse methods that account for positive selection, functional impact of mutations, regional mutation rates, and mutational processes and signatures[16]. Among protein-coding driver *p*-values of the pan-cancer cohort, 75 genes were statistically significant (FDR < 0.1; Supplementary Fig. 1) and an additional 7 genes were observed at near-significant levels (0.1 ≤ FDR < 0.25). These numbers are consistent with previous reports of a "long tail" of driver genes with few highly mutated genes and many genes with infrequent mutations across cancer types[2,18]. Non-coding mutations exhibit a similar long-tail distribution with even fewer significant genes (eight genes at FDR < 0.1 and two genes at 0.1 ≤ FDR < 0.25). No single gene has both significant or near-significant coding and non-coding driver *p*-values (FDR < 0.25), suggesting that non-coding mutations target a complementary set of genes as coding mutations.

Earlier studies have demonstrated that proteins harboring coding driver mutations interact with each other in molecular pathways and networks significantly more frequently than expected by chance[2,3,9–11,13]. We observed significant numbers of interactions between both significantly mutated coding and/or non-coding elements, suggesting that the pathway and network methods may be useful in prioritizing rare driver events that are not significant by single-element analyses (Supplementary Fig. 2; Coding and non-coding mutations cluster on networks in Supplementary file).

Pathway and network analysis of potential driver mutations. We performed a comprehensive pathway and network analysis of cancer drivers using the single-element driver *p*-values computed by the PCAWG Drivers and Functional Interpretation Working Group[16] as input. We applied seven distinct pathways and network methods (ActivePathways[19], CanIsoNet[20], Hierarchical HotNet[21], a hypergeometric analysis (Vazquez), an induced subnetwork analysis (Reyna and Raphael, in preparation), NBDI[22], and SSA-ME[23]) that each leverage information from molecular pathways or protein interaction networks (Fig. 1, Methods section) to amplify weak signals in the single-element analysis. All methods were calibrated on randomized data (Pathway and network methods in Supplementary file).

Since the prioritization of non-coding somatic mutations in cancer is not yet a solved problem, it was difficult to know in advance which analysis methodologies, if any, would be best
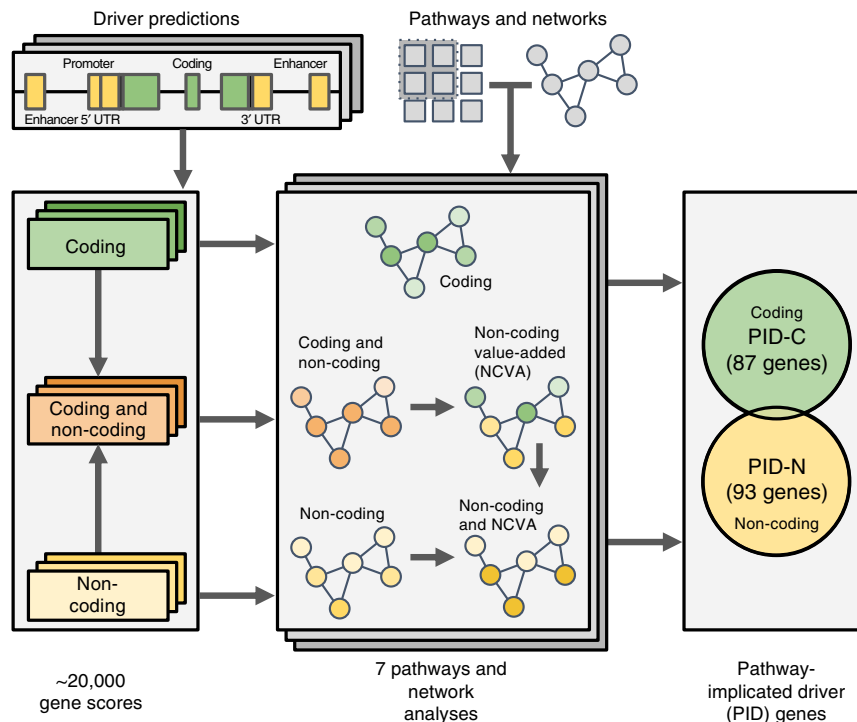
Fig. 1 Overview of the pathway and network analysis approach. Coding, non-coding, and combined gene scores were derived for each gene by aggregating driver *p*-values from the PCAWG driver predictions in individual elements, including annotated coding and non-coding elements (promoter, 5′ UTR, 3′ UTR, and enhancer). These gene scores were input to five network analysis algorithms (CanIsoNet[20], Hierarchical HotNet[21], an induced subnetwork analysis (Reyna and Raphael, in preparation), NBDI[22], and SSA-ME[23]), which utilize multiple protein–protein interaction networks, and to two pathway analysis algorithms (ActivePathways[19] and a hypergeometric analysis (Vazquez)), which utilize multiple pathway/gene-set databases. We defined a non-coding value-added (NCVA) procedure to determine genes whose non-coding scores contribute significantly to the results of the combined coding and non-coding analysis, where NCVA results for a method augment its results on non-coding data. We defined a consensus procedure to combine significant pathways and networks identified by these seven algorithms. The 87 pathway-implicated driver genes with coding variants (PID-C) are the set of genes reported by a majority (≥4/7) of methods on the coding data. The 93 pathway-implicated driver genes with non-coding variants (PID-N) are the set of genes reported by a majority of methods on non-coding data or in their NCVA results. Only five genes (*CTNNB1, DDX3X, SF3B1, TGFBR2,* and *TP53*) are both PID-C and PID-N genes.

suited to distinguish drivers from passengers by aggregating weak signals across pathways or networks. Thus, we formed a consensus of multiple methods, following the "wisdom of crowds" ensemble approach of machine learning[24] to improve the specificity of the results. We included methods that used different sources of pathway or network information and different prioritization criteria (see Supplementary Data 1 for a complete list). Each method nominated genes, and consensus sets of genes with possible coding and non-coding driver mutations were defined as the genes found by at least four of the seven methods (Supplementary Data 2–5). We use the term *pathway-implicated driver (PID) genes* to describe these candidate driver genes.

One potential concern with a consensus procedure is that the results may be dominated by a few highly correlated methods. Our pathway or network analysis methods use varied sources of prior knowledge (i.e., pathway databases or interaction networks), and input data (e.g., driver *p*-values, point mutations, and/or gene expression), and rely on different techniques to integrate these data sources. We found only a modest overlap between the output of the seven methods (Method results comparison and Consensus procedure in Supplementary file; Supplementary Data 6–8), suggesting a non-uniform weighting of the consensus to mitigate the influence of redundant methods was not necessary.

Using coding mutations alone, we identify a set of 87 pathway-implicated driver genes with coding variants (PID-C genes). The 87 PID-C genes (Supplementary Data 2, Supplementary Fig. 6a) include 68 previously identified cancer genes as catalogued by the

COSMIC Cancer Gene Census (CGC) database (v83, 699 genes from Tier 1 and Tier 2)[25] (2.98 genes expected; Fisher's exact test $p = 3.57 \times 10^{-83}$; Fig. 2a, c; Supplementary Fig. 7a). The PID-C genes have significantly higher coding gene scores than non-PID-C genes (rank-sum test $p = 1.72 \times 10^{-58}$; median rank 48 of PID-C genes), and each of the 87 PID-C genes improves the score of its network neighborhood (19.7 genes expected; $p < 10^{-6}$; Supplementary Data 9). This network neighborhood analysis shows that PID-C genes are not implicated solely by their network neighbors[14], but themselves contribute significantly to their discovery by the pathway and network methods. The 87 PID-C genes also include 31 genes that are not statistically significant (FDR > 0.1) in the PCAWG Drivers and Functional Interpretation Working Group analysis; Fig. 2a, c; Supplementary Figs. 8a, 9), illustrating that the network neighborhoods can nominate genes with infrequent mutations, i.e., those in the "long tail", as possible driver genes. Interestingly, 13 of these 31 genes with FDR > 0.1 are also known drivers according to the CGC database (3.0 genes expected; Fisher's exact test $p = 2.1 \times 10^{-14}$). Thus, the consensus pathway and network analysis recovers many known protein-coding driver mutations and identifies additional possible drivers that are infrequently mutated and thus remain below the statistical significance threshold of gene-specific driver analyses.

Using non-coding mutations alone, we identify a set of 62 genes using our consensus pathway and network analysis, resulting in fewer genes than our analysis with coding mutations. However, when we performed a joint analysis of coding and
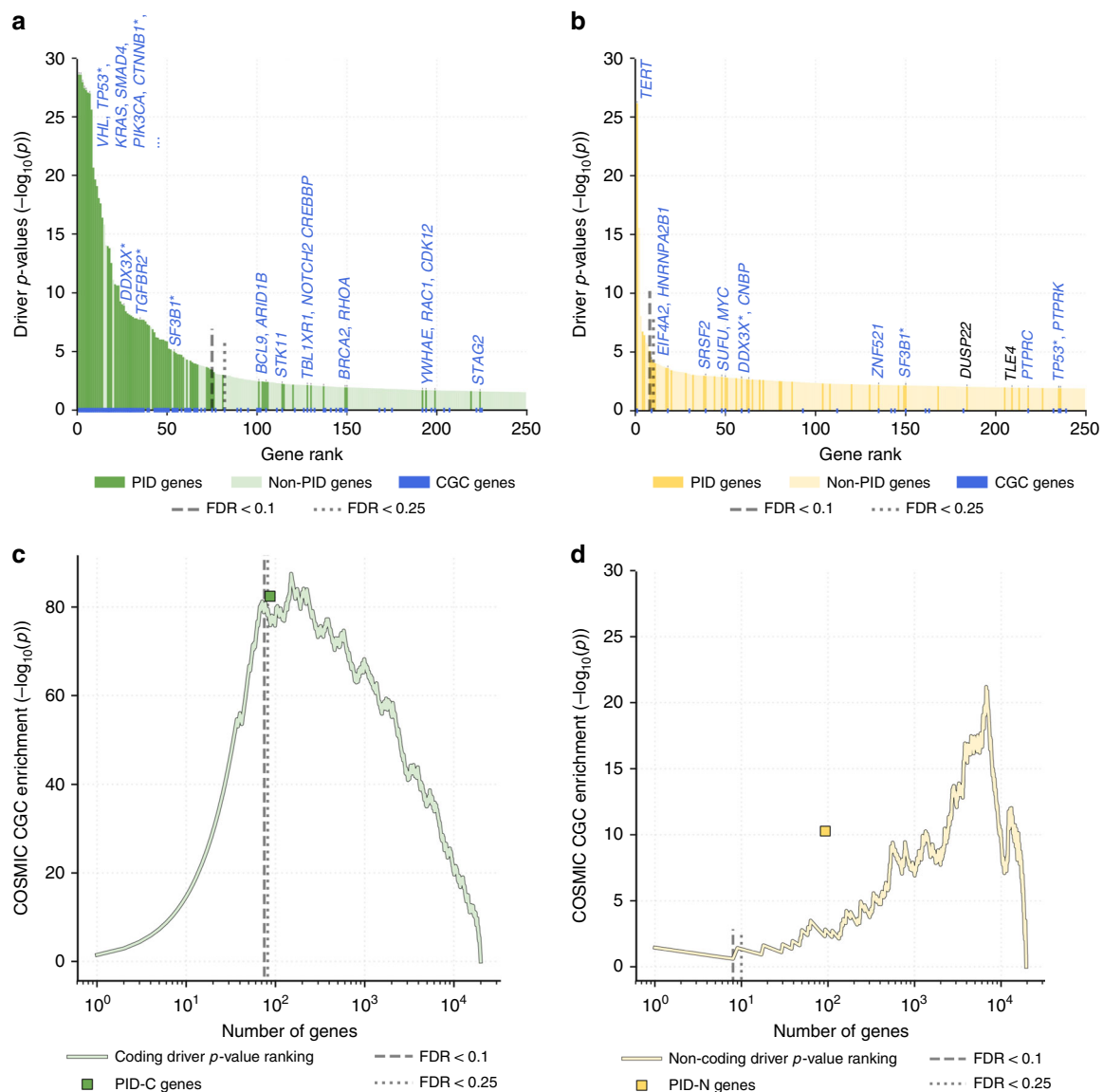
Fig. 2 Pathway and driver analysis identifies driver genes in the long tail of the driver *p*-values for coding and non-coding mutations. a Pathway and network methods identify significant coding driver mutations. Driver *p*-values on protein-coding elements for 250 genes with most significant coding driver *p*-values; dashed and dotted lines indicate FDR = 0.1 and 0.25, respectively. Dark green bars are PID-C genes, and light green bars are non-PID-C genes. Blue squares below the *x*-axis indicate COSMIC Cancer Gene (CGC) Census genes. In total, 31 of 87 PID-C genes have coding driver *p*-values with FDR > 0.1. Several PID-C genes are labeled, including all CGC genes with coding FDR > 0.1. Overlap between PID-C and PID-N genes is indicated with asterisks. Source data are provided as a Source Data file. **b** Pathway and network methods identify rare non-coding driver mutations. Driver *p*-values on non-coding elements (promoter, 5′ UTR, and 3′ UTR of gene) for 250 genes with most significant non-coding driver *p*-values; dashed and dotted lines indicate FDR = 0.1 and 0.25, respectively. Dark yellow bars are PID-N genes, and light yellow bars are non-PID-N genes. Blue squares are as above. In total, 3 (*TERT*, *HES1*, and *TOB1*) of 93 PID-N genes have non-coding driver *p*-values with FDR ≤ 0.1, while 90 have FDR > 0.1 . Several PID-N genes are labeled, including PID-N genes with significant in cis gene expression changes (see Fig. 3) and all PID-N genes with non-coding FDR > 0.25. Overlap between PID-C and PID-N genes is indicated with asterisks. Source data are provided as a Source Data file. **c** Statistical significance of overlap between top-ranked genes according to coding driver *p*-values and PID-C genes with CGC genes. Fisher's exact test *p*-values and driver FDR thresholds of 0.1 and 0.25 are highlighted. Green squares indicate overlap between PID-C genes and CGC genes. Source data are provided as a Source Data file. **d** Statistical significance of overlap of genes ranked by driver *p*-values on non-coding (promoter, 5′ UTR, 3′ UTR) elements and CGC genes. Driver FDR thresholds of 0.1 and 0.25 are highlighted. Yellow square indicates overlap between PID-N genes and CGC genes. Source data are provided as a Source Data file.

