Published in final edited form as:

Nat Genet. 2015 February ; 47(2): 106–114. doi:10.1038/ng.3168.

Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes

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

[1]Department of Computer Science, Brown University, Providence, RI, USA

[2]Center for Computational Molecular Biology, Brown University, Providence, RI, USA

[4]Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, USA

[5]The Genome Institute, Washington University in St. Louis, MO 63108, USA

[6]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02141, USA

[7]Research Unit on Biomedical Informatics, Department of Experimental and Health Sciences, University Pompeu Fabra, Barcelona, Spain

[8]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[9]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[#]Corresponding Author: Benjamin J. Raphael, Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI, USA. braphael@brown.edu, Phone: 401-863-7643.

[3]Current address: Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark.

[*]Equal contribution.

Author Contributions

MDML, FV, HW, BJR designed the HotNet2 algorithm. MDML, FV, HW, JRD, JVE, JLT, YK and BJR performed Pan-Cancer network analysis, analyzed results, and benchmarked algorithms. AP, JRD, YC and GAR analyzed mutation clusters in genes. BN, MM, LD provided MuSiC gene scores and generated mutation validation data. MSL, GG, AG-P, DT, and NL-B provided MutSigCV and Oncodrive gene scores. MDML, FV, HW, JRD, and BJR wrote the manuscript with input from all authors. BJR conceived and supervised the project.

Competing Financial Interests

A patent application related to this work has been filed.

URLs

- HI-2012 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 dataset: 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.

[11]Massachusetts General Hospital, Boston, Massachusetts 02114, USA

[12]Department of Medicine, Washington University in St. Louis, MO 63108, USA

[13]Siteman Cancer Center, Washington University in St. Louis, MO 63108, USA

Abstract

Cancers exhibit extensive mutational heterogeneity and the resulting long tail phenomenon complicates the discovery of the genes and pathways that are significantly mutated in cancer. We perform a Pan-Cancer analysis of mutated networks in 3281 samples from 12 cancer types from The Cancer Genome Atlas (TCGA) using HotNet2, a novel algorithm to find mutated subnetworks that overcomes limitations of existing single gene and pathway/network approaches.. We identify 14 significantly mutated subnetworks that include 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 genes[1–8]. 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[9,10]. This “long tail” phenomenon complicates efforts to identify cancer genes by statistical tests of recurrence, as rarely mutated cancer genes may be indistinguishable from genes containing only passenger mutations. Even recent TCGA Pan-Cancer studies[13–16] have 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/regulatory pathways and protein complexes[9,15]. Clustering of mutations on known pathways is illustrated in many cancer sequencing papers[1,2,5,8], but typically without a measure of statistical significance. While statistical tests of enrichment in known pathways or gene sets exist, such tests do not reveal novel pathways, have limited power to evaluate crosstalk between known pathways, and generally ignore the topology of interactions between genes.

We introduce a novel and complementary approach to identify pathways and protein complexes perturbed by somatic aberrations. This approach combines: (1) a new algorithm, HotNet2, for identification of mutated subnetworks in a genome-scale interaction network; (2) a large TCGA Pan-Cancer dataset of somatic single nucleotide variants, small indels, and copy number aberrations measured in 3,281 samples from 12 cancer types[14]. HotNet2 uses a directed heat diffusion model to *simultaneously* assess both the significance of mutations in individual proteins *and* the local topology of interactions among proteins, overcoming limitations of pathway-based enrichment statistics and earlier network approaches.

Our TCGA Pan-Cancer HotNet2 analysis identifies 14 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 many of the genes in all subnetworks — are rarely mutated in each cancer type, and thus revealed only by the Pan-Cancer network analysis. Many of the rarely mutated genes in the subnetworks have documented physical interactions with well-characterized cancer genes and/or mutational patterns (e.g. clustering in protein sequence/structure or an excess of inactivating mutations) that lend additional support for their role in cancer. Co-occurrence of mutations 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 Pan-Cancer HotNet2 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 dataset of exome sequencing, array copy number, and RNA-seq data from 3,281 samples from 12 cancer types, analyzing single nucleotide variants (SNVs), small indels, and copy number aberrations (CNAs) in 19,424 transcripts (Figure 1a and Supplementary Figure 1). After removing hypermutated samples and genes with low expression in all tumor types (Online Methods), the dataset contained 11,565 mutated genes in 3110 tumors. We observed that the number of samples with a mutation in a gene varied over three orders of magnitude, from 1 to 1291 mutated samples (Figure 1b). Moreover, we discovered that this broad spectrum of mutational frequencies -- from common to extremely rare mutations – posed a challenge for the identification of significantly mutated subnetworks. Specifically, our goal is to identify subnetworks according to *both* the frequency of somatic mutations in individual genes/proteins *and* the topology of the interactions between them. However, the presence of highly mutated and highly connected genes like TP53 presents difficulties for existing algorithms that attempt to achieve this goal; e.g. the HotNet algorithm[16,17] that was used for cancer network analysis in TCGA and other studies[3,4,8,18], or related network propagation approaches[19]. In the heat diffusion model used in HotNet genes like TP53 are extremely “hot” nodes and propagate this heat to their neighboring nodes. The resulting “star subnetworks” centered on the hot node (Supplementary Figure 2; Online Methods) contain many neighboring genes that are not mutated at 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 on large, broad mutation frequency spectrum datasets like Pan-Cancer (Figure 1c and Supplementary Figure 3). HotNet2 uses a modified diffusion process and considers the source, or directionality, of heat flow in the identification of subnetworks (Supplementary Figure 4). This approach reduces the artifact of star subnetworks by more than 80%, reducing the false positive rate and enabling the identification of more subtle subnetworks with rare mutations of high biological relevance (see Online Methods). We compare HotNet2 to other algorithms (Online Methods), and find 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 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 1–2) 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 Table 3), plus the condensin complex and CLASP/CLIP proteins (Supplementary Figure 5) that were significant in individual interaction networks (Supplementary Tables 6,7). Our consensus process also identifies 13 “linker” genes that are members of more than one consensus subnetwork. We developed an online interactive viewer (see URLs and Supplementary Figure 6) for Pan-Cancer HotNet2 subnetworks.

