nature genetics

Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes

Mark D M Leiserson[1,2,14], Fabio Vandin[1,2,13,14], Hsin-Ta Wu[1,2], Jason R Dobson[1–3], Jonathan V Eldridge[1], Jacob L Thomas[1], Alexandra Papoutsaki[1], Younhun Kim[1], Beifang Niu[4], Michael McLellan[4], Michael S Lawrence[5], Abel Gonzalez-Perez[6], David Tamborero[6], Yuwei Cheng[7], Gregory A Ryslik[8], Nuria Lopez-Bigas[6,9], Gad Getz[5,10], Li Ding[4,11,12] & Benjamin J Raphael[1,2]

Cancers exhibit extensive mutational heterogeneity, and the resulting long-tail phenomenon complicates the discovery of genes and pathways that are significantly mutated in cancer. We perform a pan-cancer analysis of mutated networks in 3,281 samples from 12 cancer types from The Cancer Genome Atlas (TCGA) using HotNet2, a new algorithm to find mutated subnetworks that overcomes the limitations of existing single-gene, pathway and network approaches. We identify 16 significantly mutated subnetworks that comprise well-known cancer signaling pathways as well as subnetworks with less characterized roles in cancer, including cohesin, condensin and others. Many of these subnetworks exhibit co-occurring mutations across samples. These subnetworks contain dozens of genes with rare somatic mutations across multiple cancers; many of these genes have additional evidence supporting a role in cancer. By illuminating these rare combinations of mutations, pan-cancer network analyses provide a roadmap to investigate new diagnostic and therapeutic opportunities across cancer types.

Recent whole-genome and whole-exome sequencing studies have provided an ever-expanding survey of somatic aberrations in cancer and have identified multiple new cancer-related genes[1–7]. At the same time, these studies demonstrated that most cancers exhibit extensive mutational heterogeneity, with few significantly mutated genes and many genes mutated in a small number of samples[8–10]. This 'long-tail' phenomenon complicates efforts to identify cancer-related genes by statistical tests of mutational recurrence, as rarely mutated cancer genes may be indistinguishable from genes containing only passenger mutations. Even recent TCGA pan-cancer studies[11–16] had limited power to characterize genes in the long tail, leaving an incomplete picture of the functional, somatic mutations in these samples.

A prominent explanation for the mutational heterogeneity observed in cancer is the fact that genes act together in various signaling and regulatory pathways and protein complexes[9,15]. Clustering of mutations in known pathways has been illustrated in previous TCGA studies[1–6] but typically without a measure of statistical significance. Although statistical tests of enrichment in known pathways or gene sets exist, such tests do not identify new pathways, have limited power to evaluate cross-talk between known pathways and generally ignore the topology of interactions between genes.

We introduce a new and complementary approach to identify pathways and protein complexes perturbed by somatic aberrations. This approach combines (i) a new algorithm, HotNet2, for the identification of mutated subnetworks in a genome-scale interaction network and (ii) a large TCGA pan-cancer data set of somatic single-nucleotide variants (SNVs), small indels and copy number aberrations (CNAs) measured in 3,281 samples from 12 cancer types[14]. HotNet2 uses a directed heat diffusion model to simultaneously assess the significance of mutations in individual genes and the local topology of interactions among the encoded proteins, overcoming the limitations of pathway-based enrichment statistics and earlier network approaches.

Our TCGA HotNet2 pan-cancer analysis identifies 16 significantly mutated subnetworks that encompass classic cancer signaling pathways, pathways and complexes with more recently characterized roles in cancer, and protein complexes and groups of interacting proteins with less characterized roles in cancer, such as the cohesin and condensin complexes. These latter two subnetworks—as well as many of the proteins in all subnetworks—are rarely affected by mutations in each cancer type and were thus identified only by the pan-cancer network analysis. Many of the rarely mutated proteins in

[1]Department of Computer Science, Brown University, Providence, Rhode Island, USA. [2]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, USA. [3]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, Rhode Island, USA. [4]Genome Institute, Washington University in St. Louis, St. Louis, Missouri, USA. [5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [6]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, University Pompeu Fabra, Barcelona, Spain. [7]Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [8]Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, USA. [9]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [10]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [11]Department of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA. [12]Siteman Cancer Center, Washington University in St. Louis, St. Louis, Missouri, USA. [13]Present address: Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. [14]These authors contributed equally to this work. Correspondence should be addressed to B.J.R. (braphael@brown.edu).

Received 20 March; accepted 20 November; published online 15 December 2014; doi:10.1038/ng.3168

the subnetworks have documented physical interactions with well-characterized cancer-related proteins and/or mutational patterns (for example, clustering in protein sequences or structures or an excess of inactivating alterations) that lend additional support for their role in cancer. Co-occurrence of alterations across these subnetworks supports the hypothesis that many of the subnetworks correspond to distinct biological functions.

In comparison to single-gene tests of significance, our TCGA HotNet2 pan-cancer analysis delves deeper into the long tail of rarely mutated genes and also assembles combinations of individual genes into a relatively small number of interacting networks. The mutational landscape of cancer has been proposed to consist of 'mountains' of frequently mutated genes and 'hills' of less frequently mutated genes[9]. Our pan-cancer network approach provides a richer annotation of this landscape, grouping individual peaks and mountains into mountain ranges and their associated foothills, further enabling diagnostic and therapeutic approaches in cancer care.

RESULTS

HotNet2 identifies significantly mutated subnetworks

We assembled a TCGA pan-cancer data set of exome sequencing, array copy number and RNA sequencing (RNA-seq) data from 3,281 samples from 12 cancer types, analyzing SNVs, small indels and CNAs in 19,424 transcripts (**Fig. 1a** and **Supplementary Fig. 1**). After we removed hypermutated samples and genes with low expression in all tumor types (Online Methods), the data set contained 11,565 mutated genes in 3,110 tumors. We observed that the number of samples with a mutation in a given gene varied over three orders of magnitude, with from 1 to 1,291 mutated samples (**Fig. 1b**). Moreover, we discovered that this broad spectrum of mutational frequencies—from common to extremely rare—posed a challenge for the identification of significantly mutated subnetworks. Specifically, our goal was to identify subnetworks according to both the frequency of somatic mutations in individual genes and the topology of the interactions between the corresponding proteins. However, the presence of highly mutated genes encoding highly connected proteins such as *TP53* presents difficulties for existing algorithms that attempt to achieve this goal, for example, the HotNet algorithm[16,17] that was used for cancer network analysis in TCGA and other studies[3,4,7,18] and related network-propagation approaches[19]. In the heat diffusion model used in HotNet, genes such as *TP53* are extremely 'hot' nodes and propagate this heat to their neighboring nodes. The resulting 'star' subnetworks centered on the hot node (Online Methods and **Supplementary Fig. 2**) contain many neighboring genes that are not mutated at an appreciable frequency and are of limited biological interest.

We introduce the HotNet2 (HotNet diffusion-oriented subnetworks) algorithm to address the problem of finding significantly mutated subnetworks in large, broad data sets of mutational frequency spectra such as the pan-cancer data (**Fig. 1c**,**d** and **Supplementary Fig. 3**). HotNet2 uses a modified diffusion process and considers the source, or directionality, of heat flow in the identification of subnetworks (**Supplementary Fig. 4**). This approach reduces the incidence of the artifact of star subnetworks by more than 80%, decreasing the false positive rate and enabling the identification of more subtle subnetworks with rare mutations of high biological relevance (Online Methods). We compared HotNet2 to other algorithms (Online Methods) and found that HotNet2 has higher sensitivity and specificity on both real and simulated data.

We performed HotNet2 analysis using two approaches to assign heat to individual genes according to mutational recurrence[20] and using three different interaction networks[21–24] with varying numbers of interactions (Online Methods). HotNet2 identified a significant number of subnetworks ($P < 0.01$; **Supplementary Tables 1** and **2**) for each of the two gene scores and three networks. We combined the resulting subnetworks into 14 consensus subnetworks that were found across different gene scores and networks ($P < 0.004$; **Supplementary Tables 3–5**), plus the condensin complex and a network containing the CLASP and CLIP proteins (**Supplementary Fig. 5**) that were significant in individual interaction networks (**Supplementary Tables 6** and **7**). Our consensus process also identified 13 'linker' proteins that were members of more than one consensus subnetwork. We developed an online interactive viewer for the HotNet2 pan-cancer subnetworks (see URLs and **Supplementary Fig. 6**).