non-coding mutations, we found that the much stronger signal in coding mutations dominated the combined signal in coding and non-coding mutations. To increase sensitivity to detect contributions of non-coding mutations, we devised a "non-coding value-added" (NCVA) procedure (Fig. 1; Supplementary Fig. 3; Non-coding value-added (NCVA) procedure in the Methods section). Our NCVA procedure asks if the coding mutations enhance the discovery of potential non-coding driver genes

beyond what is found with only the non-coding mutations. This procedure identified an additional set of 31 genes that, when merged with the 62 genes found with non-coding mutations alone, resulted in a set of 93 pathway-implicated driver genes with non-coding variants (PID-N) (Supplementary Fig. 4, Consensus results in the Methods section). PID-N genes appear as a robust and biologically relevant set, unbiased by any particular mutational process reflecting a particular carcinogen or DNA

damage processes (Supplementary Fig. 5, Mutational signatures in the Methods section).

The 93 PID-N genes (Supplementary Data 3, Supplementary Fig. 6b) include 19 previously identified cancer genes according to the COSMIC Cancer Gene Census (CGC) database a significant enrichment over the 3.2 genes expected ($p = 5.3 \times 10^{-11}$; Fisher's exact test, Fig. 2b, d; Supplementary Figs. 7b, c). Excluding the eight genes with individually significant non-coding elements in the PCAWG Drivers and Functional Interpretation Working Group analysis[16], 19 genes are both PID-N genes and CGC genes, a significant enrichment over the 3.1 genes expected ($p = 5.3 \times 10^{-11}$; Fisher's exact test), suggesting that non-coding mutations may alter genes with recurrent coding or structural variants in some samples. The PID-N genes have significantly higher non-coding gene scores than non-PID-N genes (rank-sum test $p = 1.47 \times 10^{-58}$; median rank 165 of PID-N genes), and 92/93 PID-N genes (except for HIST1H2BO) improve the scores of their network neighborhoods (28.5 genes expected; $p < 10^{-6}$; Supplementary Data 10). This shows that PID-N genes are not implicated solely by their network neighbors[14]. The vast majority of PID-N genes (90/93, including the 19 CGC genes) are distinct from the PCAWG Drivers and Functional Interpretation Working Group analysis (Fig. 2b; Supplementary Figs. 8b, 9), with only three genes in common: TERT, HES1, and TOB1. Of these three, only TERT is recorded as a known cancer gene in the CGC database. Moreover, the 93 PID-N genes are more strongly enriched ($p = 5.3 \times 10^{-11}$; Fisher's exact test) for COSMIC CGC genes than the 93 genes with the smallest non-coding driver p-values of promoters, 5′ UTRs, or 3′ UTRs ($p = 4.8 \times 10^{-3}$; Fisher's exact test). Thus, our consensus procedure of the pathway and network analyses appreciably augments the significantly mutated elements in the PCAWG Drivers and Functional Interpretation Working Group results[16].

Taken together, the PID-C and PID-N genes contain an additional 121 genes over what was found in the PCAWG Drivers and Functional Interpretation Working Group analysis, including 90 new possible non-coding drivers (Consensus Results in the Methods section). In total, non-coding mutations in PID-N genes cover an additional 151 samples (9.1% of samples) than PID-C genes. We found that most coding mutations in PID-C genes and most non-coding mutations in PID-N genes are clonal (median > 80% for both PID-C and PID-N genes[26]). In addition, the overwhelming majority of the PID-N genes were distinct from PID-C genes (Supplementary Fig. 4) with only five genes in common: CTNNB1, DDX3X, SF3B1, TGFBR2, and TP53. While this suggests that coding and non-coding driver mutations occur in largely distinct sets of cancer genes, we show below that both types of mutations affect genes underlying many of the same hallmark cancer processes.

Impact of non-coding mutations on gene expression. Non-coding mutations may act by altering transcription factor-binding sites or other types of regulatory sites. Thus, we evaluated whether non-coding mutations in PID-N genes were associated with in cis expression changes in the same gene. We found that five PID-N genes (FDR < 0.3) showed significant in cis expression correlations out of the 90 that could be tested using RNA-Seq data (Fig. 3; Supplementary Fig. 10; Supplementary Data 11–16). In contrast, 34 out of 87 PID-C genes exhibited significant or near-significant in cis expression correlations (FDR < 0.3) (Supplementary Data 17, 18).

Unsurprisingly, the most significant in cis expression correlation for a PID-N gene is the correlation between TERT promoter mutations and increased expression, which we find in 11 Thy-AdenoCA tumors ($p = 1.3 \times 10^{-10}$, FDR = 3.2×10^{-9}; Wilcoxon

rank-sum test), 11 CNS-Oligo tumors ($p = 6.8 \times 10^{-3}$, FDR = 9.7×10^{-2}; Wilcoxon rank-sum test), and 22 CNS-GBM tumors (Wilcoxon rank-sum test $p = 2.3 \times 10^{-2}$, FDR = 0.19; Wilcoxon rank-sum test) (Supplementary Fig. 8), consistent with previous reports[5,6,27]. Note that these associations were limited by the unavailability of RNA expression data for some samples with TERT mutations as well as the low-sequencing coverage in promoter regions that limited the detection of TERT promoter mutations. The PCAWG Drivers and Functional Interpretation Working Group investigated the latter issue for two mutation hotspot sites in the TERT promoter, and estimated that 216 mutations in these sites were likely not called[16], a large underrepresentation relative to the total of 97 samples with TERT promoter mutations (71 samples with expression data) in our analyses.

We found significant in cis expression correlations in four other PID-N genes: TP53, TLE4, TCF4, and DUSP22 (Fig. 3, Supplementary Fig. 10). TP53 shows significantly reduced expression ($p = 1.0 \times 10^{-3}$; FDR = 8.7×10^{-2}; Wilcoxon rank-sum test) across six tumors with TP53 promoter mutations from six different tumor types (Fig. 3a; Supplementary Fig. 10). The reduced expression of mutated samples is consistent with TP53's well-known role as a tumor suppressor gene, and links between TP53 promoter methylation and expression have been previously investigated[28]. This expression change was also described in the study by the PCAWG Drivers and Functional Interpretation Working Group[16]. TLE4 shows significantly reduced expression ($p = 1.7 \times 10^{-2}$; FDR = 0.20; Wilcoxon rank-sum test) in three Liver-HCC tumors with TLE4 promoter mutations (Fig. 3b; Supplementary Fig. 10). TLE4 is a transcriptional co-repressor that binds to several transcription factors[29], and TLE4 functions as a tumor suppressor gene in acute myeloid lymphoma through its interactions with Wnt signaling[30]. Furthermore, in an acute myeloid lymphoma cell line, TLE4 knockdown increased cell division rates, while forced TLE4 expression induced apoptosis[31]. However, the role of TLE4 in solid tumors is not well understood. TCF4 shows significantly reduced expression ($p = 3.4 \times 10^{-2}$; FDR = 0.27; Wilcoxon rank-sum test) in three Lung-SCC tumors with TCF4 promoter mutations (Fig. 3c; Supplementary Fig. 10). TCF4 is part of the TCF4/β-catenin complex and encodes a transcription factor that is downstream of the Wnt signaling pathway. Low TCF4 expression has been observed in Lung-SCC tumors[32]. Finally, DUSP22 has significantly reduced expressed ($p = 6.3 \times 10^{-3}$; FDR = 0.024; Wilcoxon rank-sum test) in five Lung-AdenoCA patients with DUSP22 3′ UTR mutations and significantly over-expressed ($p = 7.8 \times 10^{-4}$; FDR = 0.075; Wilcoxon rank-sum test) in three Lung-AdenoCA patients with DUSP22 5′ UTR mutations. These UTR mutations were mutually exclusive. DUSP22 encodes a phosphatase signaling protein, and was recently proposed to be a tumor suppressor in lymphoma[33].