The subnetworks and linker genes (Figure 2a) include: portions of well-known cancer pathways such as TP53, PI3K, NOTCH, and receptor tyrosine kinases (RTKs; Supplementary Figure 7), as well as pathways and complexes that have more recently been observed to be important in cancer such as SWI/SNF complex, BAP1 complex, NFE2L2-KEAP1 (Supplementary Figures 8,9), and RUNX1-CBFB core binding complex (Supplementary Figure 10). The fifth most mutated subnetwork (16.9% of samples) consists of MLL2 and MLL3 and the putative interacting protein KDM6A (Supplementary Figure 11), and was highly mutated (28.9% of samples) in TCGA Pan-Cancer squamous integrated subtype[25]. HotNet2 identified less-characterized and potentially novel subnetworks that may have also important roles in cancer including the cohesin and condensin complexes and MHC Class I proteins. The MHC Class I subnetwork (Supplementary Figure 12) is an example of the ability of HotNet2 ability to identify rarely mutated cancer genes; all of the genes in the subnetwork are mutated in fewer than 35 samples (1.1%), yet four of the five genes have recently been proposed as novel cancer genes[13]. The sections below further detail a subset of these subnetworks. Additional analyses are in the Supplementary Note.

Many of the subnetworks exhibit a significant enrichment for mutations in a subset of cancer types, including many previously unreported associations (Supplementary Tables 6–18). We also identify genes within these subnetworks enriched for mutations in particular cancer types. In addition, the HotNet2 Pan-Cancer analysis provides a clearer and more robust summary of subnetworks and novel genes than HotNet2 analysis of individual cancer types (Supplementary Table 19).

These subnetworks and linkers include a total of 147 genes, including many well-known cancer genes and pathways, but also including genes with mutations that are too rare to be significant by the single-gene tests (Supplementary Table 20). In total, 92 genes in the HotNet2 subnetworks are *not* reported by any of five single-gene tests (MutSigCV[20], Oncodrive-FM[26] and –CIS[27], MuSiC[28], or GISTIC2[29]) or listed as a known driver gene in Vogelstein *et al.*[9], while an additional 13 genes are reported in only one such list. Many of these genes have literature evidence supporting a potential role in cancer, while others are in biological processes that suggest these genes warrant further study. Table 1 lists a subset of promising candidates, with the full list and associated references 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 protein sequence (using NMC[30]) or in protein structure (using iPAC[31]; Supplementary Figures 13,14 and Supplementary Tables 21,22). We find that genes in HotNet2 consensus subnetworks are enriched for inactivating mutations ($P < 0.0001$) or mutation clusters ($P < 0.0001$) compared to genes not in subnetworks (Supplementary Table 6–18 and Supplementary Note Section 5.1). Finally, we evaluated a subset of the mutations in these genes using RNA-Seq and whole-genome sequencing (WGS) data from the same samples, and found RNA-Seq and/or WGS reads that validated 39 mutations in these novel genes (Supplementary Note Section 6 and Supplementary Table 23). These genes may represent novel 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 find that 4 pairs of subnetworks, including TP53 and NOTCH signaling, TP53 and RTK signaling, PI3K signaling and cohesin complex, and PI3K and ASCOM complex exhibit significant co-occurrence ($P < 0.05$, multiple hypotheses corrected) across the Pan-Cancer cohort (Figure 2b) or in individual cancer types (Figure 2c). Multiple pairs of genes within these subnetworks show co-occurring mutations (Supplementary Table 24). In contrast, mutual exclusive mutations are typically expected within a pathway, and not across pathways[32,33]. We observe significant mutual exclusivity within 4 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 exclusivity was largely due to LUAD samples with mutually exclusive *EGFR* and *KEAP1* mutations (Supplementary Figure 15). This observation is consistent with reports of exclusivity between *EGFR* mutations and *NFE2L2* expression in LUAD[34] and also that *NFE2L2* expression is downstream of *EGFR* signaling[35]. Examining individual cancers, we find a modest but not statistically significant enrichment for co-occurrence or exclusivity in a few cancer types. 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 – contain many well-known cancer genes (Supplementary Tables 8–10 and Supplementary Figures 16,17). Linker genes join these three subnetworks, demonstrating the extensive crosstalk between well-annotated cancer pathways. Most of these linker genes encode signaling proteins that have known cancer-related functions (e.g. *WT1*, *NOTCH2*, *PIK3R1*, *MAP2K4*, *MAP3K1*, *HRAS*, *ATM*, and *STK11*). Taken together, 81.9% of the samples contain at least one mutation in these three large subnetworks and linker genes.

HotNet2 Pan-Cancer analyses also revealed a number of novel genes (Supplementary Table 20) within these three subnetworks. These genes have documented interactions with well-known cancer 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, includes *CUL9*. *CUL9* sequesters p53 in the cytoplasm, and we find a cluster of 45 missense mutations ($P = 1.32 \times 10^{-8}$) as well as a cluster in protein structure (FDR = 0.025). Another gene of interest is *IWS1*, which is involved in transcriptional elongation and mRNA surveillance. Half (8/16) of the mutations in this gene are inactivating, and it also has a cluster of mutations ($P = 0.013$). This subnetwork also contains *CHD8*, an ATP-dependent chromatin-remodeling factor that regulates a wide range of genes[36]. We find three independent signals of *CHD8* inactivation across samples: *CHD8* is deleted in 9 samples in a focal peak from GISTIC; 18/58 (31%) of its mutations as inactivating; and has a wide cluster of missense mutations ($P = 6.37 \times 10^{-5}$). In the *NOTCH* subnetwork, we find rare mutations in *JAG1* and *DLL1*, which interact with the NOTCH receptors and have some reports of a role in cancer[37]. Moreover, 11/24 mutations in *JAG1* are inactivating. The *NOTCH* subnetwork also includes *SHPRH*, which has a significant ($P < 8\times10^{-5}$) cluster of missense mutations (Supplementary Figure 18).

SWI/SNF complex

The sixth most mutated HotNet2 Pan-Cancer subnetwork (16.8% of samples) includes multiple members of the SWI/SNF chromatin-remodeling complex (Figure 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 demonstrates the prevalence of mutations in SWI/SNF: at least 1.5% of the samples from each of the 12 cancer types contain a mutation in this subnetwork. KIRC ($P<10^{-15}$), UCEC ($P=7\times10^{-10}$), and BLCA ($P=1.8\times10^{-8}$) were enriched for mutations in this subnetwork and several genes were enriched for mutations in specific cancer types including *PBRM1* in KIRC ($P<10^{-15}$) and *ARID1A* in both BLCA ($P=4.8\times10^{-8}$) and UCEC ($P<10^{-15}$). The subnetwork also contains *ARID1B*, which is reported to have somatic mutations in juvenile neuroblastoma[41] and germline mutations in Coffin-Siris syndrome[42].

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

mutations contribute to tumorigenesis. Thus, HotNet2 analyses broaden the view of mutations in SWI/SNF to additional cancer types and additional interacting proteins.