The subnetworks and linker genes (**Fig. 2a**) included portions of well-known cancer pathways such as the TP53, phosphoinositide 3-kinase (PI3K), NOTCH and receptor tyrosine kinases (RTK) signaling pathways (**Supplementary Fig. 7**), as well as pathways and complexes that have more recently been observed to be important in cancer, such as the SWI/SNF and BAP1 chromatin-remodeling complexes, NFE2L2-KEAP1 (**Supplementary Figs. 8** and **9**) and the RUNX1-CBFB core binding complex (**Supplementary Fig. 10**). The fifth most mutated subnetwork (16.9% of samples) consisted of MLL2 and MLL3 and the putative interacting protein KDM6A (**Supplementary Fig. 11**) and was highly mutated (28.9% of samples) in the TCGA pan-cancer squamous integrated subtype[25]. HotNet2 identified less characterized and potentially new subnetworks that might also have important roles in cancer, including the cohesin and condensin complexes and major histocompatibility complex (MHC) class I proteins. The MHC class I subnetwork (**Supplementary Fig. 12**) exemplifies the ability of HotNet2 to identify rarely mutated cancer-related genes: all of the genes constituting the subnetwork were mutated in fewer than 35 samples (1.1%), yet 4 of the 5 genes have recently been proposed as new cancer-relevant genes[11]. The sections below further detail a subset of these subnetworks. Additional analyses are provided in the **Supplementary Note**.

Many of the subnetworks exhibited a significant enrichment for mutations in a subset of cancer types, including many previously unreported associations (**Supplementary Tables 6–18**). We also identified genes within these subnetworks enriched for mutations in particular cancer types. In addition, the HotNet2 pan-cancer analysis provided a clearer and more robust summary of subnetworks and newly involved genes than HotNet2 analysis of individual cancer types (**Supplementary Table 19**).

These subnetworks and linkers included a total of 147 genes, with many well-known cancer genes and pathways but also genes with mutations that were too rare to be significant by single-gene tests (**Supplementary Table 20**). In total, 92 genes in the HotNet2 subnetworks were not reported by any of 5 single-gene tests (MutSigCV[20], Oncodrive-FM[26], Oncodrive-CIS[27], MuSiC[28] or GISTIC2[29]) or listed as a known driver gene in Vogelstein *et al.*[9], and an additional 13 genes were reported in only one such list. Many of these genes have literature evidence supporting a potential role in cancer, and others are in biological processes that suggest that these genes warrant further study. A subset of the promising candidates is listed in **Table 1**, with the full list and associated references appearing in **Supplementary Table 20**.

To obtain additional support for these genes, we examined whether they had either an excess of inactivating mutations[9] or a cluster of missense mutations in the encoded protein sequence (using NMC[30]) or structure (using iPAC[31]; **Supplementary Figs. 13** and **14**, and **Supplementary Tables 21** and **22**). We found that the genes in HotNet2 consensus subnetworks were enriched for inactivating

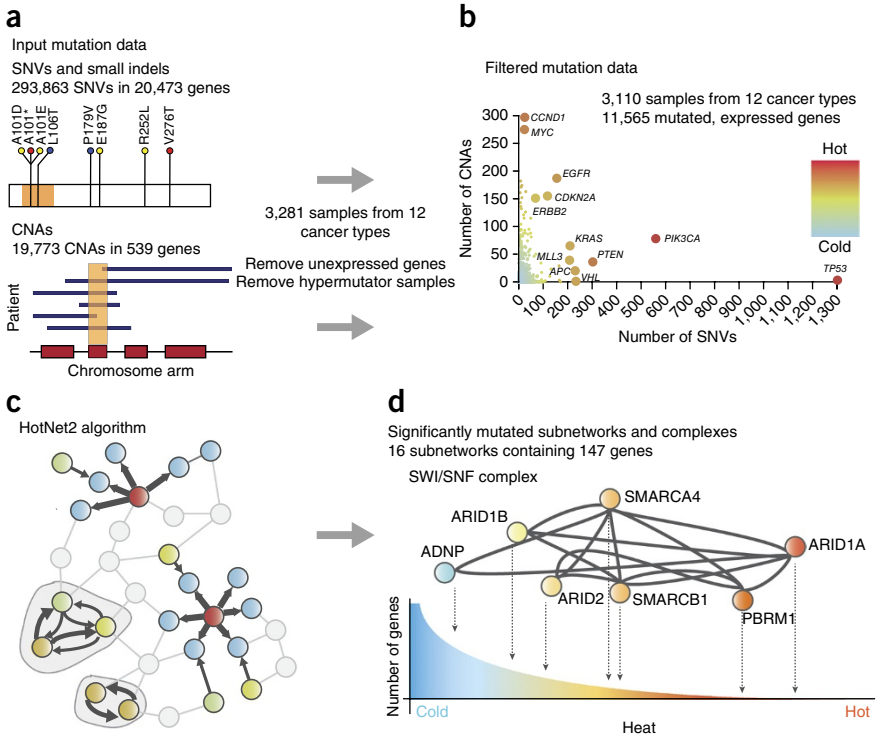

Figure 1 HotNet2 pan-cancer analysis. (**a**) The pan-cancer mutation data comprise SNVs (nonsynonymous SNVs and small indels) and CNAs (amplifications and deletions) in 19,459 genes in 3,281 samples. (**b**) Removing hypermutator samples and genes with few RNA-seq reads in all tumor types leaves 11,565 genes in 3,110 samples for analysis, with a wide range in the number of samples having an SNV (*x* axis) or a CNA (*y* axis) in these genes. The number of samples with SNVs and/or CNAs is shown for each gene, with points colored by the total. (**c**) HotNet2 finds significantly mutated subnetworks using a diffusion process on a protein-protein interaction network. Each node (protein) is assigned a score (heat) according to the frequency and significance of SNVs or CNAs in the corresponding gene. Heat diffuses across the edges of a network. Subnetworks containing nodes that both send and receive a significant amount of heat (outlined) are reported. (**d**) Subnetworks identified by HotNet2 include genes with a wide range of heat scores, including both frequently mutated, known cancer-related genes (hot genes) and rarely mutated genes (cold genes) that are implicated because of their interactions with other cancer types. Thus, HotNet2 delves into the long tail of rarely mutated genes by the analysis of combinations of interacting genes.

mutations ($P < 0.0001$) or mutation clusters ($P < 0.0001$) in comparison to genes not in subnetworks (**Supplementary Tables 6–18** and **Supplementary Note**). Finally, we evaluated a subset of the mutations in these genes using RNA-seq and whole-genome sequencing data from the same samples and found RNA-seq and/or whole-genome sequencing reads that validated 39 mutations in these genes (**Supplementary Table 23** and **Supplementary Note**). These genes might represent new biomarkers for the classification of patients for treatment regimens.

Co-occurrence and mutual exclusivity of mutations in subnetworks

Cancer cells are thought to harbor multiple driver mutations that perturb multiple biological functions[15]. Consistent with this model, we found that four pairs of subnetworks, including TP53 and NOTCH signaling, TP53 and RTK signaling, PI3K signaling and the cohesin complex, and PI3K signaling and the ASCOM complex, exhibited significant co-occurrence ($P < 0.05$, multiple-hypotheses corrected) across the pan-cancer cohort (**Fig. 2b**) or in individual cancer types (**Fig. 2c**). Multiple pairs of genes within these subnetworks showed co-occurring mutations (**Supplementary Table 24**). In contrast, mutually exclusive mutations are typically expected within a pathway and not across pathways[32,33]. We observed significant mutual exclusivity within four of our subnetworks (**Supplementary Table 25**). Intriguingly, the RTK signaling and NFE2L2-KEAP1 subnetworks were the only pair with significant mutual exclusivity across the pan-cancer cohort. This mutual exclusivity was largely due to lung adenocarcinoma (LUAD) samples with mutually exclusive *EGFR* and *KEAP1* mutations (**Supplementary Fig. 15**). This observation is consistent with reports of mutual exclusivity between *EGFR* mutations and *NFE2L2* expression in LUAD[34] and also with *NFE2L2* expression being downstream of EGFR signaling[35]. Examining individual cancers, we found a modest but not statistically significant enrichment for co-occurrence or mutual exclusivity in a few cancer types (**Supplementary Table 26**). Neither within-subnetwork mutual exclusivity nor across-subnetwork co-occurrence is explicitly programmed into the HotNet2 algorithm. These observations support the hypothesis that the HotNet2 subnetworks represent distinct biological functions that are mutated in samples.

TP53, PIK3CA and NOTCH networks

The three largest subnetworks—including a TP53 subnetwork, a PIK3CA subnetwork and a NOTCH subnetwork—contained many well-known cancer-related genes (**Supplementary Figs. 16** and **17**, and **Supplementary Tables 8–10**). Linker genes joined these three subnetworks, demonstrating the extensive cross-talk between well-annotated cancer pathways. Most of these linker genes encoded signaling proteins that have known cancer-related functions (for example, WT1, NOTCH2, PIK3R1, MAP2K4, MAP3K1, HRAS, ATM and STK11). Taken together, 81.9% of the samples contained at least one mutation in these three large subnetworks and linker genes.