While these gene expression correlations provide additional support for a subset of PID-N genes, the variant allele frequency of a mutation and the copy number of the gene are additional covariates for gene expression. We found that these covariates did not play a role in of the correlations that we identified: the majority of mutations in each PID gene were clonal (Supplementary Fig. 11) and copy-number changes did not affect the expression correlations for the five PID-N genes described above (Fig. 3; Supplementary Fig. 10). In addition, the low number of PID-N genes exhibiting associated gene expression changes is explained by the low number of samples with mutations in PID-N genes, the uneven availability of expression data across the tumor types, and decreased sequence coverage in promoter regions[16]. These issues further reduced the number of samples with non-coding mutations and RNA expression, limiting the power of in cis gene expression correlation analysis.
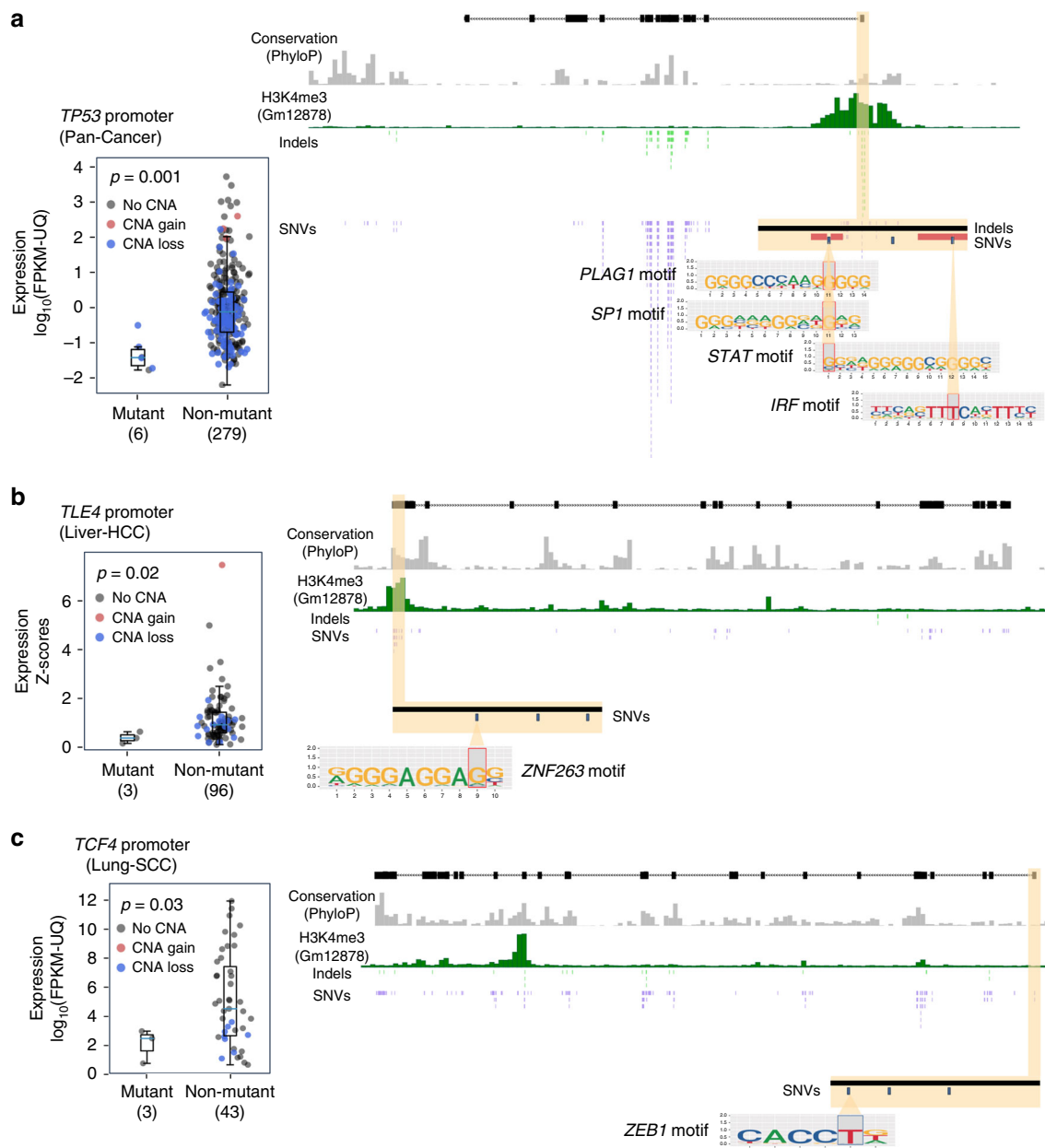
Fig. 3 Gene expression changes are correlated with mutations in PID-N genes. Evolutionary conservation of genomic elements estimated with PhyloP are shown as gray features. H3 histone lysine 4 tri-methylation sites (H3K4me3) measured in GM12878 HapMap B-lymphocytes cell lines are highlighted in the green track, indicating active promoter regions near transcription start sites[49]. Boxplot center lines show the median, boxplot bounds show the first quartile Q1 and the third quartile Q3, and whiskers show 1.5 (Q3–Q1) below and above Q1 and Q3, respectively. **a** TP53 promoter. *TP53* coding and non-coding genomic loci with zoomed-in view of the *TP53* promoter region. *TP53* promoter mutations (six mutations in Biliary-AdenoCA, ColoRect-AdenoCA, Kidney-ChRCC, Lung-SCC, Ovary-AdenoCA, and Panc-AdenoCA cancer types) correlate significantly (Wilcoxon rank-sum test $p = 0.001$, FDR = 0.087) with reduced *TP53* gene expression, where FPKM-UQ is upper quartile normalized fragments per kilobase million. Samples with copy-number gains and losses in the *TP53* promoter region are annotated in red and blue, respectively. Two of the six *TP53* promoter mutations overlap with transcription factor-binding sites (with one mutation matching three motifs). Source data are provided as a Source Data file. **b** TLE4 promoter. *TLE4* coding and non-coding genomic loci with zoomed-in view of *TLE4* promoter region. *TLE4* promoter mutations in Liver-HCC samples (three mutations) correlate (Wilcoxon rank-sum test $p = 0.02$, FDR = 0.2) with lower *TLE4* gene expression. Samples with copy-number gains and losses annotated in red and blue, respectively. One of the three *TLE4* promoter mutations has a transcription factor-binding site for *ZNF263*. Source data are provided as a Source Data file. **c** TCF4 promoter. *TCF4* coding and non-coding genomic loci with zoomed-in view of *TCF4* promoter region. *TCF4* promoter mutations in Lung-SCC samples (three mutations) correlate (Wilcoxon rank-sum test $p = 0.03$, FDR = 0.27) with lower *TCF4* gene expression. Samples with copy-number gains and losses annotated in red and blue, respectively. One of the three *TCF4* promoter mutations has a transcription factor-binding site for *ZEB1*. Source data are provided as a Source Data file.

Modular organization of coding and non-coding mutations. We identified specific protein–protein interaction subnetworks and biological pathways that were altered by coding mutations, non-coding mutations, or a combination of both types of mutations. We found significantly more interactions between PID-C genes that expected by chance using a node-degree preserving permutation test (64 interactions observed vs. 40 interactions expected, $p < 10^{-6}$), a near-significant number of

interactions between PID-N genes (18 vs. 12 expected, $p = 6.8 \times 10^{-2}$), and significantly more interactions between both PID-C and PID-N genes (67 vs. 40 expected, $p = 6 \times 10^{-4}$), demonstrating an interplay between coding and non-coding mutations on physical protein–protein interaction networks (Network annotation in the Methods section). We organized the interacting subnetworks involving PID-C and PID-N genes into five biological processes: core drivers, chromatin organization, cell proliferation, development, and RNA splicing (Fig. 4a). While the high frequency of molecular interactions between PID-C and PID-N genes is expected since such interactions were used as a signal in pathway and network methods, the organization of these interactions illustrates the relative contributions of coding and non-coding mutations in individual subnetworks.

We further characterized the molecular pathways enriched among our PID-C and PID-N using the g:Profiler web server[34] (Fig. 4b; Supplementary Fig. 12, Supplementary Data 19–24, Pathway annotation in the Methods section). Overall, 63 pathways were enriched for PID-C genes and 13 pathways were enriched for PID-N genes (FDR < 10^{-6}). Since our gene-prioritization methods use pathway databases and interaction networks as prior knowledge, it is not surprising that PID-C and PID-N genes are enriched in multiple molecular pathways. However, the enrichment results provide clues about the modular organization of the pathways and allow us to assess the relative contributions of coding and non-coding mutations in each pathway.

We further grouped these molecular pathways into 29 modules using overlaps between annotated pathways recorded in the pathway enrichment map (Supplementary Fig. 12). For each enriched module, we examined whether PID-C, PID-N, or both types of genes were responsible for the observed enrichment. This produced a clustering of modules and PID genes into four biological processes: chromatin organization, cell proliferation, development, and RNA splicing (Fig. 4b).

We found that pathways in the chromatin and cell proliferation processes—including chromatin remodeling and organization, histone modification, apoptotic signaling, signal transduction, Ras signaling, and cell growth—were altered primarily by coding mutations in PID-C genes. This is not surprising as these pathways contain many well-known cancer genes, such as TP53, KRAS, BRAF, cyclin-dependent kinase inhibitors, EGFR, PTEN, and RB1.

At the same time, we found that multiple signaling pathways include significant numbers of both PID-C and PID-N genes, suggesting that non-coding mutations provide an alternative to coding mutations in disrupting these pathways. In particular, the Wnt signaling pathway (FDR = 6.8×10^{-13}), which was predominantly targeted by coding mutations, was also targeted by non-coding mutations in several PID-N genes, including TERT (103 mutations), HNF1A/B (24 mutations), TLE4 (32 mutations), TCF4 (93 mutations), and CTNNB1 (17 mutations) (Supplementary Fig. 13a). The Notch signaling pathway (FDR = 6.8×10^{-7}) was associated with comparable numbers of PID-C and PID-N genes, including the PID-N genes JAG1 and MIB1 that encode ligands and the PID-N transcription factors ACL1, HES1, and HNF1B (66 non-coding mutations in total) (Supplementary Fig. 13b). The TGF-β signaling pathway (FDR = 3.2×10^{-7}) also contained both PID-C and PID-N genes, including the PID-N genes HES1, HNF1A/B, HSPA5, MEF2C, and the genes TGFBR2 and CTNNB1, which are both PID-C and PID-N genes (214 coding mutations and 166 non-coding mutations).