BAP1 Complex and Interactors

Another HotNet2 Pan-Cancer subnetwork (mutated in 7.1% of samples) overlaps the BAP1 complex (Figure 3b and Supplementary Table 13). This subnetwork includes *BAP1*, *ASXL1*, *ASXL2*, *FOXK1*, *FOXK2*, all members of the BAP1 core complex[45], as well as two additional interacting proteins: *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 revealed by HotNet2 Pan-Cancer analysis. This subnetwork is mutated in at least 6 samples from each cancer type, demonstrating the breadth of mutations in the BAP1 complex.

BAP1 inactivation has been reported in several cancers[45]. We find the subnetwork enriched for mutations in KIRC ($P=2\times10^{-4}$), as previously reported[46]. Consistent with Peña-Llopis *et al.*[46], we find that mutations in the *BAP1* gene are mutually exclusive ($P<7.2\times10^{-3}$) of mutations in the *PBRM1* gene in KIRC. We find that mutations in the SWI/SNF and BAP1 complexes show even greater mutual exclusivity ($P=9.4\times10^{-5}$) in KIRC because of mutations in additional genes in these complexes besides *BAP1* and *PBRM1*, respectively (Supplementary Note Section 5.8.1). This mutual exclusivity suggests that mutations in these complexes define different subtypes of kidney cancer. Supporting this hypothesis, we observe that inactivating mutations in the BAP1 complex are enriched ($P<3.4\times10^{-8}$) for samples in the third mRNA expression subtype from[3] (Figure 3c).

We find that a large fraction of the mutations in *BAP1*, *ASXL1*, and *ASXL2* in different cancer types are inactivating mutations, demonstrating alternative strategies for inactivation of the BAP1 complex. In addition, 6/13 missense mutations in *FOXK2* are in the forkhead transcription factor domain or forkhead associated domain, which may inactivate the DNA-binding properties of FOXK2. Finally, we examined the mutations in *KDM1B*, a gene that is involved in H3K4-methylation[47], but not considered a core part of the BAP1 complex. We find that 12/19 mutations in *KDM1B* (including 10/16 missense mutations) fall in the C-terminal amino-oxidase domain that is important for lysine-specific demethylation of histones[48]. Moreover, 2 of the 3 *KDM1B* mutations in LUSC and LUAD are inactivating, and these are also exclusive of *BAP1* inactivating mutations, suggesting that *KDM1B* mutations might play a role in cancer.

Cohesin and condensin

HotNet2 Pan-Cancer analysis identifies 4/5 members of the cohesin complex as a significantly mutated subnetwork (7.3% of samples, Figure 4a and Supplementary Table 15). While 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 cohesin 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 of mutations complicates the identification of recurrent mutations in

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

Mutations in some of these genes have recently been reported to be significant in several cancers. We find enrichment for mutations in the subnetwork in BLCA (P=7×10^{-4}); this enrichment derives largely from enrichment for mutations in *STAG2* in BLCA (P=0.005), which was recently reported[53]. *STAG2* has a significantly higher fraction of inactivating mutations than other genes in the subnetwork (53% for STAG2 compared to 28% for the subnetwork as a whole); these inactivating mutations are not only in BLCA, but also across multiple cancer types with multiple inactivating mutations in LAML and COADREAD. In addition, BLCA samples without STAG2 inactivating mutations harbor rare inactivating mutations in several other cohesin genes. All mutations in *RAD21* in LAML samples were inactivating, and BRCA and KIRC harbor inactivating mutations in *STAG1*. In addition, we observed a significant clustering of missense mutations in *STAG1* (P=6×10^{-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 thus mutation of *STAG1* may play another role in cancer apart from 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 identifies two subnetworks containing six proteins in the condensin complex, in HotNet2 runs from individual interaction networks. The combined subnetwork is mutated in 4.2% of samples (Figure 4b and Supplementary Table 6). Only *SMC4* was reported significant by at least one of the individual gene scores. A subnetwork consisting of *NCAPD2*, *SMC2*, and *SMC4*, both members of Condensin I form of the complex, was significantly mutated in BLCA (P= 6.2 × 10^{-6}). Condensin I is thought to primarily be involved in the 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 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 are 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* (Figure 4b), implying that mutations in this region of unknown function may be important for the deregulation of condensin. We also note that it was recently observed that expression of *NCAPD3* was positively associated with recurrence-free survival[56]. Finally, RNA-seq and whole-genome sequencing data from the same samples provide further validation of the somatic mutations in *SMC2*, *SMC4*, *NCAPD2*, *NCAPD3*, *NCAPH2*, and *NCAPG2* and show that some of these mutations are expressed (Supplementary Note Section 6 and Supplementary Table 39). 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 novel approach for identifying combinations of somatic aberrations in different cancer types using our HotNet2 algorithm to analyze a high-quality Pan-Cancer dataset of 3281 samples from 12 cancer types. This analysis represents the largest network analysis of somatic aberrations across multiple cancer types. We recover many classic cancer pathways like TP53, PI3K, NOTCH, and RTK automatically from a large-scale interaction network, demonstrating the power of the Pan-Cancer network approach. Second, we highlight the extensive crosstalk 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 was only appreciated recently such as the SWI/SNF chromatin-remodeling complex[38] and BAP1 complex[45]. Fourth, we find that several pairs of HotNet2 subnetworks have co-occurring mutations, while within subnetworks mutations are mostly exclusive. This supports the hypothesis that these subnetworks represent distinct biological functions that are mutated in samples. Finally, we identify a number of novel 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 revealed 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 (Supplementary Tables 6–18). In addition, we find that well-characterized mutations in a single gene in one cancer type (e.g. inactivating mutations *BAP1* in KIRC) are replaced in other cancer types by rare mutations in other members of the same complex (e.g. inactivating mutations in *ASXL1*, *ASXL2*, *FOXK2*, *KDM1B*). Such observations suggest that Pan-Cancer network analyses may 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 analyze only one type of aberration[11–13] or restrict attention to recurrent aberrations[57] (Supplementary Note Section 8.3 and Supplementary Table 27). The HotNet2 Pan-Cancer network approach identifies combinations of rare and common mutations in groups of interacting genes; combinations that were not apparent by 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 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 input data. Further power is anticipated by including additional samples[13], additional types of genetic and epigenetic aberrations, and better interaction networks. For example, structural variants, non-coding variants and methylation data were not included, the first two being unavailable for most TCGA samples. This lack of data, plus false negatives in the analyzed data (e.g. due to difficulties in identification of indels and subclonal variants) imply that our analysis likely underestimates the number and frequency of mutated subnetworks across cancer types. On the other hand, we note that some genes that are highly significant by individual gene scores are not reported in our network analysis; often this is due 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 – are 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 of identification of combinations of genetic variants with a statistically significant association to a phenotype. With an appropriate gene score, the HotNet2 algorithm can be applied to such data.