HotNet2 pan-cancer analyses also identified a number of newly involved genes (**Supplementary Table 20**) within these three subnetworks. These genes have documented interactions with well-known cancer-relevant genes and similar functions but with somewhat lower mutational frequency (~1%) and were not marked as significant by single-gene tests[20,26–29]. For example, the TP53 subnetwork included *CUL9*. CUL9 sequesters p53 in the cytoplasm, and we found a cluster of 45 missense mutations ($P = 1.32 \times 10^{-8}$) as well as a cluster in the protein structure (false discovery rate (FDR) = 0.025). Another gene of interest was *IWS1*, which is involved in transcriptional elongation and mRNA surveillance. Half (8/16) of the mutations in this gene were inactivating, and this gene also had a cluster of mutations ($P = 0.013$). This subnetwork also contained *CHD8*, encoding an ATP-dependent chromatin-remodeling factor that regulates a wide range of genes[36]. We found three independent signals of *CHD8* inactivation across samples: *CHD8* was deleted in 9 samples in a focal peak from GISTIC; 18 of 58 (31%) of its mutations were inactivating; and this gene had a wide cluster of missense mutations ($P = 6.37 \times 10^{-5}$).

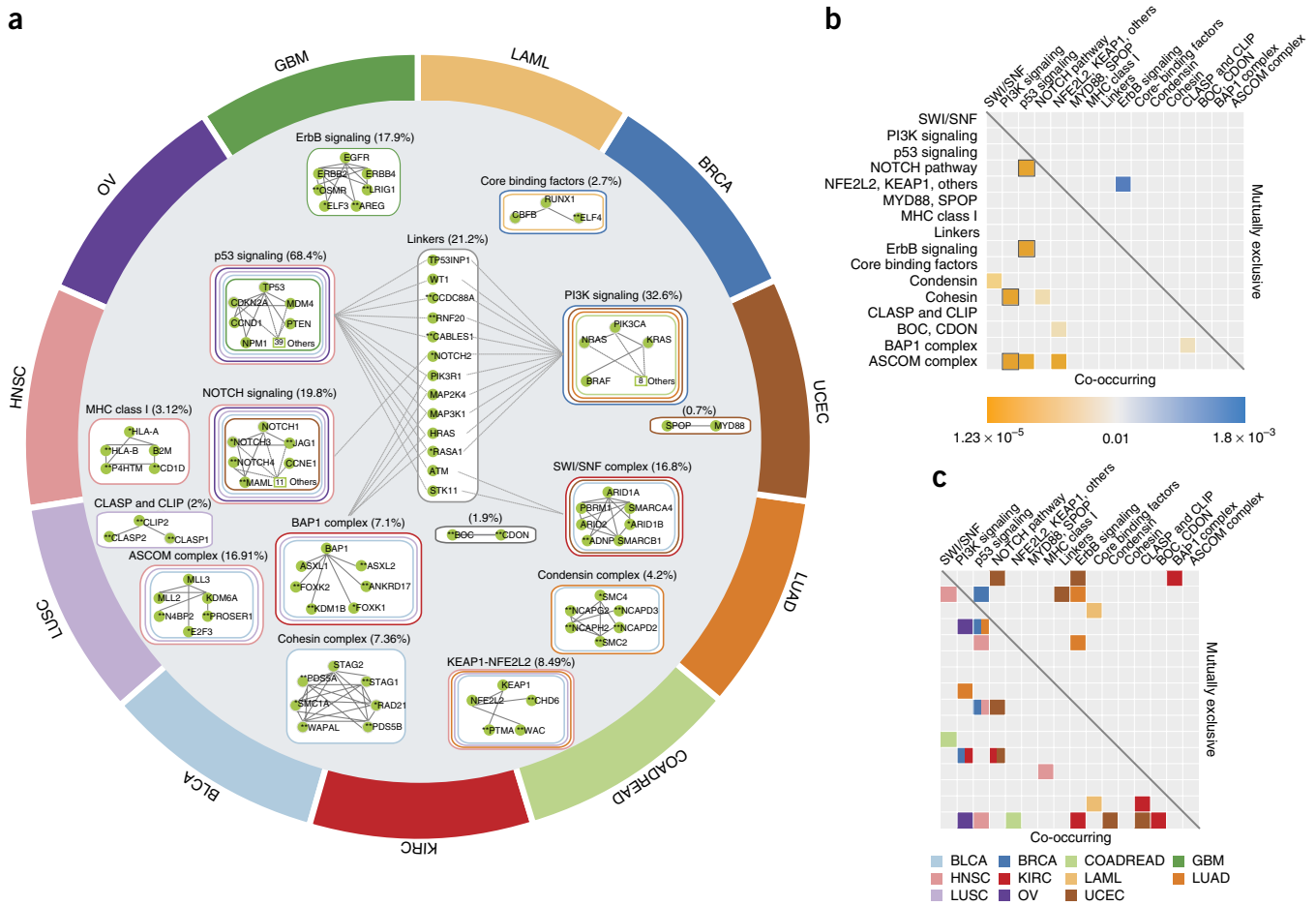

Figure 2 Overview of HotNet2 pan-cancer results. (**a**) HotNet2 consensus subnetworks are arranged near the cancer types where they are enriched for mutations using a force-directed layout (BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; COADREAD, colon adenocarcinoma and rectum adenocarcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LAML, acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; UCEC, uterine corpus endometrioid carcinoma). Colored outlines surrounding each network indicate the cancer types that are enriched for mutations (corrected $P < 0.05$). Interactions between proteins in a subnetwork are derived from the three interaction networks used in our pan-cancer analysis. In the center, there are 13 linkers that are members of more than one consensus subnetwork; lines between linkers and other consensus subnetworks indicate protein-protein interactions between them. Genes with a single asterisk were significant by exactly one of GISTIC2, MuSiC, MutSigCV, Oncodrive or the list of driver genes in ref. 9, whereas genes with two asterisks were not reported by any of these methods. (**b**) Heat map of significant co-occurrence (yellow; lower triangle) and mutual exclusivity (blue; upper triangle) of mutations across all pan-cancer samples in the most frequently mutated HotNet2 pan-cancer consensus and condensin subnetworks ($P < 0.01$, Cochran-Mantel-Haenszel test). Black outlines indicate pairs of subnetworks that have $P < 0.05$ after multiple-hypothesis correction. (**c**) Mutual exclusivity and co-occurrence ($P < 0.01$, Fisher's exact test) within individual cancer types using the same color scheme as in **a**.

In the NOTCH subnetwork, we found rare mutations in *JAG1* and *DLL1*, whose gene products interact with the NOTCH receptors and have some reports of a role in cancer[37]. Moreover, 11 of 24 mutations in *JAG1* were inactivating. The NOTCH subnetwork also included *SHPRH*, which had a significant ($P < 8 \times 10^{-5}$) cluster of missense mutations (**Supplementary Fig. 18**).

SWI/SNF complex

The sixth most mutated HotNet2 pan-cancer subnetwork (16.8% of samples) included multiple members of the SWI/SNF chromatin-remodeling complex (**Fig. 3a** and **Supplementary Table 12**). Mutations in this complex have previously been reported in several cancers[38,39], including TCGA samples[40]. Our HotNet2 pan-cancer analysis demonstrated the prevalence of mutations in SWI/SNF components: at least 1.5% of the samples from each of the 12 cancer types contained a mutation in this subnetwork. Kidney renal clear cell carcinoma (KIRC) ($P < 1 \times 10^{-15}$), uterine corpus endometrial carcinoma (UCEC) ($P = 7 \times 10^{-10}$) and bladder urothelial carcinoma (BLCA) ($P = 1.8 \times 10^{-8}$) were enriched for mutations in this subnetwork, and several genes were enriched for mutations in specific cancer types, including *PBRM1* in KIRC ($P < 1 \times 10^{-15}$) and *ARID1A* in both BLCA ($P = 4.8 \times 10^{-8}$) and UCEC ($P < 1 \times 10^{-15}$). The subnetwork also contained *ARID1B*, which has been reported to have somatic mutations in juvenile neuroblastoma[41] and germline mutations in Coffin-Siris syndrome[42].

