From Parts to Mechanisms: Research Heuristics for Addressing Heterogeneity in Cancer Genetics

William Bechtel
Department of Philosophy
University of California, San Diego

Abstract

A major approach to cancer research in the late 20th century was to search for genes that, when altered, initiated the development of a cell into a cancerous state (oncogenes) or failed to stop this development (tumor suppressor genes). But as researchers acquired the capacity to sequence tumors and incorporated the resulting data into databases, it became apparent that for many tumors no genes were frequently altered and that the genes altered in different tumors in the same tissue type were often distinct. To address this heterogeneity problem, many researchers looked to a higher level of organization—to mechanisms in which gene products (proteins) participated. They proposed to reduce heterogeneity by recognizing that multiple gene alterations affect the same mechanism and that it is the altered mechanism that is responsible for the cell developing one or more hallmarks of cancer. I examine how mechanisms figure in this research and focus on two heuristics researchers use to integrate proteins into mechanisms, one focusing on pathways and one focusing on clusters in networks.

1. Introduction

In many contexts in both science and medicine, research advances as investigators target lower levels of organization. Philosophical accounts of mechanistic explanation emphasize strategies for decomposing a mechanism into its component parts and determining how these components contribute to the phenomenon associated with the mechanism (Craver & Darden, 2001; Bechtel & Abrahamsen, 2005). But there are situations in which research advances by moving from the level of the components of mechanisms to the mechanisms themselves. I focus on one situation. Many cancer researchers in the 1980s and 1990s had hoped that cancer could be explained in terms of altered genes, but by the early 2000s they confronted the heterogeneity of proteins identified as altered in cancer cells. This has led some researchers to focus not on the genes and proteins themselves but on the mechanisms in which the proteins function in cells. In the process, they have developed a new set of research heuristics that integrate proteins into mechanisms so that they can then appeal to the mechanisms as wholes to explain cancers.

As with research on many biological phenomena, research on diseases often begins by seeking a single causal factor relevant to the generation of the disease. In the case of cancer, this has often taken the form of identifying genes that when altered[1] contribute to the transformation of a healthy cell into a tumor cell. The identification of the first oncogenes and tumor suppressor

[1] I will speak mostly of alterations, not mutations, since many studies consider other forms of genetic change, such as altered copy number or chromosomal inversions. I will use *mutations* when the study is focused specifically on mutations.

genes in the 1980s motivated further pursuit of this endeavor. With the development of new tools for gene sequencing in the 1990s, investigators started to sequence tumors to identify the genes frequently altered in them. This, however, revealed a serious problem of heterogeneity: no gene was altered in all samples of what were assumed to be the same type of cancer and only a few genes were altered in a significant portion of tumors of that type. I discuss the discovery of heterogeneity and the challenges it has presented for identifying causes of the transformation of a cell into cancer in section 2.

One response to this heterogeneity is to reject the idea that genes are the relevant causal agents in cancer and appeal, for example, to the tissues in which cancer cells reside (Soto & Sonnenschein, 2011).[2] Many cancer geneticists, however, have adopted a different strategy, one that also involves moving to a higher level of organization, but one still within the cell. They target cellular mechanisms as the relevant units and seek to explain the transition of a cell into a cancerous state in terms of the altered functioning of these mechanisms. A mechanism can be disrupted by any number of altered genes that code for the mechanism's constitutive proteins. The guiding idea is that one can overcome the heterogeneity problem by identifying and focusing on the disrupted mechanisms that are responsible for the various hallmarks of cancer.

Philosophers interested in discovery often focus on heuristic strategies: fallible reasoning strategies that reduce the search space of potential explanations (Newell & Simon, 1972; Wimsatt, 2007; Bechtel & Richardson, 1993/2010; Darden, 2006). In this paper I focus on new heuristic strategies that are employed in advancing from components to the mechanisms in which they participate so as to invoke those in explanations. In section 3 I differentiate two such approaches—integrating components into pathways and identifying clusters through network analyses. In subsequently sections I present examples of cancer research that employ these strategies to identify mechanisms through which altered genes contribute to cancer. These strategies are heuristic in the same sense as decomposition and localization, discussed by Bechtel and Richardson (1993/2010), and forward and backward chaining, discussed by Craver and Darden (2013)—they are strategies for developing mechanistic hypotheses which are not guaranteed of success. They must be further tested. One measure for evaluating them is further mechanistic research on the proposed mechanisms themselves, showing how they function in healthy cells and how they can induce hallmarks of cancer when disrupted. Another, invoked in several of the examples discussed below, is to show that they enable better stratification of patients in terms of outcomes and responses to therapies.