Online Methods

Somatic aberration data

SNVs, indels, and splice-site mutations were extracted from TCGA Pan-Cancer analysis on Synapse (syn1710680), and copy number aberrations (CNAs) from GISTIC2 output via Firehose. We restricted attention to the 3276 samples containing both SNV and CNA data. We removed 71 samples identified as ultramutators in syn1729383 and 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 Figure 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 3 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 dataset contained aberrations in 11,565 genes and 3110 samples (Supplementary Figure 1). We used genes scores from: mutation frequency and MutSigCV $-\log_{10} q$-values. Nonsense, frame shift indels, nonstop, or splice site mutations were classified as inactivating following[11]. We used three interaction networks: HINT+HI2012, a combination of HINT network[21] and the HI-2012[22] set of protein-protein interactions; MultiNet[23]; iRefIndex[24]. Additional details of the datasets are in the Supplementary Note.

HotNet2

We developed the HotNet2 (HotNet diffusion oriented subnetworks) algorithm to identify subnetworks of a genome-scale interaction network that are mutated more than expected by chance. While 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 (Supplemental Note Section 1.1). 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/proteins as well as the topology of interactions between genes/proteins (Supplementary Figure 3).

The input to HotNet2 is: a heat vector $\vec{h}$ that contains the scores (e.g., mutation frequency) for each gene g; and a graph $G = (V, E)$, where each node corresponds to a gene/protein and each edge corresponds to an interaction between the corresponding genes/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; the amount of heat on each node at equilibrium thus depends on its initial heat, the local topology of the network around the node, and the value β. If a unit heat source is placed at node j (e.g. 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:

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

where

$$W_{ij}=\begin{cases}\frac{1}{\deg(j)}, & \text{if node } i \text{ interacts with node } j\\ 0, & \text{otherwise.}\end{cases}$$

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 described in terms of a random walk with restart (Supplementary Note Section 1.2). 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 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 unit heat is placed on nodes. To jointly analyze network topology and gene scores given by the initial heat vector $\vec{h}$, we define the exchanged heat matrix E:

$$E=F\,D_{\vec{h}},$$

where D_h is the diagonal matrix with entries h. $\vec{E}(i, j) = F(i, j)h(\vec{j})$ is the amount of heat that diffuses from node g_j to node g_i on the network when $h(\vec{j})$ heat is placed on g_j, which we interpret as the similarity of g_j, g_i. Since the diffusion matrix F is not symmetric and in general $h(\vec{i}) \neq h(\vec{j})$, the similarity $E(i, j)$ is also not symmetric (Supplementary Note Section 1.2.1).

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 Note Section 1.3, Supplementary Figures 20–23 and Supplementary Table 28).

HotNet2 is available online (See URLs).

HotNet2 has two parameters β and δ, and 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 Note Section 1.4.1, Supplementary Figure 24, and Supplementary Table 29). We evaluated the sensitivity of the HotNet2 results to the value of β and found that varying β ±10% has only a minor effect on the results, with at most 7 genes (3.8% of total) added/removed from the subnetworks (Supplementary Table 28). 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 Note Section 1.4.2, Supplementary Figure 25, and Supplementary Table 30). We evaluated the sensitivity of the HotNet2 results to the value of δ, and found that varying δ ±5% changed at most 35 genes (12.3% of total) in the subnetworks (Supplementary Table 29).

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, g_j to derive a directed measure of similarity $E(i, j)$ between them, while HotNet derives a symmetric influence. Third, HotNet2 identifies strongly connected components in the directed graph H, while HotNet computes connected components in an undirected graph. These differences enable HotNet2 to effectively detect significant subnetworks in datasets 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 Figure 2.

Expanding on this third point, when undirected diffusion algorithms like HotNet or related network propagation algorithms[19] are run on large datasets containing a wide range of gene scores (e.g. the Pan-Cancer dataset), many of the resulting subnetworks are "hot" *star graphs* determined by a single high-scoring node and the immediate neighbors of this node (Supplementary Figure 2). Star graphs, or more generally spider graphs, have one central node connected to multiple neighboring nodes that are not interconnected. While the hot, center node in these star graphs is typically a significant gene, the neighboring nodes are often artifacts.

We found that HotNet2 returns >80% fewer hot stars/spiders than HotNet on the Pan-Cancer datasets (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 Table 32,33), while HotNet2 finds statistically significant results on all six runs. The HotNet2 subnetworks also have a higher fraction of interactions with proteins other than a hot central node (Supplementary Note Section 7.1). These differences are explained by the undirected vs. 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 Note Section 7.2 and Supplementary Figure 26).

To further demonstrate the advantages of HotNet2 on the Pan-Cancer mutation frequency dataset, 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 Section 8.1). We also show that HotNet2 has much higher specificity than HotNet, DAVID, and GSEA in identifying genes satisfying the 20/20 rule[9] (Supplementary Note Section 8.1.4, Supplementary Figure 27, and Supplementary Tables 34–36). Finally, we find that HotNet2 was more stable than HotNet in identifying 20/20 genes using cross-validation (Supplementary Note Section 7.3 and Supplementary Figure 28).

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 difference between HotNet2 and MEMo. Namely, HotNet2 (1) analyzes the mutations and network topology *simultaneously*; (2) is not restricted to analyzing exclusive mutations and can analyze co-occurring mutations, and (3) can use input heat scores that capture additional information (e.g. functional significance) about the mutations. We found that MEMo was unable to run on the Pan-Cancer mutation frequency dataset, consistent with the authors' recommendation that MEMo should be run only on a small number of significant mutations (details in Supplementary Note Section 8.2).

Finding consensus subnetworks and linkers

We ran HotNet2 on each combination of gene scores (mutation frequency and MutSigCV[20] q-values; see Supplementary Note Section 2.2) and interaction networks (HINT

+HI2012[21,22], iRefIndex[23], and Multinet[24]; Supplementary Note Section 2.5 and Supplementary Figure 29). 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 Supplementary Note Section 1.5.

We evaluated the statistical significance of the HotNet2 consensus subnetworks using the HotNet2 statistical test on consensus networks found in randomly permuted data. We generate the null distribution of consensus networks by permuting tuples containing the mutation frequency and MutSigCV scores of genes over each of the networks. Thus, the permutation preserves the relationship between the mutation frequency and MutSigCV score. We then ran HotNet2 on the three networks using the permuted mutation frequency and MutSigCV scores 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 in Supplementary Note Section 1.3.