In addition to known members of the SWI/SNF complex, the subnetwork included *ADNP*. *ADNP* mutations have not previously been reported in cancer and were not considered significant by the three individual gene-scoring methods applied. However, ADNP has a known interaction with the SWI/SNF complex[43] and protects against oxidative stress in neuronal cells[44], suggesting that in rare cases *ADNP* mutations contribute to tumorigenesis. Thus, HotNet2

Table 1 A subset of candidate cancer genes identified by HotNet2 but not by single-gene tests of significance

Gene	SNVs	CNAs	Cancer enrichment (*S*)	Function
ADNP	21	0		Homeobox transcription factor with nine zinc fingers found in the SWI/SNF complex; mediates neuroprotective responses to cellular growth and regulates cancer cell proliferation
ASXL2	30	0		BAP1 complex–mediated chromatin modulation and transcriptional regulation; has an opposing role to ASXL1
CCDC88A	38	0		Girdin family member with a key role in the PI3K and AKT signaling pathways that might be involved in metastasis when overexpressed
CHD8[a]	49	9		DNA helicase that acts as a chromatin-remodeling factor and suppresses transcription; suppresses *TP53* and negatively regulates β-catenin in WNT signaling; is essential for embryonic development
CUL9	48	0		Involved in p53 localization; critical regulator of the cell cycle and quiescence
ELF3[a]	19	0	BLCA, COADREAD	Transcriptional activator that binds ETS motifs; might be a downstream effector of the ERBB2 signaling pathway
EPHA3[a]	50	3		RTK with possible roles in BRCA, COADREAD, GBM, HNSC, lung and pancreatic cancers
FOXK2	13	12		Forkhead transcription factor whose functions are cell cycle regulated; recruits AP-1 and functions in DNA mismatch repair
IWS1	16	0		Involved in transcriptional elongation and transcriptional surveillance
JAG1	24	0		Ligand for multiple NOTCH receptors and involved in the mediation of NOTCH signaling; might have a role in AML, BRCA, COADREAD, GBM, OV and pancreatic cancers
KDM1B	14	0		Histone demethylase that acts as a corepressor; along with BAP1, regulates cell growth
KLF5	12	36	BLCA, COADREAD, HNSC	Kruppel-like transcriptional activation factor; regulates pluripotency and cellular growth
MLL5	30	0		Histone methyltransferase that acts as an important cell cycle regulator; high expression is associated with a favorable outcome in AML
NCAPH2	19	0		Non-SMC condensin II subunit; critical for mitotic chromosome assembly
NOTCH3	93	4	OV	Receptor for Jagged1, Jagged2 and Delta1 to regulate cell fate through transcriptional activation; mutations cause CADASIL
RNF20	27	0		E3 ubiquitin-protein ligase for H2BK120ub1; putative tumor suppressor
SHPRH	39	0		E3 ubiquitin-protein ligase for PCNA involved in DNA repair
SMG1	51	0		mRNA surveillance through nonsense-mediated mRNA decay
SMG7	23	0	LUSC	mRNA surveillance through nonsense-mediated mRNA decay
STAG1	31	0		Cohesin subunit involved in sister chromatid adhesion following DNA replication
WAC	19	0		Regulates cell cycle progression by linking transcription to H2BK120ub1

For each gene, the number of samples with at least one SNV or CNA in the gene and the cancers enriched for mutations ($P < 0.05$, corrected) are listed. More information on these genes as well as other candidate driver genes is provided in **Supplementary Table 20**.
[a]Listed as a cancer driver by Oncodrive or GISTIC.

analyses broaden the view of mutations in the SWI/SNF complex to additional cancer types and additional interacting proteins.

BAP1 complex and interactors

Another HotNet2 pan-cancer subnetwork (mutated in 7.1% of samples) overlapped the BAP1 complex (**Fig. 3b** and **Supplementary Table 13**). This subnetwork included *BAP1*, *ASXL1*, *ASXL2*, *FOXK1* and *FOXK2*, all encoding members of the BAP1 core complex[45], as well as two additional interacting factors, *KDM1B* and *ANKRD17*. Only *BAP1* and *ASXL1* were significant by individual-gene scores—the other genes harbored rare mutations across many cancer types, a subtle signal identified by HotNet2 pan-cancer analysis. This subnetwork was mutated in at least six samples from each cancer type, demonstrating the breadth of mutations in the BAP1 complex.

BAP1 inactivation has been reported in several cancers[45]. We found the subnetwork enriched for mutations in KIRC ($P = 2 \times 10^{-4}$), as previously reported[46]. Consistent with the results of Peña-Llopis *et al.*[46], we found that mutations in the *BAP1* gene were mutually exclusive ($P < 7.2 \times 10^{-3}$) with mutations in the *PBRM1* gene in KIRC. We found that mutations in the SWI/SNF and BAP1 complexes showed even greater mutual exclusivity ($P = 9.4 \times 10^{-5}$) in KIRC because of mutations in additional genes in these complexes in addition to *BAP1* and *PBRM1*, respectively (**Supplementary Note**). This mutual exclusivity suggests that mutations in these complexes define different subtypes of kidney cancer. Supporting this hypothesis, we observed that inactivating mutations in the BAP1 complex were enriched ($P < 3.4 \times 10^{-8}$) for samples in the third mRNA expression subtype described in ref. 3 (**Fig. 3c**).

We found that a large fraction of the mutations in *BAP1*, *ASXL1* and *ASXL2* in different cancer types were inactivating mutations, demonstrating alternative strategies for the inactivation of the BAP1 complex. In addition, 6 of 13 missense mutations in *FOXK2* affected the forkhead transcription factor domain or forkhead-associated domain, which might inactivate the DNA-binding properties of FOXK2. Finally, we examined the mutations in *KDM1B*, which encodes a protein that is involved in histone H3 lysine 4 (H3K4) methylation[47] but is not considered to be a core part of the BAP1 complex. We found that 12 of 19 mutations in *KDM1B* (including 10 of 16 missense mutations) fell in the C-terminal amino-oxidase domain that is important for lysine-specific demethylation of histones[48]. Moreover, two of the three *KDM1B* mutations in lung squamous cell carcinoma (LUSC) and LUAD were inactivating, and these were also mutually exclusive with *BAP1* inactivating mutations, suggesting that *KDM1B* mutations might have a role in cancer.

Cohesin and condensin

HotNet2 pan-cancer analysis identified four of five members of the cohesin complex as a significantly mutated subnetwork (7.3% of samples; **Fig. 4a** and **Supplementary Table 15**). Although named for its role in sister chromatid cohesion, the cohesin complex has recently been implicated more broadly in gene regulation[49–51], and its role in myeloid leukemia was only recently reported[52]. We found that the cohesin complex was universally mutated across cancer types (>4% of samples in each cancer type). Moreover, the mutations in the complex were spread uniformly across the genes, with no gene in the complex mutated in more than 1.9% of samples. This pattern