a
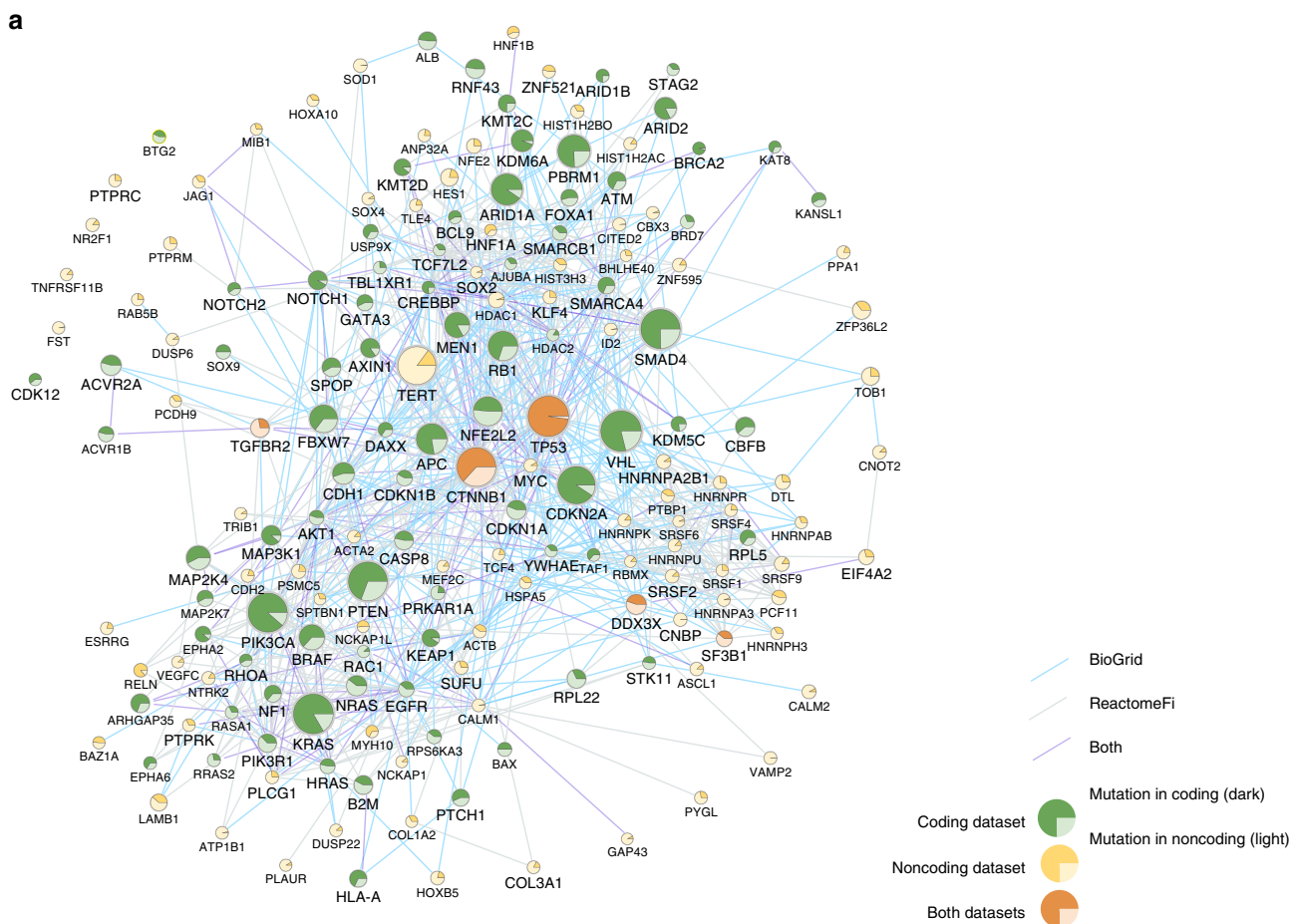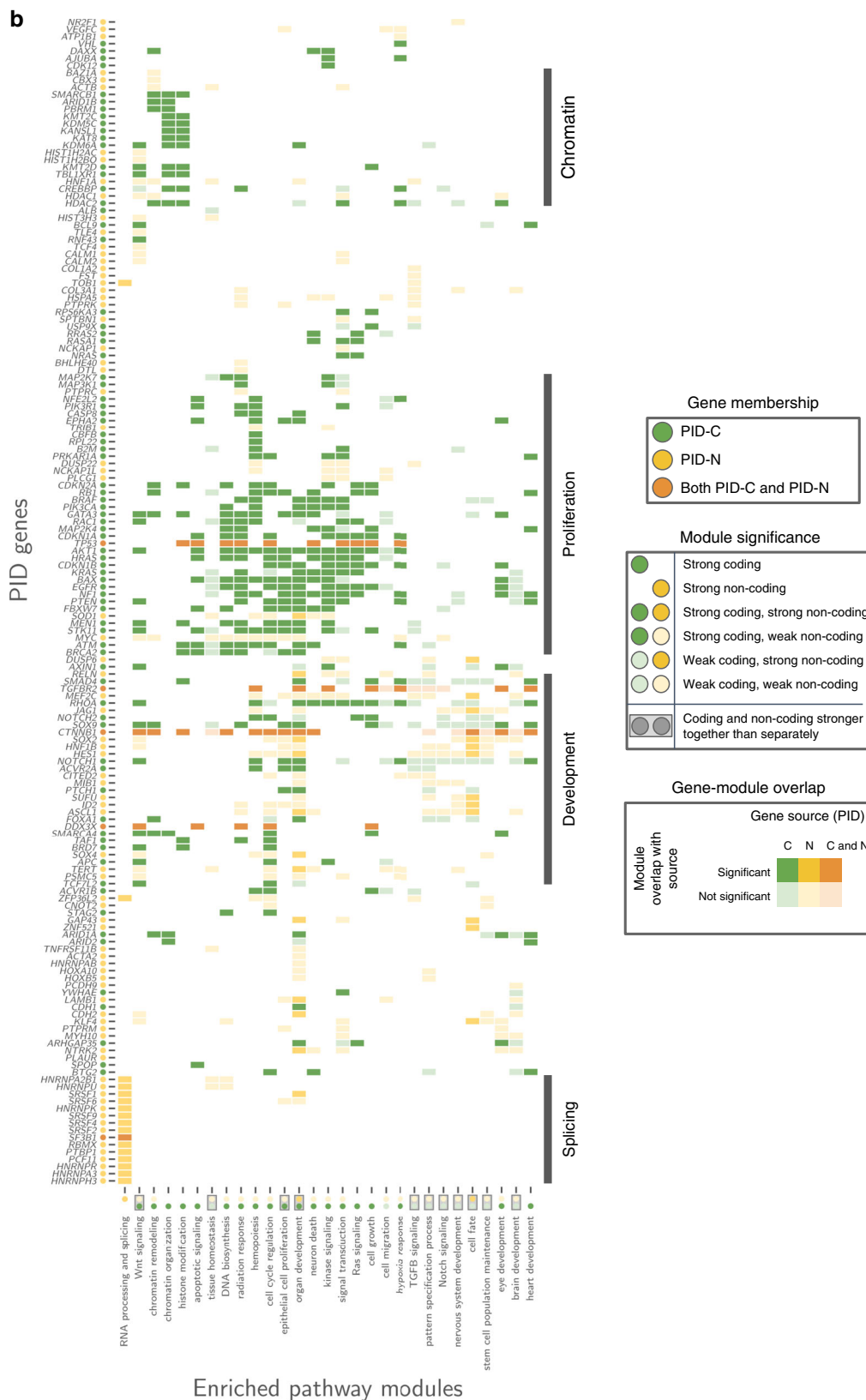
We found that several developmental processes were altered by significant numbers of both PID-C and PID-N genes. Cell fate determination (FDR = 2.0×10^{-7}) was predominantly affected by non-coding mutations in the PID-N genes *DUSP6*, *MEF2C*, *JAG1*, *SOX2*, *HES1*, *ACL1*, *ID2*, *SUFU,* and *KLF4* (total 191 non-coding

mutations), but also by coding mutations in PID-C genes *BRAF*, *GATA3*, and *NOTCH1/2*. Pathways related to nervous system development (FDR = 5.8×10^{-8}) were enriched for the PID-N genes *ASCL1*, *CTNNB1*, *ID2*, *SUFU*, and *TERT* that have known roles in cancer[35,36], complementing the PID-C genes *NOTCH1*,

Fig. 4 Pathway and network modules containing PID-C and PID-N genes. a Network of functional interactions between PID-C and PID-N genes. Nodes represent PID-C and PID-N genes and edges show functional interactions from the ReactomeFI network (gray), physical protein–protein interactions from the BioGRID network (blue), or interactions recorded in both networks (purple). Node color indicates PID-C genes (green), PID-N genes (yellow), or both PID-C and PID-N genes (orange); node size is proportional to the score of the gene; and the pie chart diagram in each node represents the relative proportions of coding and non-coding mutations associated with the corresponding gene. Dotted outlines indicate clusters of genes with roles in chromatin organization and cell proliferation, which predominantly contain PID-C genes; development, which includes comparable amounts of PID-C and PID-N genes; and RNA splicing, which contains PID-N genes. A core cluster of genes with many known drivers is also indicated. **b** Pathway modules containing PID-C and PID-N genes. Each row in the matrix corresponds to a PID-C or PID-N gene, and each column in the matrix corresponds to a pathway module enriched in PID-C and/or PID-N genes (see the Methods section). A filled entry indicates a gene (row) that belongs to one or more pathways (column) colored according to gene membership in PID-C genes (green), PID-N genes (yellow), or both PID-C and PID-N genes (orange). A dark colored entry indicates that a PID-C or PID-N gene belongs to a pathway that is significantly enriched for PID-C or PID-N genes, respectively. A lightly colored entry indicates that a PID-C or PID-N gene belongs to a pathway that is significantly enriched for the union of PID-C and PID-N genes, but not for PID-C or PID-N genes separately. Enrichments are summarized by circles adjacent each pathway module name and PID gene name. Boxed circles indicate that a pathway module contains a pathway that is significantly more enriched for the union of the PID-C and PID-N genes than the PID-C and PID-N results separately. The enriched modules and PID genes are clustered into four biological processes: chromatin, development, proliferation, and RNA splicing as indicated.

PTEN, and *RHOA* that also have known cancer roles. The pattern specification process (FDR $= 8.8 \times 10^{-8}$) was also affected by both coding and non-coding mutations, including the PID-N genes *ASCL1*, *SUFU*, and *RELN* and the PID-C genes *ATM* and *SMAD4*. In these cases, non-coding mutations complement the coding mutations that disrupt these pathways, covering additional patients.

Intriguingly, we find that RNA splicing pathways were affected primarily by non-coding mutations (FDR $= 7.6 \times 10^{-9}$). A total of 17 PID-N genes involved in splicing-related pathways (Supplementary Fig. 13c), including several heterogeneous nuclear ribonucleoproteins (hnNRP) and serine- and arginine-rich splicing factors (SRSFs). None of these PID-N genes were significantly mutated according to single-element tests used in the PCAWG Drivers and Functional Interpretation Working Group analysis[16].

As we did not find any significant (FDR < 0.1) in cis associations between non-coding mutations and altered expression in splicing-related PID-N genes, we explored potential in trans effects between non-coding mutations in these genes and expression of other genes. We found that non-coding mutations in splicing-related PID-N genes largely recapitulate a recently published association from a TCGA PanCanAtlas analysis[37] between coding mutations in several splicing factors and differential expression of 47 pathways (Fig. 5). In particular, we identified three clusters of mutations in RNA splicing genes (C1, C2, and C3; Fig. 5a, b) using hierarchical clustering of differential expression patterns across these pathways. A highly overlapping set of clusters was found using *t*-distributed stochastic neighbor embedding (top annotation bar in Fig. 5a) showing that the clusters were robust to the choice of the clustering method. Further support for robustness of clusters was found via silhouette scores and bootstrapping (Supplementary Fig. 14). Each of these clusters contained at least one coding mutation in the splicing genes *SF3B1*, *FUBP1*, and *RBM10*, as reported previously[37], along with non-coding mutations in splicing-related PID-N genes, demonstrating that both types of mutations resulted in similar gene expression signatures. The joint analysis of coding and non-coding mutations in splicing factors also recovered the two groups of enriched pathways[37] (P1 and P2 in Fig. 5a; Supplementary Fig. 15). One group (P1) is characterized by immune cell signatures and the other group (P2) reflects mostly cell-autonomous gene signatures of cell cycle, DDR, and essential cellular machineries[37]. This similarity between the gene expression signatures for non-coding mutations in several PID-N splicing factors and the signatures previously reported for coding mutations in splicing factor genes[37] supports a functional role for splicing-related PID-N genes in altering similar gene expression programs.

In addition to the above modules, we also found that transcription factors were well represented among both the PID-C and PID-N genes. In total, nine PID-C genes are transcription factors (*ARHGAP35*, *ARID2*, *FOXA1*, *GATA3*, *NFE2L2*, *SMAD4*, *SOX9*, *TCF7L2*, *TP53*; FDR $= 2.1 \times 10^{-10}$), while 19 PID-N genes are transcription factors (*ASCL1*, *BHLHE40*, *ESRRG*, *HES1*, *HNF1A*, *HNF1B*, *HOXA10*, *HOXB5*, *KLF4*, *MEF2C*, *MYC*, *NFE2*, *NR2F1*, *SOX2*, *SOX4*, *TCF4*, *TP53*, *ZNF521*, *ZNF595*; FDR $= 4.1 \times 10^{-20}$). This observation suggests that non-coding mutations may affect transcriptional regulatory networks.