Expression and Germline Filtering

Most of the subnetworks (12/14) identified by HotNet2 were also found when we remove the requirement for RNA-Seq expression (Supplementary Table 37). This result demonstrates the robustness and scalability of the HotNet2, as the unfiltered mutation data includes 19,459 genes. Notable among the additional subnetworks identified when we remove the requirement for RNA-Seq expression is a subnetwork (Supplementary Table 25) containing members of the telomerase complex (including *TERT* and *TEP1*) that has a well-studied role in cancer[66] (Supplementary Figure 30 and Supplementary Table 38). While the lack of RNA-Seq reads from these genes is a concern, we note that the RNA-Seq expression criteria was 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 altered by somatic aberrations in these samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Fritz Roth for his assistance in constructing HINT+HI2012 interaction network. We gratefully acknowledge the contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group. This work is supported by NSF grant IIS-1016648 and NIH grants R01HG005690, 1R01HG007069, and 1R01CA180776 to BJR, and National Human Genome Research Institute grant U01HG006517 to LD. BJR 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). MDML is supported by NSF GRFP DGE 0228243. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. HI-2012 data created by The Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute is supported by The National Human Genome Research Institute (NHGRI) of NIH, The Ellison Foundation, Boston, MA, and The Dana-Farber Cancer Institute Strategic Initiative.

References

1. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]
2. The Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–8. [PubMed: 18772890]
3. Creighton CJ, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 201310.1038/nature12222
4. The Cancer Genome Atlas Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med. 201310.1056/NEJMoa1301689
5. Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–25. [PubMed: 22960745]
6. Kandoth C, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013; 497:67–73. [PubMed: 23636398]
7. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–24. [PubMed: 19360079]
8. The Cancer Genome Atlas Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–15. [PubMed: 21720365]
9. Vogelstein B, et al. Cancer genome landscapes. Science. 2013; 339:1546–58. [PubMed: 23539594]
10. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013; 153:17–37. [PubMed: 23540688]
11. Kandoth C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502:333–339. [PubMed: 24132290]
12. Zack TI, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013; 45:1134–1140. [PubMed: 24071852]
13. Lawrence MS, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501. [PubMed: 24390350]
14. Weinstein JN, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–20. [PubMed: 24071849]
15. Hanahan D, Weinberg Ra. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–74. [PubMed: 21376230]
16. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011; 18:507–22. [PubMed: 21385051]
17. Vandin F, Clay P, Upfal E, Raphael BJ. Discovery of Mutated Subnetworks Associated with Clinical Data in Cancer. Pacific Symp Biocomput. 2012:55–66.
18. Grasso CS, et al. The mutational landscape of lethal castration-resistant prostate cancer. Nature. 2012; 487:239–243. [PubMed: 22722839]
19. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 201310.1038/nmeth.2651
20. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–218. [PubMed: 23770567]
21. Das J, Yu H. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol. 2012; 6:92. [PubMed: 22846459]
22. Yu H, Tardivo L, Tam S, Weiner E, Gebreab F, et al. Next-generation sequencing to generate interactome datasets. Nature methods. 2011; 8:478–480. [PubMed: 21516116]
23. Khurana E, Fu Y, Chen J, Gerstein M. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol. 2013; 9:e1002886. [PubMed: 23505346]
24. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008; 9:405. [PubMed: 18823568]
25. Hoadley KA, et al. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. Cell. 2014; 158:929–944. [PubMed: 25109877]
26. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012; 40:e169. [PubMed: 22904074]

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. 2013; 8:e55489. [PubMed: 23408991]
28. Dees ND, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012; 22:1589–98. [PubMed: 22759861]
29. Mermel CH, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011; 12:R41. [PubMed: 21527027]
30. Ye J, Pavlicek A, Lunney EA, Rejto PA, Teng CH. Statistical method on nonrandom clustering with application to somatic mutations in cancer. BMC Bioinformatics. 2010; 11:11. [PubMed: 20053295]
31. Ryslik GA, Cheng Y, Cheung KH, Modis Y, Zhao H. Utilizing protein structure to identify non-random somatic mutations. BMC Bioinformatics. 2013; 14:190. [PubMed: 23758891]
32. Yeang CH, McCormick F, Levine A. Combinatorial patterns of somatic gene mutations in cancer. FASEB J. 2008; 22:2605–22. [PubMed: 18434431]
33. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2011; 22:375–85. [PubMed: 21653252]
34. Solis LM, et al. Nrf2 and Keap1 abnormalities in non-small cell lung carcinoma and association with clinicopathologic features. Clin Cancer Res. 2010; 16:3743–53. [PubMed: 20534738]
35. Yamadori T, et al. Molecular mechanisms for the regulation of Nrf2-mediated cell proliferation in non-small-cell lung cancers. Oncogene. 2012; 31:4768–77. [PubMed: 22249257]
36. Thompson, Ba; Tremblay, V.; Lin, G.; Bochar, Da. CHD8 is an ATP-dependent chromatin remodeling factor that regulates beta-catenin target genes. Mol Cell Biol. 2008; 28:3894–904. [PubMed: 18378692]
37. Greife A, et al. Canonical Notch signalling is inactive in urothelial carcinoma. BMC Cancer. 2014; 14:628. [PubMed: 25167871]
38. Wilson BG, Roberts CWM. SWI/SNF nucleosome remodellers and cancer. Nat Rev Cancer. 2011; 11:481–92. [PubMed: 21654818]
39. Varela I, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature. 2011; 469:539–42. [PubMed: 21248752]
40. Kadoch C, et al. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat Genet. 2013; 45:592–602. [PubMed: 23644491]
41. Sausen M, et al. Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. Nat Genet. 2013; 45:12–7. [PubMed: 23202128]
42. Tsurusaki Y, et al. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat Genet. 2012; 44:376–8. [PubMed: 22426308]
43. Mandel S, Gozes I. Activity-dependent neuroprotective protein constitutes a novel element in the SWI/SNF chromatin remodeling complex. J Biol Chem. 2007; 282:34448–56. [PubMed: 17878164]
44. Steingart RA, Gozes I. Recombinant activity-dependent neuroprotective protein protects cells against oxidative stress. Mol Cell Endocrinol. 2006; 252:148–53. [PubMed: 16704895]
45. Carbone M, et al. BAP1 and cancer. Nat Rev Cancer. 2013; 13:153–9. [PubMed: 23550303]
46. Peña-Llopis S, et al. BAP1 loss defines a new class of renal cell carcinoma. Nat Genet. 2012; 44:751–9. [PubMed: 22683710]
47. Fang R, et al. Human LSD2/KDM1b/AOF1 regulates gene transcription by modulating intragenic H3K4me2 methylation. Mol Cell. 2010; 39:222–33. [PubMed: 20670891]
48. Shi Y, et al. Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell. 2004; 119:941–53. [PubMed: 15620353]
49. Xu H, Tomaszewski JM, McKay MJ. Can corruption of chromosome cohesion create a conduit to cancer? Nat Rev Cancer. 2011; 11:199–210. [PubMed: 21326324]
50. Rubio ED, et al. CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A. 2008; 105:8309–14. [PubMed: 18550811]