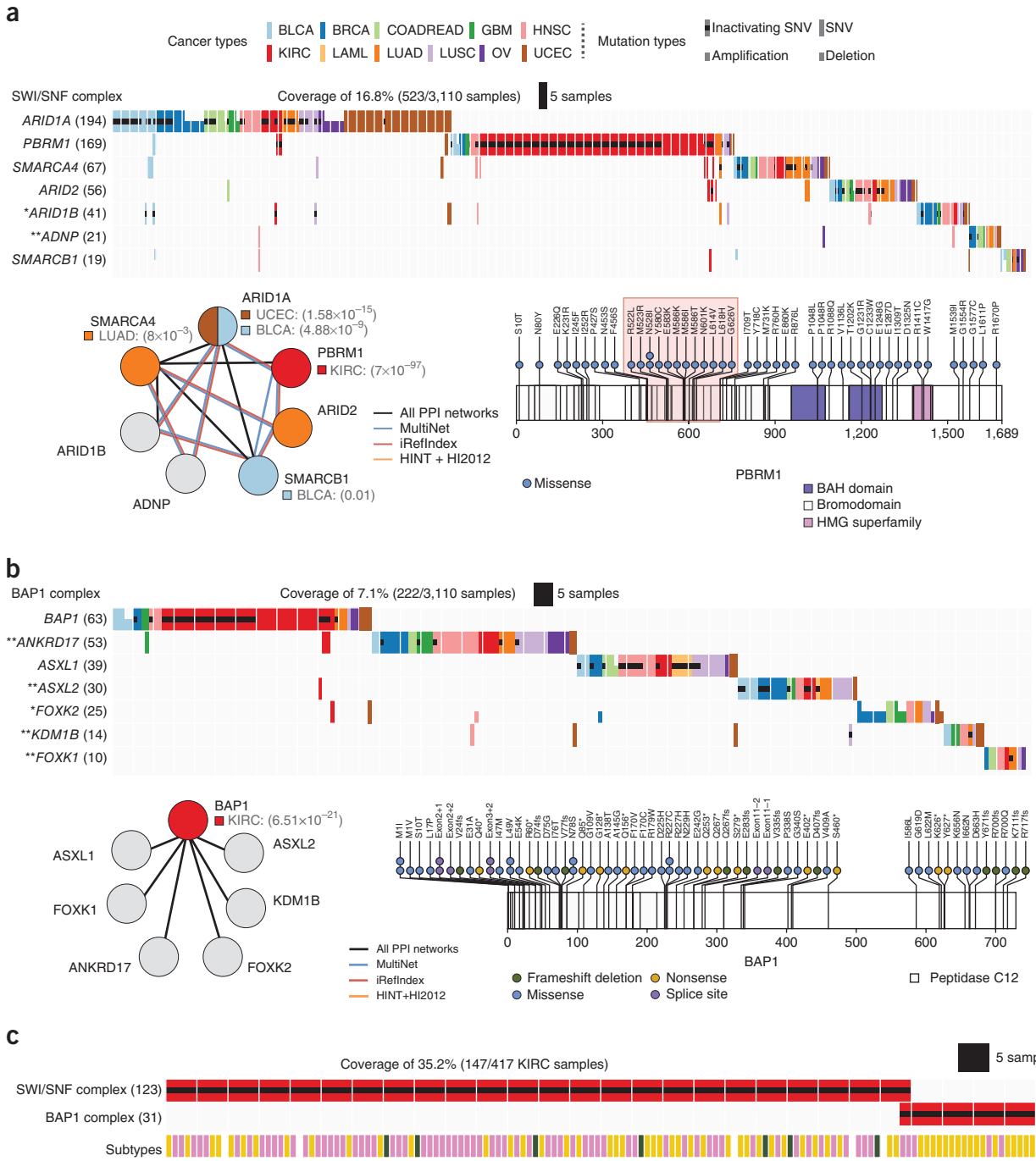

Figure 3 HotNet2 pan-cancer subnetworks overlapping SWI/SNF and BAP1 complexes. (**a**) Subnetwork containing members of the SWI/SNF complex, including the BAF proteins ARID1A and ARID1B, the PBAF proteins PBRM1 and ARID2, the catalytic core member SMARCA4, SMARCB1 and ADNP. (**a**) Top, mutation matrix showing the samples (colored by cancer type) with a mutation of the indicated type: full ticks represent SNVs, indels and splice-site mutations; upticks and downticks represent amplifications and deletions, respectively. A black dot corresponds to samples with an inactivating mutation in the gene and indicates that the genes contain at least one of the following mutations: nonsense, frameshift indel, nonstop or splice site. The number of samples with mutations in a gene is given in parentheses; marks are defined as in **Figure 2a**. Bottom left, interactions between proteins in the subnetwork from each interaction network are colored according to mutually enriched cancer type with corresponding *P* values. PPI, protein-protein interaction. Bottom right, the PBRM1 protein sequence exhibited significant clustering of missense alterations ($P = 1.6 \times 10^{-5}$) in a 105-amino-acid bromodomain, a region that was reported to be mutated in a different renal clear cell carcinoma cohort[39] but not in the TCGA KIRC publication[3]. Splice-site mutations are annotated with the nearest exon, where +1 and +2 refer to the adjacent 3′ splice donor and −1 and −2 refer to the adjacent 5′ splice acceptor. (**b**) Subnetwork containing members of the BAP1 complex, including the core PR-DUB complex, comprised of the deubiquitinating enzyme BAP1 and the Polycomb group proteins ASXL1 and ASXL2, as well as the BAP1-interacting proteins: ANKRD17, FOXK1, FOXK2 and KDM1B. Colors, marks and panel organization are as in **a**. (**c**) Inactivating mutations across samples (columns) in the SWI/SNF and BAP1 complexes (rows) in KIRC. The bottom row shows the mRNA expression classification of each sample[3]. The mutations in these complexes are surprisingly mutually exclusive in KIRC ($P < 3.6 \times 10^{-4}$, Fisher's exact test, corrected), and the BAP1 complex is significantly enriched in mutations in the third expression subtype ($P < 3.4 \times 10^{-8}$, Fisher's exact test).

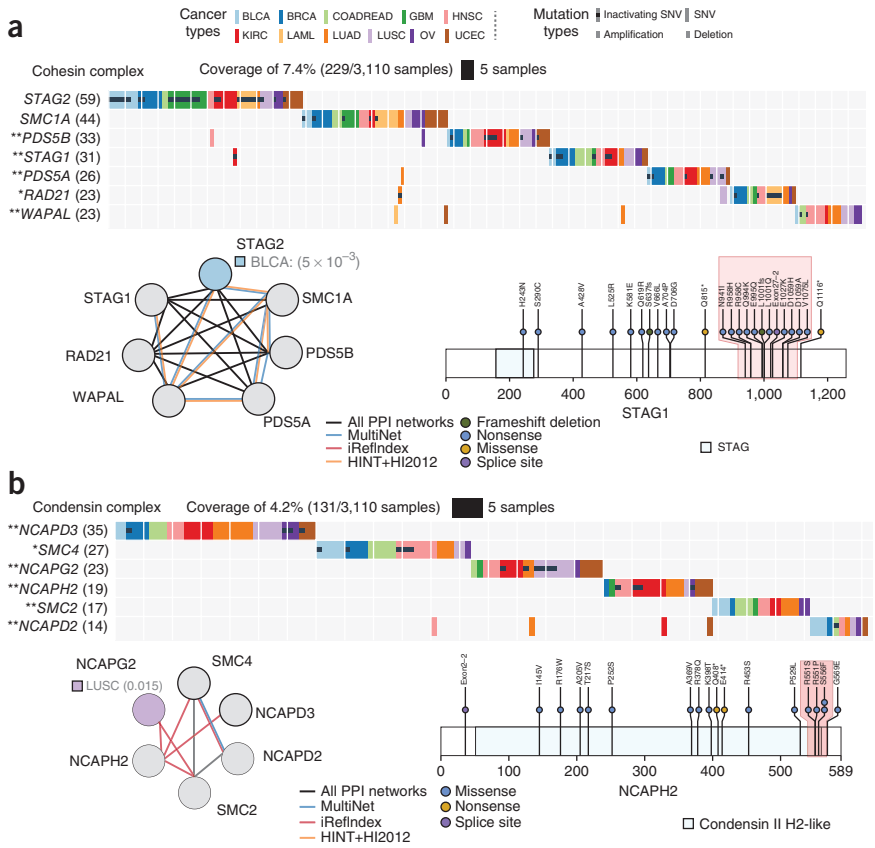

Figure 4 HotNet2 pan-cancer subnetworks overlapping the cohesin and condensin complexes. (**a**) Cohesin consensus subnetwork and its mutations. Colors and marks are as defined in **Figure 3a**. None of the genes are mutated in more than 1.9% of the samples, but the subnetwork is mutated in >4% of the samples in each cancer type. STAG1 exhibits significant ($P < 6 \times 10^{-5}$) clustering of missense alterations across 135 residues (highlighted) in the Pfam-B domain (PFAM, PB002581), a pattern suggesting inactivation of the corresponding domain. (**b**) Condensin consensus subnetwork and its mutations. Top, mutation matrix showing five genes in the condensin I and condensin II complexes. Only one gene, *SMC4*, was significant by individual gene scores. Bottom left, a subnetwork consisting of NCAPD2 and SMC4, both members of condensin I, was significantly mutated in BLCA, and a subnetwork consisting of NCAPD3, NCAPG2 and NCAPH2, all members of condensin II, was significantly mutated in LUAD and LUSC. At the gene level, *NCAPD2* was significantly mutated in BLCA, *SMC4* was significantly mutated in BLCA and HNSC, *NCAPD3* was significantly mutated in LUAD and *NCAPG2* was significantly mutated in LUSC. Bottom right, NCAPH2 shows a significant ($P < 2.6 \times 10^{-4}$) cluster of missense alterations between Arg551 and Ser556.

of mutations complicates the identification of recurrent mutations in individual genes; indeed, only half of the genes in the complex (*STAG2*, *SMC1A* and *RAD21*) were significant by at least one of the three gene scores applied.

Mutations in some of these genes have recently been reported to be significant in several cancers. We found enrichment for mutations in the subnetwork in BLCA ($P = 7 \times 10^{-4}$); this enrichment derived largely from enrichment for mutations in *STAG2* in BLCA ($P = 0.005$), which was recently reported[53]. *STAG2* had a significantly higher fraction of inactivating mutations than other genes in the subnetwork (53% for *STAG2* in comparison to 28% for the subnetwork as a whole); these inactivating mutations were not only in BLCA but also across multiple cancer types, with multiple inactivating mutations in acute myeloid leukemia (LAML) and colon and rectal adenocarcinoma (COADREAD). In addition, BLCA samples without *STAG2* inactivating mutations harbored rare inactivating mutations in several other cohesin genes. All mutations in *RAD21* in LAML samples were inactivating, and BRCA and KIRC samples harbored inactivating mutations in *STAG1*. In addition, we observed a significant clustering of missense mutations in *STAG1* ($P = 6 \times 10^{-5}$), and the broad span of the cluster (135 residues) is indicative of inactivation. STAG1 has been shown to function as a transcriptional coactivator[50,51], and mutation of *STAG1* might thus have another role in cancer apart from affecting genome stability. Together, these results show that mutational inactivation of the cohesin complex occurs broadly across cancer types and across genes within the complex.