Discussion

We present an integrative pathway and network analysis that expands the list of genes with possible non-coding driver mutations into the "long tail" of rarely mutated elements that are not significant by single-element analysis. We find that genes harboring both coding or non-coding mutations overlap in pathways and networks; thus, pathway databases and interaction networks serve as useful sources of prior knowledge to implicate additional mutated elements beyond those identified by single-element tests. In total, our integrative pathway and network analysis identified 87 pathway-implicated driver genes with coding variants (PID-C) and 93 pathway-implicated driver genes with non-coding variants (PID-N). Importantly, 90 PID-N genes were not statistically significant (FDR > 0.1) by single-element tests on non-coding mutation data, and these genes are key candidates for future experimental characterization. Among them, we find that promoter mutations in *TP53*, *TLE4*, and *TCF4* are associated with reduced expression of these genes.

We find that coding and non-coding driver mutations largely target different genes and make varying contributions to pathways and networks perturbed in cancer. While some cancer pathways are targeted by both coding and non-coding mutations, such as the Wnt and Notch signaling pathways, other pathways appear to be predominantly altered by one class of mutations. In particular, we find non-coding mutations in multiple genes in the RNA splicing pathway, and samples with these mutations exhibit gene expression signatures that are concordant with gene expression changes observed in samples with coding mutations splicing factors *SF3B1*, *FUBP1*, and *RBM10*[37]. Together these results demonstrate that rare non-coding mutations may result in similar perturbations to both common and complementary biological processes.

There are several caveats to the results reported in this study. First, there is relatively low power to detect non-coding mutations in the cohort, particularly in cancer types with small numbers of
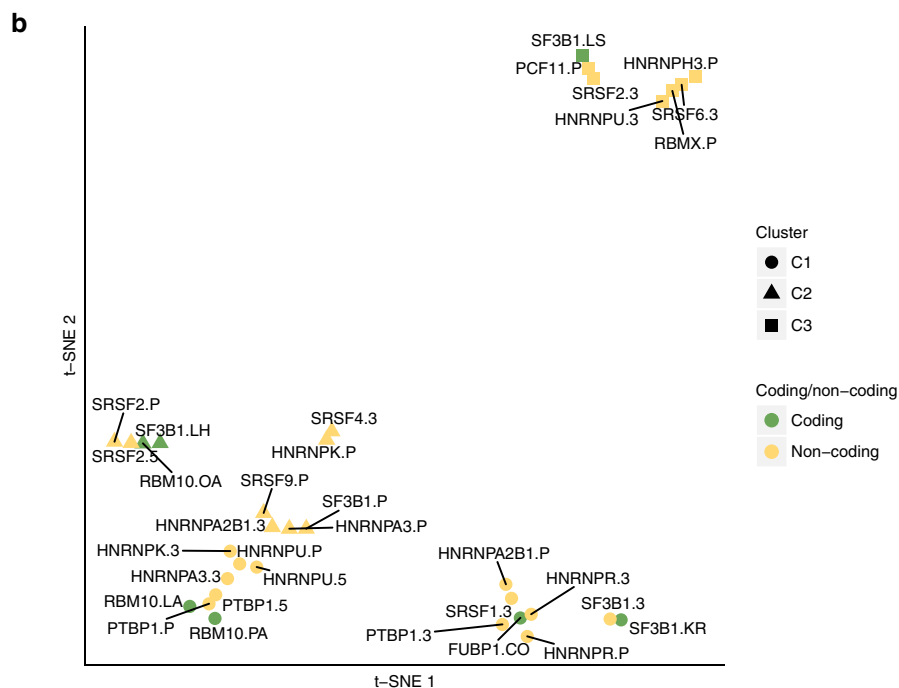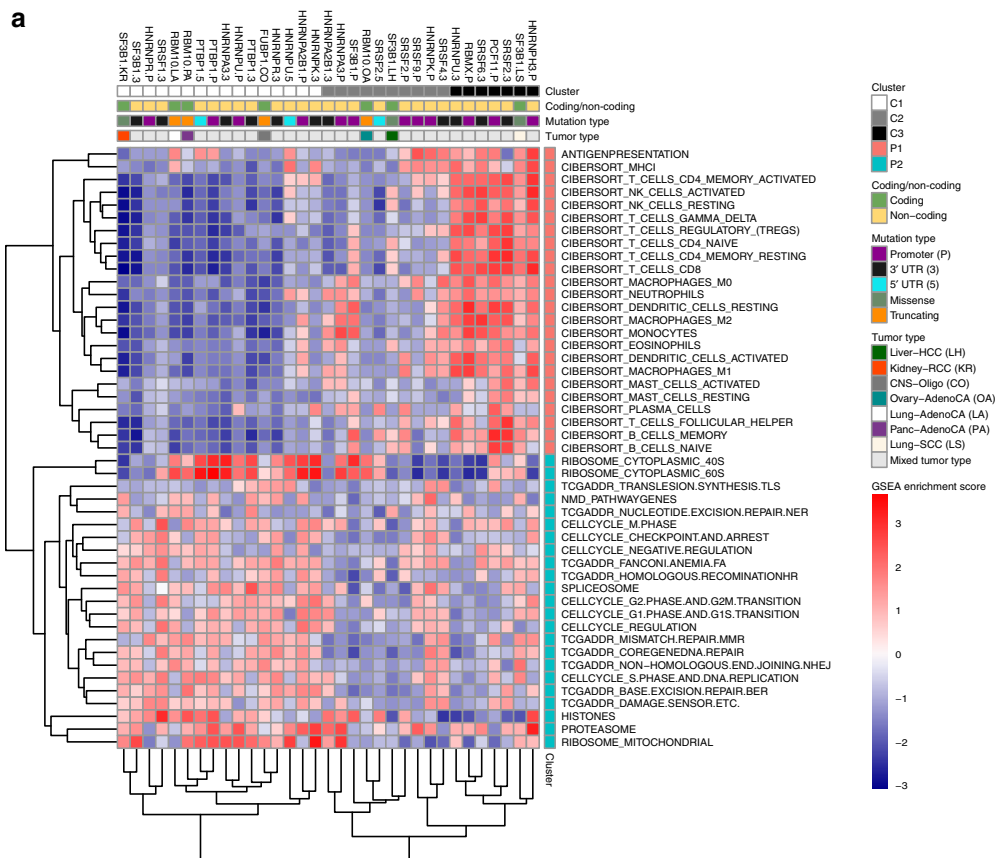
While this analysis accounts for regional variations in the background mutation rate across the genome, it is possible that these corrections are incomplete. Furthermore, if the uncorrected confounding variables are correlated with gene membership in pathways and subnetworks, then the false positive rates in our

Fig. 5 RNA splicing factors are targeted primarily by non-coding mutations and alter expression of similar pathways as coding mutations in splicing factors. a Heatmap of Gene Set Enrichment Analysis (GSEA) Normalized Enrichment Scores (NES). The columns of the matrix indicate non-coding mutations in splicing-related PID-N genes and coding mutations in splicing genes reported in ref. [37], and the rows of the matrix indicate 47 curated gene sets[37]. Red heatmap entries represent an upregulation of the pathway in the mutated samples with respect to the non-mutated samples, and blue heatmap entries represent a downregulation. The first column annotation indicates mutation cluster membership according to common pathway regulation. The second column annotation indicates whether a mutation is a non-coding mutation in a PID-N gene or a coding mutation, with the third column annotation specifies the aberration type (promoter, 5′ UTR, 3′ UTR, missense, or truncating). The fourth column annotation indicates the cancer type for coding mutations. The mutations cluster into three groups: C1, C2, and C3. The pathways cluster into two groups: P1 and P2, where P1 contains an immune signature gene sets and P2 contains cell-autonomous gene sets. **b** t-SNE plot of mutated elements. Gene expression signatures for samples with non-coding mutations clusters in splicing-related PID-N genes with gene expression signatures for coding mutations in previously published splicing factors. The shape of each point denotes the mutation cluster assignment (C1, C2, or C3), and the color represents whether the corresponding gene is a PID-N gene with non-coding mutations or a splicing factor gene with coding mutations.

analysis may be higher than estimated. All of these factors, plus other unknown confounding variables, make it difficult to assess the false discovery rate of our predictions, particularly for PID-N genes. Further experimental validation of these predictions is necessary to determine the true positives from false positives in our PID gene lists.

Because of limited power in individual cancer types, our pathway and network analysis focused on associations found across cancer and tissue types. Thus, we primarily utilized generic, tissue-independent network and pathway information. However, it is well known that gene–gene interactions vary across tissues and that cancer cells rewire signaling and regulatory networks. Future investigations that consider the variable landscape of regulatory and physical interactions across tissues may offer a new perspective on the data. Specifically, each cell type has a different epigenetic wiring and regulatory machinery, and non-coding mutations may target cell type-dependent vulnerabilities. Approaches that incorporate tissue-specific, cancer-specific, or patient-specific gene–gene regulatory information may reveal new classes of drivers unexplored with our current approaches.

Our pathway- and network-driven strategies enable us to interpret the coding and non-coding landscape of tumor genomes to discover driver mechanisms in interconnected systems of genes. This approach has multiple benefits. First, by broadening our mutation analysis from single genomic elements to pathways and networks of multiple genes, we identify new components of known cancer pathways that are recurrently altered by both coding and non-coding mutations, and thus likely to be important in cancer. Second, we identify new pathways and subnetworks that would remain unseen in an analysis focusing on coding sequences. Investigation of the coding and non-coding mutations that perturb these pathways and networks will enable more accurate patient-stratification strategies, pathway-focused biomarkers, and therapeutic approaches.

Methods

Mutation and pathway data. We used gene scores derived from the PCAWG Drivers and Functional Interpretation Working Group analysis[16] and combined several pathways and interaction networks for our pathway and network analyses. Here, we use the term "pathway methods" to refer to approaches that utilize sets of related genes for their analysis and use the term "network methods" to refer to approaches that utilize pairwise interactions among genes and/or their products.

Somatic mutation data. We obtained consensus driver *p*-values (syn8494939) from the PCAWG Drivers and Functional Interpretation Working Group analysis[16] for coding and non-coding (core promoter, 5′ UTR, 3′ UTR, enhancers) genomic elements for the Pancan-no-skin–melanoma–lymph cohort. We removed driver *p*-values for several elements (*H3F3A* and *HIST1H4D* coding; *LEPROTL1*, *TBC1D12*, and *WDR74* 5′ UTR; and chr6:142705600-142706400 enhancer, which targets *ADGRG6*) that the PCAWG Drivers and Functional Interpretation Working Group analysis had manually examined and discarded. We included enhancers with ≤ 5 gene targets (syn7201027), which covered 89% of enhancers elements from the PCAWG Drivers and Functional Interpretation Working Group analysis[16]. In cases where the PCAWG Drivers and Functional Interpretation

Working Group analysis reported multiple *p*-values for the same genomic element, we used the smallest reported *p*-value for that element.