51. Schmidt D, et al. A CTCF-independent role for cohesin in tissue-specific transcription. Genome Res. 2010; 20:578–88. [PubMed: 20219941]
52. Kon A, et al. Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. Nat Genet. 2013; 45:1232–7. [PubMed: 23955599]
53. Solomon, Da, et al. Frequent truncating mutations of STAG2 in bladder cancer. Nat Genet. 2013; 45:1428–30. [PubMed: 24121789]
54. Wood AJ, Severson AF, Meyer BJ. Condensin and cohesin complexity: the expanding repertoire of functions. Nat Rev Genet. 2010; 11:391–404. [PubMed: 20442714]
55. Hirano T. Condensins: universal organizers of chromosomes with diverse functions. Genes Dev. 2012; 26:1659–78. [PubMed: 22855829]
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. 2008; 32:205–9. [PubMed: 18223322]
57. Ciriello G, et al. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013; 45:1127–1133. [PubMed: 24071851]
58. Mitra K, Carvunis A-R, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013; 14(17):719–32. [PubMed: 24045689]
59. Chung F. The heat kernel as the pagerank of a graph. Proc Natl Acad Sci. 2007; 104:19735–19740.
60. Berkhin P. Bookmark-Coloring Algorithm for Personalized PageRank Computing. Internet Math. 2006; 3:41–62.
61. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009; 4:44–57. [PubMed: 19131956]
62. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37:1–13. [PubMed: 19033363]
63. Mootha VK, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34:267–73. [PubMed: 12808457]
64. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–50. [PubMed: 16199517]
65. Ciriello G, Cerami EG, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2011; 22:3980496.
66. Shay JW, Zou Y, Hiyama E, Wright WE. Telomerase and cancer. Hum Mol Genet. 2001; 10:677–686. [PubMed: 11257099]

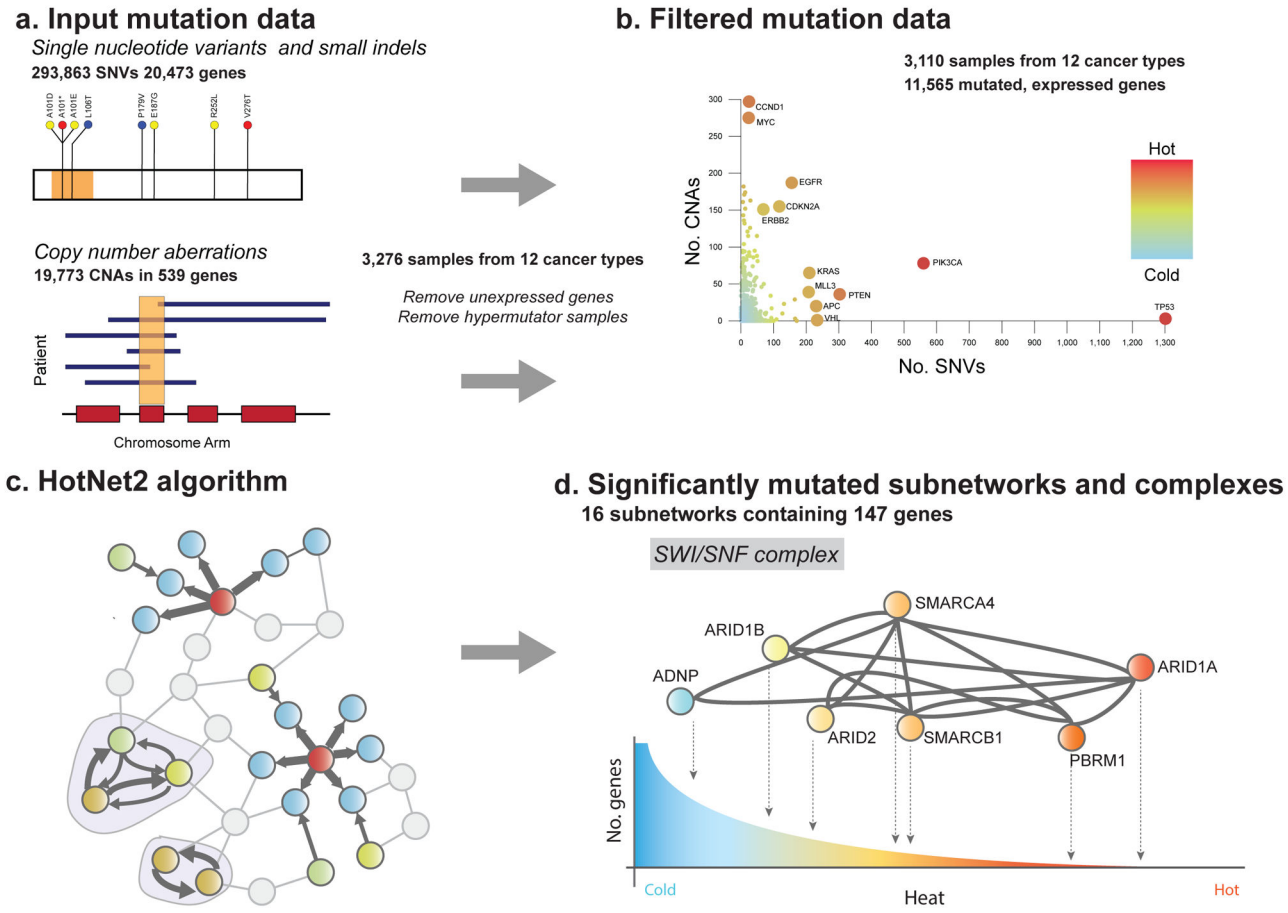

Figure 1.

HotNet2 Pan-Cancer analysis (a) The Pan-Cancer mutation data combines SNVs (nsSNVs and small indels) and CNAs (amplifications and deletions) in 19,459 genes in 3,281 samples. The number of samples with SNVs/CNAs is shown for each gene, with points colored by the total. (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 CNA (y-axis) in these genes. (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/significance of SNVs or CNAs in the corresponding gene. Heat diffuses across edges of 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 wide range of heat scores, including both frequently mutated, known cancer genes (hot genes) and rarely mutated genes (cold genes) that are implicated due to their interactions with other cancer types. Thus, HotNet2 delves into long tail of rarely mutated genes by analysis of combinations of interacting genes.