HotNet2 also identified two subnetworks containing six proteins in the condensin complex, in HotNet2 runs from individual interaction networks. The combined subnetwork was mutated in 4.2% of samples (**Fig. 4b** and **Supplementary Table 6**). Only *SMC4* was reported to be significant by at least one of the individual-gene scores. A subnetwork consisting of *NCAPD2*, *SMC2* and *SMC4*, all members of the condensin I form of the complex, was significantly mutated in BLCA ($P = 6.2 \times 10^{-6}$). Condensin I is thought to primarily be involved in sister chromatid condensation during mitosis[54,55], suggesting that these mutations promote genome instability. In contrast, a subnetwork consisting of *NCAPD3*, *NCAPG2* and *NCAPH2*, all members of the condensin II form of the complex, was significantly mutated in LUAD ($P = 0.04$) and LUSC ($P = 0.002$), and the majority (4/7) of *NCAPG2* mutations in LUSC were inactivating. Condensin II is generally involved in gene regulatory processes[54,55], suggesting a different phenotype for these mutations. In addition, we found a significant ($P = 0.002$) cluster of missense mutations in *NCAPH2* (**Fig. 4b**), implying that mutations in this region of unknown function might be important for the deregulation of condensin. We also note that it was recently observed that expression of *NCAPD3* is positively associated with recurrence-free survival[56]. Finally, RNA-seq and whole-genome sequencing data from the same samples provided further validation of the somatic mutations in *SMC2*, *SMC4*, *NCAPD2*, *NCAPD3*, *NCAPH2* and *NCAPG2* and showed that some of these mutations were expressed (**Supplementary Table 23** and **Supplementary Note**). Our HotNet2 pan-cancer analysis suggests that multiple cancer types harbor rare mutations in the cohesin and condensin complexes, supporting a proposed tumor-suppressor role for these complexes[49,54,55].

DISCUSSION

We present a new approach for identifying combinations of somatic aberrations in different cancer types, using our HotNet2 algorithm to analyze a high-quality pan-cancer data set of 3,281 samples from 12 cancer types. This analysis represents the largest network analysis of somatic aberrations across multiple cancer types. First, we recover

many classic cancer pathways, such as TP53, PI3K, NOTCH and RTK signaling, automatically from a large-scale interaction network, demonstrating the power of the pan-cancer network approach. Second, we highlight the extensive cross-talk between these pathways, overlaps that are often overlooked in analyses that treat pathways as distinct gene lists. Third, we find pathways and complexes whose role in cancer has only recently been appreciated, such as the SWI/SNF chromatin-remodeling complex[38] and the BAP1 complex[45]. Fourth, we find that several pairs of HotNet2 subnetworks have co-occurring mutations, although mutations are mostly mutually exclusive within subnetworks. This finding supports the hypothesis that these subnetworks represent distinct biological functions that are mutated in samples. Finally, we identify a number of new mutated subnetworks with potential roles in cancer, including the cohesin and condensin complexes[54], MHC class I proteins and the telomerase complex. These subnetworks have rare mutations in nearly all cancer types, making them difficult to detect without a sensitive pan-cancer network approach that examines combinations of genes across multiple cancer types.

The HotNet2 subnetworks contain 92 genes that are rarely mutated, both in individual cancer types and across the pan-cancer cohort, and are not reported as significant by single-gene tests. Nearly all of the subnetworks contain such genes, which are identified by the combination of their mutations and interactions across cancer types. Some of these rarely mutated genes are inevitably false positive predictions of the analysis, but many (including *SHPRH*, *CUL9*, *CHD8*, *RNF20*, *JAG1*, *ELF3*, *STAG1*, *NCAPH2* and others) exhibit either mutational clustering or protein interactions that support a role for the observed somatic aberrations in cancer (**Supplementary Tables 6–18**). In addition, we find that well-characterized mutations in a single gene in one cancer type (for example, inactivating mutations in *BAP1* in KIRC) are replaced in other cancer types by rare mutations in other members of the same complex (for example, inactivating mutations in *ASXL1*, *ASXL2*, *FOXK2* and *KDM1B*). Such observations suggest that pan-cancer network analyses might prove useful in translating diagnostic or therapeutic approaches that were developed in one cancer type to other cancer types.

Our analysis complements other recent pan-cancer analyses, including studies that analyzed only one type of aberration[11–13] or restricted their focus to recurrent aberrations[57] (**Supplementary Table 27** and **Supplementary Note**). The HotNet2 pan-cancer network approach identifies combinations of rare and common mutations in groups of interacting genes—combinations that were not apparent by the analysis of single genes, known pathways or single cancer types. Indeed, we observe that many of the identified subnetworks contain genes altered by both SNVs and CNAs, demonstrating that integrating data on multiple types of aberrations is beneficial when jointly analyzing multiple cancer types that might have different mutational landscapes. Pan-cancer network analysis of multiple aberration types thus provides an alternative approach to prioritize rare mutations for further experimental characterization.

As with any computational approach, our findings are limited by the quality and quantity of the input data. Further power is anticipated by including additional samples[11], additional types of genetic and epigenetic aberrations, and better interaction networks. For example, structural variants, noncoding variants and methylation data were not included, with the first two being unavailable for most TCGA samples. This lack of data in combination with false negatives in the analyzed data (for example, due to difficulties in the identification of indels and subclonal variants) implies that our analysis likely underestimates the number and frequency of mutated subnetworks across cancer types. In contrast, we note that some genes that are highly significant by individual-gene scores are not reported in our network analysis; often this is owing to problems with the interaction network. Improved knowledge of the human interactome—including more systematic efforts to record known interactions, measure additional interactions and determine the tissue specificity of interactions—is needed to increase coverage and reduce possible ascertainment bias.

Finally, the HotNet2 algorithm introduced here is suitable for other applications, both biological and non-biological. In particular, genome-wide association studies (GWAS) and other studies of genetic diseases face an analogous problem in identifying combinations of genetic variants with a statistically significant association with a phenotype. With an appropriate gene score, the HotNet2 algorithm can be applied to such data.

URLs. HI2012 interactome, http://interactome.dfci.harvard.edu/; HotNet2 pan-cancer analysis website, http://compbio.cs.brown.edu/pancancer/hotnet2/; RNA expression data used for the TCGA pan-cancer data set, https://www.synapse.org/#!Synapse:syn1734155; pan-cancer mutations with additional germline variant filtering, https://www.synapse.org/#!Synapse:syn1729383; HotNet2 software release, http://compbio.cs.brown.edu/software.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank F. Roth for his assistance in constructing the HINT+HI2012 interaction network. We gratefully acknowledge the contributions of the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group. This work is supported by US National Science Foundation (NSF) grant IIS-1016648 and US National Institutes of Health (NIH) grants R01HG005690, R01HG007069 and R01CA180776 to B.J.R. and by National Human Genome Research Institute (NHGRI) grant U01HG006517 to L.D. B.J.R. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship and an NSF CAREER Award (CCF-1053753). M.D.M.L. is supported by NSF fellowship GRFP DGE 0228243. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Data for HI2012, created by the Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute, are supported by the NHGRI of the US NIH, the Ellison Foundation and the Dana-Farber Cancer Institute Strategic Initiative.