2. The Discovery of Heterogeneity

The quest to find altered genes as the causes of cancerous states within cells was galvanized in the 1980s by the identification of two different classes of genes that were discovered to be altered in tumors—oncogenes, which were hypothesized to generate cancer when mutated, and tumor suppressor, hypothesized normally to prevent the transition to cancer but allow it when altered. The discovery of the first oncogenes, *Hras* (*H* for *H*arvey and *ras* for *ra*t *s*arcoma) and

[2] For discussion of the opposition between what has been dubbed the somatic mutation theory and the tissue organization field theory, see Bertolaso (2016), (Plutynski, 2018) and (Green, forthcoming).

Kras (*K* for Kirsten) (Ellis, Defeo, Shih et al., 1981), grew out of research that viewed tumors as induced by viruses but ended up focusing attention on gene alterations as causes of cancer.[3] The proposal that some genes normally suppress tumors but allow them when mutated grew out of Knudson's (1971) hypothesis that development of some cancers require two independent mutations (two hits), where, in some cases, the first hit involves a gene whose function is normally to suppress development of cancer. Although it does not function in the two-hit scenario, *TP53*, mutated in more than 30% of human tumors, is the best-known tumor suppressor gene.

The discoveries of oncogenes and tumor suppressor genes encouraged researchers to seek genes in which alterations caused cancer, an endeavor that was much enhanced with the development of high-throughput gene sequencing techniques in the 1990s. To make this growing body of data on cancer genes available to the larger community, Futreal, Coin, Marshall et al. (2004) conducted what they termed "a census of human cancer genes." One of the questions Futreal et al. faced in determining which genes to include in the census was differentiating altered genes that play a causal role in cancer (which for them meant conferring "a clonal growth advantage")[4] from what they identified as passenger or bystander mutations ("Somatic mutations that are found in cancer cells that are not involved in generating the neoplastic phenotype"). To avoid including passenger mutations, Futreal et al. simply "excluded genes in which fewer than five unambiguous somatic mutations have been reported in primary neoplasms," assuming that genes that do not play a causal role are more likely to vary than those that do. Even using this criterion, Futreal et al. identified 291 genes, all coding for proteins. They found this number surprisingly large as it amounts to somewhat more than 1% of known coding genes in humans. A further surprise was that even genes that were frequently mutated were not mutated in all tumors affecting a given tissue. This began to draw attention to heterogeneity of gene alterations in tumors as a serious challenge in identifying genes responsible for cancer.

In the same year as Futreal et al.'s census was published, another group of researchers at the Sanger Institute in London made public the Catalogue of Somatic Mutations in Cancer (COSMIC) online database (Bamford, Dawson, Forbes et al., 2004). Initially, COSMIC selected four cancer genes, *Hras*, *Kras2*, *Nras*, and *Braf*. The curators searched PubMed and extracted information from the identified publications about samples, experimental methods, and mutations. Within a year COSMIC had expanded to include 28 known cancer genes. In addition to published sequences, the researchers also included data from their own Cancer Genome Project that by 2005 had re-sequenced known cancer genes in 728 publicly available cell lines with a goal of identifying novel oncogenes. Altogether, that expanded coverage to 538 genes and 124,367 tumors with 23,157 mutations (Forbes, Clements, Dawson et al., 2006). In the ensuing decade, COSMIC has continued to expand rapidly and provides further evidence about just how heterogenous is the set of gene altered in cancer.

---

[3] For historical reviews of the discovery of the *Ras* genes and the development of the oncogene framework, see Morange (1993, 1997, 2001); Malumbres and Barbacid (2003).

[4] Cancer researchers now generally refer to those genes that play a causal role in cancer as *drivers* (Greenman, Stephens, Smith et al., 2007; Stratton, Campbell, & Futreal, 2009).

The heterogeneity of genes implicated in cancer became even more apparent with a paper by Wood, Parsons, Jones et al. (2007).[5] These researchers sequenced about 13,000 genes from 11 breast and 11 colorectal cancer patients and reported significant mutations in almost 200, with a mean of 76 mutations resulting in altered amino acids in proteins in individual breast cancer tumors and 84 mutations in colorectal cancer tumors. The well-known oncogenes and tumor suppressor genes were among the frequently mutated, but there were many samples in which no frequently mutated gene was found. This led the authors of offer a new vision of cancer genome landscapes: "They are composed of a handful of commonly mutated gene 'mountains' but are dominated by a much larger number of infrequently mutated gene 'hills'."[6] The challenge was to make sense of how mutations in the genes constituting the hills contributed to cancer.