Derivation of gene scores. Pathway databases and gene interaction networks typically record information at the level of individual genes. Thus, we formed coding and non-coding gene scores by combining PCAWG driver *p*-values across coding and/or non-coding (core promoter, 5′ UTR, 3′ UTR, enhancer) genomic elements as follows. Let $p_x(g)$ be the driver *p*-value for element x of gene g from the PCAWG Drivers and Functional Interpretation Working Group analysis[16]. We combined *p*-values from multiple elements using Fisher's method, where we selected the minimum *p*-value $\min(p_{\text{promoter}}(g), p_{5'\text{UTR}}(g))$ for overlapping core promoter and 5′ UTR elements on gene g and the minimum *p*-value $p_{\text{enhancer}}(g)$ of all enhancers targeting gene g. Using this approach, we defined the following gene scores on coding (GS-C), non-coding, (GS-N), and combined coding and non-coding (GS-CN) genomic elements:

1. GS-C: $p_C(g) = p_{\text{coding}}(g)$
2. GS-N: $p_N(g) = \text{fisher}(\min(p_{\text{promoter}}(g), p_{5'\text{UTR}}(g)), p_{3'\text{UTR}}(g), p_{\text{enhancer}}(g))$
3. GS-CN: $p_{CN}(g) = \text{fisher}(p_{\text{coding}}(g), \min(p_{\text{promoter}}(g), p_{5'\text{UTR}}(g)), p_{3'\text{UTR}}(g), p_{\text{enhancer}}(g))$

Here, $p = \text{fisher}(p_1, \dots, p_k) = -2\sum_{i=1}^{k} \ln(p_i) \sim \chi^2_{2k}$, is Fisher's method for combining *p*-values, where $2k$ is the degrees of freedom in the calculation. When the driver *p*-value for a genomic element was undefined, we omitted that element from the calculation and reduced the number of degrees of freedom.

For the pathway and networks methods that analyze individual mutations, we used mutations from the PCAWG MAF (syn7364923) on the same genomic elements as the PCAWG Drivers and Functional Interpretation Working Group analysis, i.e., coding, core promoter, 5′ UTR, 3′ UTR, and enhancer. We removed melanoma and lymphoma samples as well as 69 hypermutated samples with over 30 mutations/MB (syn7894281, syn7814911). We also removed mutations in elements that the PCAWG Drivers and Functional Interpretation Working Group analysis had manually examined and discarded (see above), resulting in lists of mutations used for later assessing biological relevance of our results (syn8103141, syn9684700).

Pathway and network databases. Pathway methods used gene sets from six databases: CORUM[38], GO[39], InterPro[40], KEGG[41], NCI Nature[42], and Reactome[43] (syn3164548, syn11426307), where small (<3 genes) and large (>1000 genes) pathways were removed.

Network methods used interactions from three interaction networks: the largest connected subnetwork of the ReactomeFI 2015 interaction network[44] (syn3254781) with high-confidence (≥ 0.75 confidence score) interactions, which we treated as undirected; the largest connected subnetwork of the iRefIndex14 interaction network[45], which we augmented with interactions from the KEGG pathway database[41] (syn10903761). The BioGRID interaction network[46] (syn3164609) was also used to evaluate and annotate results.

Individual pathway and network algorithms. We applied seven pathway and network methods to the gene scores and mutation data. We used two pathway methods: ActivePathways[19] and a hypergeometric analysis (Vazquez). We also used five network methods: CanIsoNet[20], Hierarchical HotNet[21], an induced subnetwork analysis (Reyna and Raphael, in preparation), NBDI[22], and SSA-ME[23]. Table 1 shows pathway databases and interaction networks used by each method.

Using these pathway and network databases, we ran each method on the GS-C, GS-N, and GS-CN gene scores to identify three corresponding lists of genes. Each method evaluated the statistical significance of its results on each data set.

Non-coding value-added (NCVA) procedure. The GS-CN results leverage both coding and non-coding mutation data, improving the detection of weaker pathway and network signals. We devised a non-coding value-added (NCVA) procedure to separate the coding and non-coding signals in this combined analysis, resulting in a

Table 1 Summary of pathway database and interaction network data for each method.

Method	Pathway databases or interaction networks
ActivePathways	Gene Ontology (GO)[39] biological processes, Reactome[43] pathways
CanIsoNet	STRING v10[50], DIMA[51], 3did[52]
Hierarchical HotNet	ReactomeFI 2015[43], iRefIndex14+KEGG[41, 45]
Hypergeometric analysis	GO biological processes; CORUM[38], KEGG[41], InterPro[41], Nature NCI[42] pathways
Induced subnetwork analysis	ReactomeFI 2015[43], iRefIndex14+KEGG[41, 45]
NBDI	ReactomeFI 2015[43]
SSA-ME	ReactomeFI 2015[43]

set of NCVA genes for which the non-coding mutation data make a statistically significant contribution to their discovery in the GS-CN results. Specifically, we evaluated the statistical significance of genes in the GS-CN results using a permutation test where the driver p-values for coding elements were fixed and the driver p-values for non-coding elements were permuted. This procedure identified the subset of the GS-CN results that were reported infrequently ($p < 0.1$) on permuted data, and thus more likely to be true positives. Each method's NCVA results were added to that method's overall set of non-coding results (GS-N).

Consensus results for pathway and network methods. We defined a consensus set of genes for each set of results: GS-C results, GS-N results, GS-CN results, and GS-N combined with NCVA results, across our seven pathway and network methods. Specifically, we defined a gene to be a consensus gene if it was found by a majority ($\geq 4/7$) of the pathway and network methods. For our analysis, we focused on the consensus GS-C results, which we call the pathway-implicated driver genes with coding variants (PID-C), and the consensus from the GS-N results combined with NCVA results, which we call the pathway-implicated driver genes with non-coding variants (PID-N). We defined PID-C genes as the 87 genes in the consensus of the GS-C results, and we defined PID-N genes as the 93 genes in the consensus of each method's GS-N results combined with its NCVA results. We performed several analyses to assess the biological relevance of PID-C and PID-N genes.

Identification of mutational signatures of PID genes. We performed a permutation-based enrichment test for mutation signatures from PCAWG mutation signatures analysis[47]. We identified the most likely mutation signature for each non-coding mutation in PID-N genes and compared them to randomly chosen non-coding mutations in non-PID-N genes.

Improved network neighborhood scores of PID genes. To assess the extent to which gene scores on PID genes contribute to their detection by pathway and network methods, we considered the contribution of each PID gene's score to the score of its network neighborhood in the BioGRID interaction network.

For each PID gene g, we used Fisher's method to combine the gene scores of the first-order network neighbors of g both with and without the score of g itself. In particular, for gene g, let $p(g)$ be the gene score for g and $N(g)$ be the network neighborhood of g. Then

$$p_{N(g)}^{\text{with}} = \text{fisher}(p(v) : v \in N(g) \cup \{g\})$$

is a score for the network neighborhood of g when including gene g and

$$p_{N(g)}^{\text{without}} = \text{fisher}(p(v) : v \in N(g))$$

is a score for the network neighborhood of g when excluding gene g.

If the network neighborhood of g has a smaller p-value with g than without g, i.e., $p_{N(g)}^{\text{with}} < p_{N(g)}^{\text{without}}$, then gene g improves the score of the network neighborhood, suggesting that the gene score of g plays a role in its detection by pathway and network methods. Alternatively, if the network neighborhood of g has a larger p-value with g than without g, i.e., $p_{N(g)}^{\text{with}} > p_{N(g)}^{\text{without}}$, then gene g worsens the score of the network neighborhood, suggesting that the gene scores of the network neighbors of g are predominantly responsible for the detection of g by pathway and network methods.

We performed this test for every PID-C gene with GS-C gene scores and every PID-N gene with GS-N gene scores. We also sampled genes uniformly at random from the network (87 for PID-C genes and 93 for PID-N genes; 10^6 trials) to ascertain whether significantly more PID genes that improved the scores of their network neighborhoods than expected by chance.

Expression analysis of PID genes. We evaluated whether mutation status of each PID gene was correlated with RNA expression. We used PCAWG-3 gene expression data (syn5553991), which was averaged from TopHat2 and STAR-based

alignments, with FPKM-UQ normalization. Tumor type and copy-number aberrations are known to be covariates for gene expression, so we conditioned on tumor types and annotated copy-number aberrations.

We used the following procedure to assess expression correlations on individual tumor types. We only considered cases with at least three mutated samples and three non-mutated samples to restrict our analysis to cases with sufficient statistical power. For each PID-C gene or each non-coding element in a PID-N gene, we partitioned the samples with expression data into a set A of samples with mutation(s) in the element and a set B of samples without mutations in the element. We performed the Wilcoxon rank-sum test for the expression of the gene in sets A and B and performed the Benjamini–Hochberg correction on each coding or non-coding element to provide FDRs.

We used the following procedure to assess expression correlations across tumor types. We only considered cases with at least one mutated sample and one non-mutated sample to restrict our analysis to cases with sufficient statistical power. For each PID-C gene and each non-coding element in a PID-N gene, we partitioned the samples with expression data into sets A_c of samples in cohort c with mutation(s) in the element and sets B_c of samples in cohort c without mutations in the element. We converted the expression values into z-scores using the expression from non-mutated samples in cohort c, and we computed the Wilcoxon rank-sum test on the expression of the gene in sets from $A = \bigcup_{c \in C} A_c$ and $B = \bigcup_{c \in C} B_c$, where C is the set of all cohorts containing samples with mutation(s) in the element. We then performed the Benjamini–Hochberg correction on each coding or non-coding element to provide FDRs.

Network annotation of PID genes. We performed a permutation test to evaluate the statistical significance of the number of interactions in the BioGRID high-confidence interaction network between PID-C genes, the number of interactions between PID-N genes, and the number of interactions between PID-C and PID-N genes, i.e., when a PID-C gene interacts with a PID-N gene. To compute the permutation p-value, we sampled random networks uniformly at random from the collection of networks with the same degree sequence as the BioGRID network.

We found connected subnetworks of 46 PID-C genes (31 genes expected, $p = 9 \times 10^{-4}$) and 16 PID-N genes (10 genes expected, $p = 6.1 \times 10^{-2}$) in the high-confidence BioGRID[48] protein–protein interaction (PPI) network. The union of the PID-C and PID-N genes formed a larger connected subnetwork of 73 genes (Fig. 4a). These connected subnetworks were significantly larger than expected by chance according to this permutation test (57 genes expected, $p = 2.2 \times 10^{-3}$). Furthermore, we observed statistically significant numbers of protein–protein interactions between PID-C and PID-N genes (67 interactions observed vs. 45 expected, $p = 6 \times 10^{-4}$), suggesting that the associated mutations may target an overlapping set of underlying pathways. The PID-C genes were connected by significantly more interactions than expected (64 vs. 40 expected, $p < 10^{-4}$), and the PID-N genes were interconnected at a sub-significant level (18 vs 12 expected, $p = 6.8 \times 10^{-2}$). Thus, certain pathways are affected by either coding or non-coding mutations, but some pathways are affected by a complement of both coding and non-coding mutations.

Pathway annotation of PID genes. Using g:Profiler[34], we performed a pathway enrichment analysis for PID genes and 12,061 gene sets representing GO biological processes and Reactome pathways. We used the Benjamini–Hochberg correction to control the FDR of the results.

Characterization of PID genes in RNA splicing. GSEA enrichment analysis was performed with the default parameters using the curated pathway gene lists[37] for samples harboring non-synonymous coding mutations in five genes (*FUBP1*, *RBM10*, *SF3B1*, *SRSF2*, and *U2AF1*) with confirmed on-target splicing deregulation. Due to limited number of samples with RNA-seq data in individual tumor types, we restricted our analysis to missense mutations in *SF3B1*, truncating mutations in *RBM10*, and truncating mutations in *FUBP1* for tumor types contained at least three samples with these classes of mutations. Each tumor type containing such mutations was considered separately[37].