AUTHOR CONTRIBUTIONS

M.D.M.L., F.V., H.-T.W. and B.J.R. designed the HotNet2 algorithm. M.D.M.L., F.V., H.-T.W., J.R.D., J.V.E., J.L.T., Y.K. and B.J.R. performed pan-cancer network analysis, analyzed results and benchmarked algorithms. A.P., J.R.D., Y.C. and G.A.R. analyzed mutation clusters in genes. B.N., M.M. and L.D. provided MuSiC gene scores, assisted with figures and generated mutation validation data. M.S.L., G.G., A.G.-P., D.T. and N.L.-B. provided MutSigCV and Oncodrive gene scores. M.D.M.L., F.V., H.-T.W., J.R.D. and B.J.R. wrote the manuscript with input from all authors. B.J.R. conceived and supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
2. Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

3. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **99**, 43–49 (2013).
4. Cancer Genome Atlas Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
6. Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
7. Cancer Genome Atlas Network. Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
8. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
9. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
10. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
11. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
12. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
13. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**, 1134–1140 (2013).
14. Weinstein, J.N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
15. Hanahan, D. & Weinberg, R.a. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
16. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* **18**, 507–522 (2011).
17. Vandin, F., Clay, P., Upfal, E. & Raphael, B.J. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac. Symp. Biocomput.* **2012**, 55–66 (2012).
18. Grasso, C.S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
19. Hofree, M., Shen, J.P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013).
20. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
21. Das, J. & Yu, H. HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6**, 92 (2012).
22. Yu, H. *et al.* Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8**, 478–480 (2011).
23. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLOS Comput. Biol.* **9**, e1002886 (2013).
24. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).
25. Hoadley, K.A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
26. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
27. Tamborero, D., Lopez-Bigas, N. & Gonzalez-Perez, A. Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS ONE* **8**, e55489 (2013).
28. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
29. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
30. Ye, J., Pavlicek, A., Lunney, E.A., Rejto, P.A. & Teng, C.-H. Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics* **11**, 11 (2010).
31. Ryslik, G.A., Cheng, Y., Cheung, K.-H., Modis, Y. & Zhao, H. Utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics* **14**, 190 (2013).
32. Yeang, C.-H., McCormick, F. & Levine, A. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.* **22**, 2605–2622 (2008).
33. Vandin, F., Upfal, E. & Raphael, B.J. *De novo* discovery of mutated driver pathways in cancer. *Genome Res.* **22**, 375–385 (2012).
34. Solis, L.M. *et al.* Nrf2 and Keap1 abnormalities in non–small cell lung carcinoma and association with clinicopathologic features. *Clin. Cancer Res.* **16**, 3743–3753 (2010).
35. Yamadori, T. *et al.* Molecular mechanisms for the regulation of Nrf2-mediated cell proliferation in non-small-cell lung cancers. *Oncogene* **31**, 4768–4777 (2012).
36. Thompson, B.A., Tremblay, V., Lin, G. & Bochar, D.A. CHD8 is an ATP-dependent chromatin remodeling factor that regulates β-catenin target genes. *Mol. Cell. Biol.* **28**, 3894–3904 (2008).
37. Greife, A. *et al.* Canonical Notch signalling is inactive in urothelial carcinoma. *BMC Cancer* **14**, 628 (2014).
38. Wilson, B.G. & Roberts, C.W.M. SWI/SNF nucleosome remodellers and cancer. *Nat. Rev. Cancer* **11**, 481–492 (2011).
39. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene *PBRM1* in renal carcinoma. *Nature* **469**, 539–542 (2011).
40. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat. Genet.* **45**, 592–601 (2013).
41. Sausen, M. *et al.* Integrated genomic analyses identify *ARID1A* and *ARID1B* alterations in the childhood cancer neuroblastoma. *Nat. Genet.* **45**, 12–17 (2013).
42. Tsurusaki, Y. *et al.* Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. *Nat. Genet.* **44**, 376–378 (2012).
43. Mandel, S. & Gozes, I. Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. *J. Biol. Chem.* **282**, 34448–34456 (2007).
44. Steingart, R.A. & Gozes, I. Recombinant activity-dependent neuroprotective protein protects cells against oxidative stress. *Mol. Cell. Endocrinol.* **252**, 148–153 (2006).
45. Carbone, M. *et al.* BAP1 and cancer. *Nat. Rev. Cancer* **13**, 153–159 (2013).
46. Peña-Llopis, S. *et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nat. Genet.* **44**, 751–759 (2012).
47. Fang, R. *et al.* Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. *Mol. Cell* **39**, 222–233 (2010).
48. Shi, Y. *et al.* Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell* **119**, 941–953 (2004).
49. Xu, H., Tomaszewski, J.M. & McKay, M.J. Can corruption of chromosome cohesion create a conduit to cancer? *Nat. Rev. Cancer* **11**, 199–210 (2011).
50. Rubio, E.D. *et al.* CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. USA* **105**, 8309–8314 (2008).
51. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20**, 578–588 (2010).
52. Kon, A. *et al.* Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat. Genet.* **45**, 1232–1237 (2013).
53. Solomon, D.A. *et al.* Frequent truncating mutations of *STAG2* in bladder cancer. *Nat. Genet.* **45**, 1428–1430 (2013).
54. Wood, A.J., Severson, A.F. & Meyer, B.J. Condensin and cohesin complexity: the expanding repertoire of functions. *Nat. Rev. Genet.* **11**, 391–404 (2010).
55. Hirano, T. Condensins: universal organizers of chromosomes with diverse functions. *Genes Dev.* **26**, 1659–1678 (2012).
56. Lapointe, J. *et al.* hCAP-D3 expression marks a prostate cancer subtype with favorable clinical behavior and androgen signaling signature. *Am. J. Surg. Pathol.* **32**, 205–209 (2008).
57. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).

ONLINE METHODS

Somatic aberration data. SNVs, indels and splice-site mutations were extracted from TCGA pan-cancer analysis on Synapse (syn1710680), and CNAs were extracted from GISTIC2 output via Firehose. We restricted our focus to the 3,276 samples containing both SNV and CNA data. We removed 71 samples identified as ultramutators in syn1729383 and an additional 95 samples with an unusually high number of aberrations (>400 SNVs or CNAs). We selected the threshold of 400 aberrations per sample, as the derivative of the number of mutations per sample starts increasing rapidly beyond this value (**Supplementary Fig. 19**). We removed genes without CNAs that contained SNVs in >2% of samples but were not identified as significant ($q < 0.05$) by MutSigCV[20]. Finally, we used only those genes that had at least three reads from RNA-seq data in at least 70% of samples of at least one of the cancer types, as described in syn1734155 (see URLs). The resulting data set contained aberrations in 11,565 genes and 3,110 samples (**Supplementary Fig. 1**). We used gene scores from mutation frequency and MutSigCV $-\log_{10} q$ values. Nonsense mutations, frameshift indels, nonstop mutations or splice-site mutations were classified as inactivating according to ref. 12. We used three interaction networks: HINT+HI2012, a combination of the HINT network[21] and the HI-2012 (ref. 22) set of protein-protein interactions; MultiNet[23]; and iRefIndex[24]. Additional details on the data sets are provided in the **Supplementary Note**.

HotNet2. We developed the HotNet2 (HotNet diffusion-oriented subnetworks) algorithm to identify subnetworks of a genome-scale interaction network that were mutated more than expected by chance. Although interaction networks have proven useful in analyzing various types of genomic data[58], statistically robust identification of significantly mutated subnetworks is a difficult problem with several major challenges (**Supplementary Note**). HotNet2 addresses these challenges and identifies significantly mutated subnetworks of a genome-scale interaction network, using an insulated heat diffusion process that considers both the scores on individual genes as well as the topology of interactions between the corresponding proteins (**Supplementary Fig. 3**).

The input to HotNet2 is a heat vector $\vec{h}$ that contains the score (for example, mutation frequency) for each gene g and a graph $G = (V,E)$, where each node corresponds to a gene and each edge corresponds to an interaction between the encoded proteins. HotNet2 performs the following steps.

1. Heat diffusion. HotNet2 employs an insulated heat diffusion process[59,60] that captures the local topology of the interaction network surrounding a protein. At each time step, nodes in the graph pass to and receive heat from their neighbors but also retain a fraction β of their heat, governed by an insulating parameter β. The process is run until equilibrium is reached; the amount of heat on each node at equilibrium thus depends on the initial heat for the node, the local topology of the network around the node and the β value. If a unit heat source is placed at node j (for example, a mutation in g_j in one sample), then the amount of heat on node i is given by the (i,j) entry of the diffusion matrix F defined by:

$$F = \beta\left(I - (1-\beta)W\right)^{-1}$$

where

$$W_{ij} = \frac{1}{\deg(j)}$$

if node i interacts with node j, 0 otherwise. Thus, W is a normalized adjacency matrix of the graph G. We interpret $F(i,j)$ as the influence that a heat source placed on g_j has on g_i. The insulated heat model can also be described in terms of a random walk with restart (**Supplementary Note**). Note that the insulated diffusion process is generally asymmetric, i.e., $F(i,j) \neq F(j,i)$. The diffusion matrix F depends only on the graph G and not on the heat vector $\vec{h}$. Therefore, the influence (for a given β) needs to be computed only once for a given interaction network.