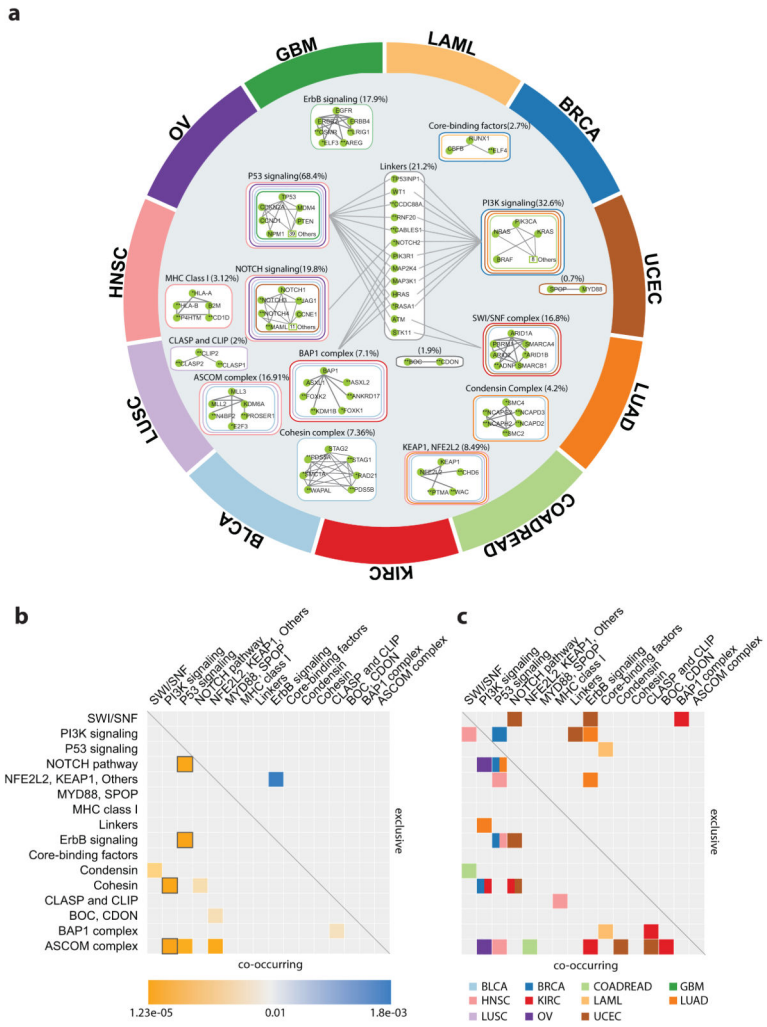

Figure 2.

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 “linker” genes that are members of more than one consensus subnetwork; dotted lines between linkers and other consensus subnetworks indicate protein-protein interactions between them. (b) Heat map of significant co-occurrence (yellow, lower triangular) and exclusivity (blue, upper triangular) 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) Exclusivity/co-occurrence ($P < 0.01$, Fisher's exact test) within individual cancer types using the same color scheme as part (a).

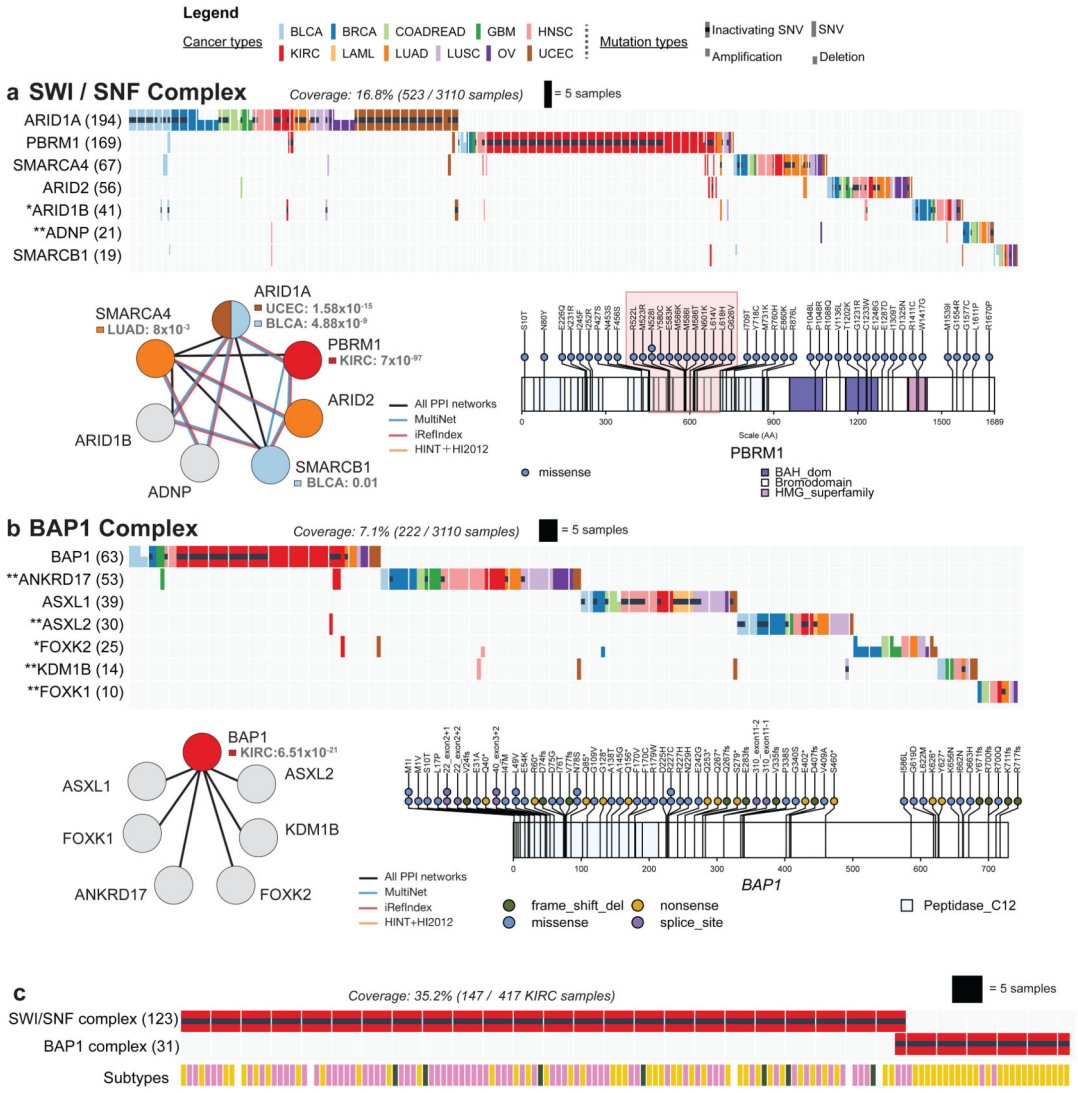

Figure 3.