We performed the same GSEA analysis for non-coding mutations in 17 PID-N genes that were annotated as involved in RNA splicing. Due to limited number of samples from individual tumor types containing mutations in these genes (often there was only one per tumor type), we performed GSEA analysis jointly on all tumor types containing mutations in an individual PID-N gene, restricting the non-mutated group to samples from the same tumor types as the mutant samples. The GSEA Normalized Enrichment Scores (NES) were clustered using hierarchical complete linkage clustering on the Euclidean distance between the NES scores. Separately, we computed a 2D projection of NES scores using t-Distributed Stochastic Neighbor Embedding (t-SNE).

Ethical review. Sequencing of human subjects' tissue was performed by ICGC and TCGA consortium members under approval of local Institutional Review Boards (IRBs). Informed consent was obtained from all human participants. All data were deidentified for this study, and data access for participating researchers was obtained through data access agreements between local institutions, the ICGC Data Access Compliance Office (DACO), and the NIH dbGaP.

Data availability

Raw data are available through the PCAWG data portal https://dcc.icgc.org/pcawg. Processed data from the PCAWG Networks and Pathways working groups as well as other PCAWG working groups are available on www.synapse.org at the Synapse links provided in this section. The source data underlying Figs. 2a–c, 3a, b, c, and Supplementary Figs. 1a–c, 7a–c, 10a–h are provided as a Source Data file. Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described here[15] and available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance

with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier, which does not require access approval. To access potentially identification information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login) for access to the TCGA portion of the data set, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single-nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization.

Derived data sets described specifically in this paper can be found at these locations:

Label	Synapse ID	ICGC DCC URL	ICGC DCC file name	Access (open/controlled)
PCAWG driver p-values	syn8494939	https://dcc.icgc.org/releases/PCAWG/networks/	final_integration_results_2017_03_16.tar.gz	Open
Enhancer-gene mappings	syn7201027	https://dcc.icgc.org/releases/PCAWG/networks/	map.enhancer.gene.txt.gz	Open
Somatic MAF file	syn7364923	https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel/	final_consensus_passonly.snv_mnv_indel.icgc.public.maf.gz	Open
Somatic MAF file	syn7364923	https://dcc.icgc.org/releases/PCAWG/consensus_snv_indel/	final_consensus_passonly.snv_mnv_indel.tcga.controlled.maf.gz	Controlled
Hypermutated donors	syn7894281	https://dcc.icgc.org/releases/PCAWG/networks/	Hypermutated_spls_removed_ActiveDriver2_AllScores_211216.txt	Open
Hypermutated samples	syn7814911	https://dcc.icgc.org/releases/PCAWG/networks/	Hypermutated_spls_removed_ActiveDriver2_AllScores_291116.aliquotid.txt	Open
Mutations to coding and noncoding elements	syn8103141	https://dcc.icgc.org/releases/PCAWG/networks/	PCAWG_mutations_to_elements.icgc.public.txt.gz	Open
Mutations to coding and noncoding elements	syn8103141	https://dcc.icgc.org/releases/PCAWG/networks/	PCAWG_mutations_to_elements.tcga.controlled.txt.gz	Controlled
Mutation matrix	syn9684700	https://dcc.icgc.org/releases/PCAWG/networks/	PCAWG.gene_status.all.tsv.gz	Controlled
Primary pathway databases	syn3164548	https://dcc.icgc.org/releases/PCAWG/networks/	Gene_sets_pathways_processes_functions.zip	Open
Secondary pathway databases	syn11426307	https://dcc.icgc.org/releases/PCAWG/networks/	PCAWG-5.pathway.data.CNIO.tar.gz	Open
ReactomeFI 2015 network	syn3254781	https://dcc.icgc.org/releases/PCAWG/networks/	Functional_interaction_network_Reactome_FI_Network_2015.zip	Open
iRefIndex14 network	syn10903761	https://dcc.icgc.org/releases/PCAWG/networks/	irefindex14-kegg.tsv.gz	Open
BioGRID network	syn3164609	https://dcc.icgc.org/releases/PCAWG/networks/	Protein_Protein_interaction_network_BIOGRID_filtered.zip	Open
STRING v10 network	syn11712027	https://dcc.icgc.org/releases/PCAWG/networks/	string10_ppi_high_confident_edges.tsv	Open
PCAWG gene expression data	syn5553991	https://dcc.icgc.org/releases/PCAWG/transcriptome/gene_expression/	tophat_star_fpkm_uq.v2_aliquot_gl.tsv.gz	Controlled
PCAWG pathway and network method results	syn21413360	https://dcc.icgc.org/releases/PCAWG/networks/	pathway_and_network_method_results.tar.gz	Open
PCAWG pathway and network consensus results	syn11654843	https://dcc.icgc.org/releases/PCAWG/networks/	method_results_2017_10_10.tar.gz	Open
Coding and noncoding elements	syn21416282	https://dcc.icgc.org/releases/PCAWG/networks/	gene-coding-and-non-coding-elements.tar.gz	Open
Transcript expression data (Counts)	syn7536588	https://dcc.icgc.org/releases/PCAWG/transcriptome/transcript_expression/	pcawg.rnaseq.transcript.expr.counts.tsv.gz	Controlled
Transcript expression data (FPKM)	syn7536589	https://dcc.icgc.org/releases/PCAWG/transcriptome/transcript_expression/	pcawg.rnaseq.transcript.expr.fpkm.tsv.gz	Controlled
eQTL data	syn17096221	https://dcc.icgc.org/releases/PCAWG/transcriptome/eQTL/summarystats/somatic/	all_somatic_eqtl.tsv.tar.gz	Controlled
Gene-level copy-number data	syn8291899	https://dcc.icgc.org/releases/PCAWG/consensus_cnv/gene_level_calls/	all_samples.consensus_CN.by_gene.170214.txt.gz	Open
CanIsoNet PCAWG Ensembl transcripts	syn7536587	https://dcc.icgc.org/releases/PCAWG/transcriptome/transcript_expression/	pcawg.rnaseq.transcript.expr.tpm.tsv.gz	Open
CanIsoNet GTEx Ensembl transcripts	syn7596599	https://dcc.icgc.org/releases/PCAWG/transcriptome/transcript_expression/	GTEX_v4.pcawg.transcripts.tpm.tsv.gz	Open
CanIsoNet filtered PCAWG samples	syn7416381	https://dcc.icgc.org/releases/PCAWG/transcriptome/metadata/	rnaseq.extended.metadata.aliquot_id.V4.tsv	Open
CanIsoNet filtered GTEx samples	syn7596611	https://dcc.icgc.org/releases/PCAWG/transcriptome/metadata/	GTEX_v4.metadata.tsv.gz	Open
CanIsoNet protein–protein isoforms	syn10245952	https://dcc.icgc.org/releases/PCAWG/networks/	isoNet.tsv.gz	Open
CanIsoNet shortest path results	syn9770515	https://dcc.icgc.org/releases/PCAWG/networks/	string_cosmic_neighbourhood_min900_shell3_20160527.tsv.xz	Open
CanIsoNet functional regions	syn7345646	https://dcc.icgc.org/releases/PCAWG/networks/	allCombined.zip	Open
CanIsoNet results (noncoding region)	syn9765614	CanIsoNet results (noncoding region)	non_canIsoNet_mdi_results_noLymNoMel.tsv	Open
CanIsoNet results (coding region)	syn9765615	CanIsoNet results (coding region)	cds_canIsoNet_mdi_results_noLymNoMel.tsv	Open

Code availability

Code for the contributing methods in this analysis can be found from their respective papers or by request to the contributing author. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v3.0, which allows for reuse and distribution.

References

1. Gonzalez-Perez, A. et al. Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10**, 723–729 (2013).
2. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
3. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
4. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
5. Horn, S. et al. TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
6. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
7. Zhang, W. et al. A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet.* **50**, 613–620 (2018).
8. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
9. Creixell, P. et al. Pathway and network analysis of cancer genomes. *Nat. Methods* **12**, 615–621 (2015).
10. Vandin, F., Upfal, E. & Raphael, B. J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
11. Paull, E. O. et al. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (TieDIE). *Bioinformatics* **29**, 2757–2764 (2013).
12. Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci. Rep.* **3**, 2651 (2013).
13. Leiserson, M. D. M. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
14. Horn, H. et al. NetSig: network-based discovery from cancer genomes. *Nat. Methods* **15**, 61–66 (2018).
15. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
16. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1965-x (2020).
17. Fredriksson, N. J. et al. Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *PLoS Genet.* **13**, e1006773 (2017).
18. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
19. Paczkowska, M. et al. Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* https://doi.org/10.1038/s41467-019-13983-9 (2020).
20. Kahraman, A. & von Mering, C. Pathogenic impact of isoform switches in 1209 cancer samples covering 27 cancer types using an isoform-specific interaction network. Preprint at https://www.biorxiv.org/content/10.1101/742379v1 (2019).
21. Reyna, Matthew A., Leiserson, Mark D. M. & Raphael, Benjamin J. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i982 (2018).
22. Verbeke, L. P. C. et al. Pathway relevance ranking for tumor samples through network-based data integration. *PLoS One* **10**, e0133503 (2015).
23. Pulido-Tamayo, S., Weytjens, B., De Maeyer, D. & Marchal, K. SSA-ME detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. *Sci. Rep.* **6**, 36257 (2016).
24. Marbach, D. et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
25. Forbes, S. A. et al. The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* **10**, Unit 10.11 (2008).
26. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* https://doi.org/10.1038/s41586-019-1907-7 (2020).
27. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
28. Jesionek-Kupnicka, D. et al. TP53 promoter methylation in primary glioblastoma: relationship with TP53 mRNA and protein expression and mutation status. *DNA Cell Biol.* **33**, 217–226 (2014).
29. Fisher, A. L. & Caudy, M. Groucho proteins: transcriptional corepressors for specific subsets of DNA-binding transcription factors in vertebrates and invertebrates. *Genes Dev.* **12**, 1931–1940 (1998).
30. Shin, T. H., Brynczka, C., Dayyani, F., Rivera, M. N. & Sweetser, D. A. TLE4 regulation of Wnt-mediated inflammation underlies its role as a tumor suppressor in myeloid leukemia. *Leuk. Res.* **48**, 46–56 (2016).
31. Dayyani, F. et al. Loss of TLE1 and TLE4 from the del(9q) commonly deleted region in AML cooperates with AML1-ETO to affect myeloid cell proliferation and survival. *Blood* **111**, 4338–4347 (2008).
32. Cadigan, K. M. & Waterman, M. L. TCF/LEFs and Wnt signaling in the nucleus. *Cold Spring Harb. Perspect. Biol.* **4**, a007906 (2012).
33. Mélard, P. et al. Molecular alterations and tumor suppressive function of the DUSP22 (dual specificity phosphatase 22) gene in peripheral T-cell lymphoma subtypes. *Oncotarget* **7**, 68734–68748 (2016).
34. Reimand, J. et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–9 (2016).
35. Augustyn, A. et al. ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung cancers. *Proc. Natl Acad. Sci. USA* **111**, 14788–14793 (2014).
36. Lasorella, A., Benezra, R. & Iavarone, A. The ID proteins: master regulators of cancer stem cells and tumour aggressiveness. *Nat. Rev. Cancer* **14**, 77–91 (2014).
37. Seiler, M. et al. Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Rep.* **23**, 282–296.e4 (2018).
38. Ruepp, A. et al. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **36**, D646–D650 (2008).
39. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
40. Finn, R. D. et al. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45**, D190–D199 (2017).
41. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. Schaefer, C. F. et al. PID: the pathway interaction database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
43. Croft, D. et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39**, D691–D697 (2011).
44. Croft, D. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477 (2014).
45. Razick, S., Magklaras, G. & Donaldson, I. M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinforma.* **9**, 405 (2008).
46. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379 (2017).
47. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* https://doi.org/10.1038/s41586-020-1943-3 (2020).
48. Stark, C. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
49. Consortium, T. E. P., The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
50. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
51. Luo, Q., Pagel, P., Vilne, B. & Frishman, D. DIMA 3.0: domain interaction map. *Nucleic Acids Res.* **39**, D724–D729 (2011).
52. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **42**, D374–D379 (2014).