2. Exchanged heat matrix. The insulated heat diffusion process described above encodes the local topology of the network, assuming that unit heat is placed on nodes. To jointly analyze the network topology and gene scores given by the initial heat vector h, we define the exchanged heat matrix E as:

$$E = FD_{\vec{h}}$$

where $D_{\vec{h}}$ is the diagonal matrix with entries $\vec{h}$. $E(i,j) = F(i,j)\vec{h}(j)$ is the amount of heat that diffuses from node g_j to node g_i on the network when $\vec{h}(j)$ heat is placed on g_j, which we interpret as the similarity of g_j,g_i. As the diffusion matrix F is not symmetric and in general $\vec{h}(i) \neq \vec{h}(j)$, the similarity $E(i,j)$ is also not symmetric (**Supplementary Note**).

3. Identification of hot subnetworks. We form a weighted directed graph H whose nodes are all measured genes. If $E(i,j) > \delta$, then there is a directed edge from node j to node i of weight $E(i,j)$. HotNet2 identifies strongly connected components in H. A strongly connected component C in a directed graph is a set of nodes such that for every pair u,v of nodes in C there is a path from u to v.

4. Statistical test for subnetworks. HotNet2 employs a statistical test to determine the significance of the number and size of the subnetworks determined in the previous step. The statistical test is the same as the two-stage statistical test introduced in the original HotNet algorithm[16,17] (**Supplementary Figs. 20–23**, **Supplementary Table 28** and **Supplementary Note**). HotNet2 is available online (see URLs).

HotNet2 parameters. HotNet2 has two parameters β and δ, and it selects values for both of these parameters using automated procedures. β is selected from the protein-protein interaction network, independently of any gene scores (**Supplementary Fig. 24**, **Supplementary Table 29** and **Supplementary Note**). We evaluated the sensitivity of the HotNet2 results to the value of β and found that varying β by ±10% had only a minor effect on the results, with at most seven genes (3.8% of the total) added or removed from the subnetworks (**Supplementary Table 29**). The value of δ is chosen such that large connected components are not found using the observed gene score distribution on random networks with the same degree distribution as the observed network (**Supplementary Fig. 25**, **Supplementary Table 30** and **Supplementary Note**). We evaluated the sensitivity of the HotNet2 results to the value of δ and found that varying δ by ±5% changed at most 35 genes (12.3% of the total) in the subnetworks (**Supplementary Table 30**).

Comparison of HotNet2 to other algorithms. HotNet2 extends our previous algorithm HotNet[17,18] in several directions. First, HotNet2 employs an insulated heat diffusion process that better encodes the local topology of the neighborhood surrounding a protein in the interaction network. Second, HotNet2 uses an asymmetric influence $F(i,j)$ between two proteins g_i and g_j to derive a directed measure of similarity $E(i,j)$ between them, whereas HotNet derives a symmetric influence. Third, HotNet2 identifies strongly connected components in the directed graph H, whereas HotNet computes connected components in an undirected graph. These differences enable HotNet2 to effectively detect significant subnetworks in data sets in which the number of samples is order(s) of magnitude larger than considered by HotNet and in which the mutational frequencies, or scores, occupy a broad range (from very common to extremely rare) (see **Supplementary Fig. 2**).

Expanding on the third point, when undirected diffusion algorithms such as HotNet or related network-propagation algorithms[19] are run on large data sets containing a wide range of gene scores (for example, the pan-cancer data set), many of the resulting subnetworks are 'hot' star graphs determined by a single high-scoring node and the immediate neighbors of this node (**Supplementary Fig. 2**). Star graphs or, more generally, spider graphs have one central node connected to multiple neighboring nodes that are not interconnected. Although the hot, center node in these star graphs is typically a significant gene, the neighboring nodes are often artifacts.

We found that HotNet2 returned >80% fewer hot stars/spiders than HotNet on the pan-cancer data sets (**Supplementary Table 31**). This is a major difference between the algorithms and is one of the reasons why HotNet fails to find statistically significant results ($P \leq 0.01$ for any subnetwork size k) on three of six runs (**Supplementary Tables 32** and **33**), whereas HotNet2 found statistically significant results on all six runs. The HotNet2 subnetworks also had a higher fraction of interactions with proteins other than a hot central node (**Supplementary Note**). These differences are explained by the undirected versus directed heat similarity measures used in HotNet versus HotNet2. We note that the goal of HotNet2 is not to eliminate hot stars/spiders but rather to reduce the number of such subnetworks that are false positives.

We also compared HotNet2 to HotNet on simulated data. In short, the results show that HotNet2 achieves higher sensitivity and specificity than HotNet (**Supplementary Fig. 26** and **Supplementary Note**).

To further demonstrate the advantages of HotNet2 on the pan-cancer mutation frequency data set, we compared HotNet2 to HotNet and to two standard tests of pathway enrichment, DAVID[61,62] and gene set enrichment analysis (GSEA)[63,64]. We find that HotNet2 provides both new insights and a simpler summary of groups of interacting genes and is a useful complement (or arguably a replacement for) other pathway tests (**Supplementary Note**). We also show that HotNet2 has much higher specificity than HotNet, DAVID and GSEA in identifying genes satisfying the 20/20 rule[9] (**Supplementary Fig. 27**, **Supplementary Tables 34–36** and **Supplementary Note**). Finally, we found that HotNet2 was more stable than HotNet in identifying 20/20 genes using cross-validation (**Supplementary Fig. 28** and **Supplementary Note**).

We attempted to compare HotNet2 to MEMo[65], an algorithm to identify groups of interacting genes with mutually exclusive mutations. First, we note several important differences between HotNet2 and MEMo. Namely, HotNet2 (i) analyzes mutations and network topology simultaneously; (ii) is not restricted to analyzing mutually exclusive mutations and can analyze co-occurring mutations, and (iii) can use input heat scores that capture additional information (for example, functional relevance) about the mutations. We found that MEMo was unable to run on the pan-cancer mutation frequency data set, consistent with the authors' recommendation that MEMo should be run only on a small number of significant mutations (details provided in the **Supplementary Note**).

Finding consensus subnetworks and linkers. We ran HotNet2 on each combination of gene scores (mutation frequencies and MutSigCV[20] q values; **Supplementary Note**) and interaction networks (HINT+HI2012 (refs. 21,22), iRefIndex[23] and Multinet[24]; **Supplementary Fig. 29** and **Supplementary Note**). We derived 'consensus' subnetworks and 'linker' genes from the HotNet2 results on the different network and gene scores using an iterative procedure on a weighted graph. This procedure is described in the **Supplementary Note**.

We evaluated the statistical significance of the HotNet2 consensus subnetworks using the HotNet2 statistical test on consensus networks found in randomly permuted data. We generated the null distribution of consensus networks by permuting tuples containing the mutation frequency and MutSigCV score of genes over each of the networks. Thus, the permutation preserved the relationship between the mutation frequency and MutSigCV score. We then ran HotNet2 on the three networks using the permuted mutation frequency and MutSigCV score, forming a 'permuted consensus' using the same consensus procedure described above. We used these permuted consensus subnetworks to form an empirical distribution for the statistical test. Additional details of the statistical procedure are provided in the **Supplementary Note**.

Expression and germline filtering. Most of the subnetworks (12/14) identified by HotNet2 were also found when we removed the requirement for RNA-seq expression (**Supplementary Table 37**). This result demonstrates the robustness and scalability of HotNet2, as the unfiltered mutation data included 19,459 genes. Notable among the additional subnetworks identified when we removed the requirement for RNA-seq expression was a subnetwork containing members of the telomerase complex (including *TERT* and *TEP1*) that has a well-studied role in cancer[66] (**Supplementary Fig. 30** and **Supplementary Table 38**). Although the lack of RNA-seq reads from these genes is a concern, we note that the RNA-seq expression criteria were strict enough to exclude several bona fide cancer genes (see URLs). Thus, the lack of RNA-seq reads should not automatically exclude these genes from further study. We also ran HotNet2 using a more aggressive criterion to remove potential germline mutations (see URLs). We found only minor differences in the HotNet2 subnetworks (**Supplementary Table 39**), demonstrating that our reported subnetworks are largely altered by somatic aberrations in these samples.

58. Mitra, K., Carvunis, A.-R., Ramesh, S.K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* **14**, 719–732 (2013).
59. Chung, F. The heat kernel as the pagerank of a graph. *Proc. Natl. Acad. Sci. USA* **104**, 19735–19740 (2007).
60. Berkhin, P. Bookmark-Coloring algorithm for personalized PageRank computing. *Internet Math.* **3**, 41–62 (2006).
61. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* **4**, 44–57 (2009).
62. Huang, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
63. Mootha, V.K. *et al.* PGC-1α–responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
64. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
65. Ciriello, G., Cerami, E.G., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
66. Shay, J.W., Zou, Y., Hiyama, E. & Wright, W.E. Telomerase and cancer. *Hum. Mol. Genet.* **10**, 677–685 (2001).

doi:10.1038/ng.3168