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, PBAF proteins PBRM1 and ARID2, catalytic core member SMARCA4, SMARCB1 and ADNP. (a - Top) Mutation matrix shows the samples (colored by cancer type as shown in legend) 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, that the genes contain at least one of the following mutations: nonsense, frame shift indels, nonstop, or splice site. The number of samples with mutations in a gene is in parenthesis; genes with * were significant by exactly one of GISTIC2, MuSiC, MutSigCV, Oncodrive, or the list of driver genes in[9] while genes with ** were not significant by any of these methods. (a - Bottom left) Interactions between proteins in the subnetwork from each interaction network are colored according to mutually enriched cancer type with corresponding *P*-values. (a - Bottom right) PBRM1 protein sequence exhibited significant

clustering of missense mutations (P=1.6×10^{-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 TCGA KIRC publication[3]. (B) Subnetwork containing members of the BAP1 complex including core PR-DUB complex, comprised of the deubiquinating 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 structured as in panel (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 exclusive in KIRC (P<3.6×10^{-4}, Fisher's exact test, corrected), and BAP1 is significantly enriched in mutations in the third expression subtype (P<3.4×10^{-8}, Fisher's exact test).

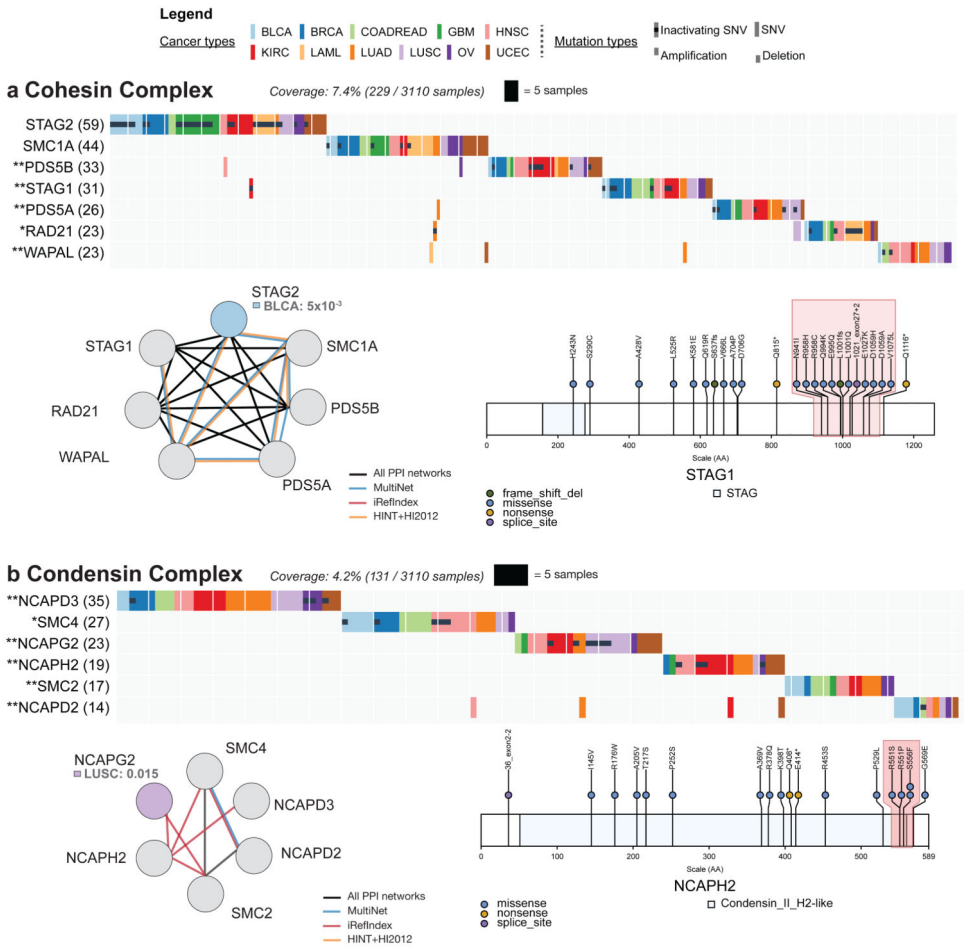

Figure 4.

HotNet2 Pan-Cancer subnetworks overlapping the cohesin and condensin complexes. (a) Cohesin consensus subnetwork and its mutations. Colors and marks as in Figure 2(a). None of the genes is 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 ($P<6\times10^{-5}$) clustering of missense mutations across 135 residues (highlighted) in the Pfam-B domain (PFAM ID: PB002581), a pattern suggesting inactivation of the corresponding domain. (b) Condensin consensus subnetwork, its mutations. (Top) Mutation matrix shows five genes in the condensin I and 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, while 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 ($P < 2.6\times10^{-4}$) cluster of missense mutations between R551 and S556.

Table 1

A subset of candidate cancer genes identified by HotNet2, but not by single-gene tests of significance (non-italicized genes are listed as a cancer driver by Oncodrive or GISTIC). For each gene, the number of samples with at least one SNV/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 in Supplementary Table 20.

GENE	SNVs	CNAS	CANCER ENRICHMENT(S)	FUNCTION
ADNP	21	0		Homeobox transcription factor with 9 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; plays an opposing role to ASXL1.
CCDC88A	38	0		Girdin family member with a key role in PI(3)K and Akt signaling pathways that may be involved in metastasis when overexpressed
CHD8	49	9		DNA helicase that acts as a chromatin remodeling factor and suppresses transcription. Suppresses *TP53* and negatively regulates β-catenin in WNT signaling. CHD8 is essential for embryonic development.
CUL9	48	0		Involved with p53 localization; critical regulator of cell cycle and quiescence.
ELF3	19	0	BLCA, COADREAD	Transcriptional activator that binds ETS motifs. May be a downstream effector of the ERBB2 signaling pathway.
EPHA3	50	3		Receptor tyrosine kinase with possible roles in BRCA, COADREAD, GBM, HNSC, lung, and pancreatic cancer.
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. May play a role in AML, BRCA, COADREAD, GBM, OV, and pancreatic cancer.
KDM1B	14	0		Histone demethylase that acts as a co-repressor; 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 MLL5 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 Jagged1/2 and Delta 1 to regulate cell fate through transcriptional activation; mutations in *NOTCH3* 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 chromatin adhesion following DNA replication
WAC	19	0		Regulates cell cycle progression by linking transcription to H2BK120ub1