Author contributions

Competing interests

Additional information

Matthew A. Reyna [1,2], David Haan[3], Marta Paczkowska[4], Lieven P.C. Verbeke [5,6], Miguel Vazquez [7,8], Abdullah Kahraman [9,10], Sergio Pulido-Tamayo[5,6], Jonathan Barenboim[4], Lina Wadi[4], Priyanka Dhingra[11], Raunak Shrestha [12], Gad Getz [13,14,15,16], Michael S. Lawrence[13,14], Jakob Skou Pedersen [17,18], Mark A. Rubin [11], David A. Wheeler[19], Søren Brunak[20,21], Jose M.G. Izarzugaza[20,21], Ekta Khurana [11], Kathleen Marchal [5,6], Christian von Mering [9], S. Cenk Sahinalp[12,22], Alfonso Valencia[7,23], PCAWG Drivers and Functional Interpretation Working Group, Jüri Reimand [4,24]*, Joshua M. Stuart [3]*, Benjamin J. Raphael [1]* & PCAWG Consortium

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA. [2]Department of Biomedical Informatics, Emory University, Atlanta, GA 30322, USA. [3]Department of Biomolecular Engineering and UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95060, USA. [4]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [5]Department of Information Technology, IDLab, Ghent University, IMEC, Ghent, the Netherlands. [6]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, the Netherlands. [7]Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain. [8]Norwegian University of Science and Technology, Trondheim, Norway. [9]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, CH-8057 Zurich, Switzerland. [10]Department of Pathology and Molecular Pathology, University Hospital Zurich, CH-8091 Zurich, Switzerland. [11]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA. [12]Vancouver Prostate Centre, 2660 Oak Street, Vancouver, BC V6H 3Z6, Canada. [13]The Broad Institute of MIT and Harvard, Cambridge, MA 02124, USA. [14]Massachusetts General Hospital Center for Cancer Research, Charlestown, MA 02129, USA. [15]Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA. [16]Massachusetts General Hospital, Department of Pathology, Boston, MA 02114, USA. [17]Department of Molecular Medicine (MOMA), Aarhus University Hospital, Aarhus, Denmark. [18]Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark. [19]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. [20]DTU Bioinformatics, Department of Bio and Health Informatics, Technical University of Denmark, Kemitorvet, 2800 Kongens Lyngby, Denmark. [21]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences,

University of Copenhagen, 2200 Copenhagen, Denmark. [22]Department of Computer Science, Indiana University, Bloomington, IN 47405, USA. [23]ICREA, Barcelona 08010, Spain. [24]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. PCAWG Drivers and Functional Interpretation Working Group authors and their affiliations appears at the end of the paper. PCAWG Consortium members and their affiliations appear in the Supplementary Information. *email: juri.reimand@utoronto.ca; jstuart@ucsc.edu; braphael@princeton.edu

PCAWG Drivers and Functional Interpretation Working Group

Federico Abascal[25], Samirkumar B. Amin[26,27,28], Gary D. Bader[29], Pratiti Bandopadhayay[30,31,32], Rameen Beroukhim[32,33,34], Johanna Bertl[17,35], Keith A. Boroevich[36,37], John Busanovich[32,38], Peter J. Campbell[25,39], Joana Carlevaro-Fita[40,41,42], Dimple Chakravarty[43], Calvin Wing Yiu Chan[44,45], Ken Chen[46], Jung Kyoon Choi[47], Jordi Deu-Pons[48,49], Klev Diamanti[50], Lars Feuerbach[51], J. Lynn Fink[7,52], Nuno A. Fonseca[53,54], Joan Frigola[48], Carlo Gambacorti-Passerini[55], Dale W. Garsed[56,57], Mark Gerstein[58,59,60], Qianyun Guo[18], Ivo G. Gut[61,62], Mark P Hamilton[63], Nicholas J. Haradhvala[32,64], Arif O. Harmanci[60,65], Mohamed Helmy[66], Carl Herrmann[44,67,68], Julian M. Hess[14,32], Asger Hobolth[18,35], Ermin Hodzic[69], Chen Hong[45,51], Henrik Hornshøj[17], Keren Isaev[24,70], Rory Johnson[40,41], Todd A. Johnson[36], Malene Juul[17], Randi Istrup Juul[17], Andre Kahles[71,72,73,74,75], Manolis Kellis[32,76], Jaegil Kim[32], Jong K. Kim[77], Youngwook Kim[78,79], Jan Komorowski[50,80], Jan O. Korbel[54,81], Sushant Kumar[59,60], Andrés Lanzós[40,41,42], Erik Larsson[71], Donghoon Lee[60], Kjong-Van Lehmann[71,72,73,74,75], Shantao Li[60], Xiaotong Li[60], Ziao Lin[32,82], Eric Minwei Liu[83,84,85], Lucas Lochovsky[27,59,60,86], Shaoke Lou[59,60], Tobias Madsen[17], Iñigo Martincorena[25], Alexander Martinez-Fundichely[83,84,87], Yosef E. Maruvka[14,32,64], Patrick D. McGillivray[59], William Meyerson[60,88], Ferran Muiños[49,89], Loris Mularoni[49,89], Hidewaki Nakagawa[37], Morten Muhlig Nielsen[17], Keunchil Park[90,91], Kiejung Park[92], Tirso Pons[93], Iker Reyes-Salazar[89], Esther Rheinbay[32,34,64], Carlota Rubio-Perez[49,89,94], Gordon Saksena[32], Leonidas Salichos[59,60], Chris Sander[71,95,96,97], Steven E. Schumacher[32,98], Mark Shackleton[56,57], Ofer Shapira[32,96], Ciyue Shen[97,99], Shimin Shuai[29,70], Nikos Sidiropoulos[100], Lina Sieverling[45,51], Nasa Sinnott-Armstrong[32,101], Lincoln D. Stein[29,70], David Tamborero[49,89], Grace Tiao[32], Tatsuhiko Tsunoda[36,102,103,104], Husen M. Umer[50,105], Liis Uuskula-Reimand[106,107], Claes Wadelius[108], Jiayin Wang[109,110,111], Jonathan Warrell[59,60], Sebastian M. Waszak[81], Joachim Weischenfeldt[81,100,112], Guanming Wu[113], Jun Yu[114,115], Jing Zhang[60], Xuanping Zhang[109,116], Yan Zhang[60,117,118], Zhongming Zhao[119] & Lihua Zou[120]

[25]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [26]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [27]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. [28]Quantitative & Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX 77030, USA. [29]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [30]Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA 02215, USA. [31]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA. [32]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [33]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [34]Harvard Medical School, Boston, MA 02115, USA. [35]Department of Mathematics, Aarhus University, Aarhus 8000, Denmark. [36]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [37]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [38]Foundation Medicine, Inc, Cambridge, MA 02141, USA. [39]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK. [40]Department for BioMedical Research, University of Bern, Bern 3008, Switzerland. [41]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, Bern 3010, Switzerland. [42]Graduate School for Cellular and Biomedical Sciences, University of Bern, Bern 3012, Switzerland. [43]Department of Genitourinary Medical Oncology - Research, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [44]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [45]Faculty of Biosciences, Heidelberg University, Heidelberg 69120, Germany. [46]University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [47]Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea. [48]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona 8003, Spain. [49]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, Barcelona 08002, Spain. [50]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, Uppsala SE-75124, Sweden. [51]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany. [52]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia. [53]CIBIO/InBIO - Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão 4485-601, Portugal. [54]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [55]University of Milano Bicocca, Monza 20052, Italy. [56]Peter MacCallum Cancer Centre and University of Melbourne, Melbourne,, VIC 3000, Australia. [57]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC 3052, Australia. [58]Department of Computer Science, Yale University, New Haven, CT 06520, USA. [59]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [60]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [61]CNAG-CRG,

Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona 08028, Spain. [62]Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain. [63]Department of Internal Medicine, Stanford University, Stanford, CA 94305, USA. [64]Massachusetts General Hospital, Boston, MA 02114, USA. [65]Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX 77030, USA. [66]The Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada. [67]Health Data Science Unit, University Clinics, Heidelberg 69120, Germany. [68]Institute of Pharmacy and Molecular Biotechnology and BioQuant, Heidelberg University, Heidelberg 69120, Germany. [69]Simon Fraser University, Burnaby, BC V5A 1S6, Canada. [70]Computational Biology Program, Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [71]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [72]ETH Zurich, Department of Biology, Wolfgang-Pauli-Strasse 27, Zürich 8093, Switzerland. [73]ETH Zurich, Department of Computer Science, Zurich 8092, Switzerland. [74]SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland. [75]University Hospital Zurich, Zurich 8091, Switzerland. [76]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [77]Research Core Center, National Cancer Centre Korea, Goyang-si 410-769, South Korea. [78]Department of Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [79]Samsung Genome Institute, Seoul 06351, South Korea. [80]Institute of Computer Science, Polish Academy of Sciences, Warsawa 01-248, Poland. [81]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg 69117, Germany. [82]Harvard University, Cambridge, MA 02138, USA. [83]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA. [84]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. [85]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [86]Yale University, New Haven, CT 06520, USA. [87]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10065, USA. [88]Yale School of Medicine, Yale University, New Haven, CT 06520, USA. [89]Institute for Research in Biomedicine (IRB Barcelona), Barcelona 08028, Spain. [90]Division of Hematology-Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [91]Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [92]Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan 31066, South Korea. [93]Spanish National Cancer Research Centre, Madrid 28029, Spain. [94]Vall d'Hebron Institute of Oncology: VHIO, Barcelona 08035, Spain. [95]cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA. [96]Dana-Farber Cancer Institute, Boston, MA 02215, USA. [97]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA. [98]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [99]cBio Center, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [100]Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, Copenhagen 2200, Denmark. [101]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. [102]CREST, Japan Science and Technology Agency, Tokyo 113-0033, Japan. [103]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo 113-8510, Japan. [104]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan. [105]Science for Life Laboratory, Department of Oncology-Pathology, Karolinska Institutet, Stockholm 17121, Sweden. [106]Department of Gene Technology, Tallinn University of Technology, Tallinn 12616, Estonia. [107]Genetics & Genome Biology Program, SickKids Research Institute, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada. [108]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala SE-75108, Sweden. [109]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710048, China. [110]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710048, China. [111]The McDonnell Genome Institute at Washington University, St Louis, MO 63108, USA. [112]Department of Urology, Charité Universitätsmedizin Berlin, Berlin 10117, Germany. [113]Oregon Health & Sciences University, Portland, OR 97239, USA. [114]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, NTHong Kong, China. [115]Second Military Medical University, Shanghai 200433, China. [116]The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [117]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA. [118]The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH 43210, USA. [119]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [120]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60637, USA