The heterogeneity problem grew steadily with the pursuit of yet another extremely large-scale research endeavor, The Cancer Genome Atlas (TCGA) was created in 2008 as a joint initiative by two institutes within the U.S. National Institutes of Health, the National Cancer Institute and the National Human Genome Research Institute. The project set out to collect, sequence, and distribute approximately 500 samples of tumors in different organs and deposit the data in publicly accessible databases. TCGA began with glioblastoma multiforme in the brain, squamous cell carcinoma of the lung, and cystadenocarcinoma of the ovary and eventually expanded to cancers affecting 33 different organs. Under the name The Cancer Genome Atlas Research Network, TCGA researchers published characterizations of many cancer types, including human glioblastoma (2008), breast (2012), lung (2012), colon and rectal (2012) cancers, clear cell renal cell carcinoma (2013), acute myeloid leukemia (2013), endometrial carcinoma (2013), urothelial bladder carcinoma (2014), and gastric adenocarcinoma (2014). These studies often revealed previously unsuspected genes implicated in cancers in particular tissues. The first study identified three previously unsuspected genes as frequently mutated in glioblastoma: *NF1*, previously implicated in neurofibromatosis, *ERBB2*, previously identified in breast cancer, and *PIK3R1*, part of the PIK3 signaling pathway that was known to be abnormally activated in a number of cancers (Cancer Genome Atlas Research Network, 2008).

TCGA revealed additional heterogeneity in the relation between genes and cancer. In addition to continually identifying additional genes mutated in cancers in different tissues, it revealed a serious problem with typing cancers by the tissues in which they occurred. This resulted both in missing important differences in terms of altered genes between cancers that affected the same tissue and commonalities between cancers that affected different tissues. For example, even though TCGA had set out to study colon and rectal cancers separately, they discovered that the genomic alterations are very similar and concluded that the two cancer types should be grouped as one (Cancer Genome Atlas Research Network, 2012a). TCGA's breast cancer study (Cancer Genome Atlas Research Network, 2012b) reaffirmed and further characterized the four subtypes

[5] Other papers of the same period reached similar conclusions: Thomas, Baker, DeBiasi et al. (2007); Annunziata, Davis, Demchenko et al. (2007); Keats, Fonseca, Chesi et al. (2007).

[6] Ideker pithily captures the problem posed by heterogeneity: "heterogeneity by definition means that recurrent patterns are not observed for most mutations. To make matters worse, patients afflicted by such unique patterns of mutations have been labeled 'N-of-1s,' to capture the idea that they cannot be joined together with any other individuals to be analyzed and treated as a larger cohort (i.e., of size N>1). Patients enduring this desultory fate stand alone, without a friend even in disease" (Ideker, 2016).

of breast cancer that had already been arrived at by earlier analyses. However, the researchers also found that the basal-like subtype exhibited a similar pattern of gene mutations to that found in serous ovarian cancer, suggesting that they constitute a common form of cancer. Similarly, the endometrial cancer study (Cancer Genome Atlas Research Network, 2013) went beyond the traditional classification of endometrial cancers into entometroid (class 1) and serous (class 2) by identifying a subset of entometroid tumors that clustered with serous tumors and showed that these manifest strong similarities to serous ovarian cancer and basal-like breast cancer. The remaining entometroid tumors formed three classes: a newly discovered group with mutations in *POLE*, those that exhibited microsatellite instability, and those with low copy number alterations. These entometroid tumors share characteristics with colorectal tumors that TCGA had previously characterized. In recognition of the fact that "cancers of disparate organs have many shared features, whereas, conversely, cancers from the same organ are often quite distinct," TCGA developed a new pan-cancer initiative that began by integrating the datasets from 12 individual cancer types already analyzed (Cancer Genome Atlas Research Network, Weinstein, Collisson et al., 2013).[7]

The official TCGA project wound down 2017,[8] but the datasets it produced have provided data for an extensive set of network studies of cancer and a sharpened recognition of how heterogeneous the genetic alterations in cancer are. Drawing upon the results of Wood et al. as well as those TCGA and COSMIC, Garraway and Lander (2013) concluded that very few genes are altered in greater than 10% of samples of a given cancer. Moreover, a very large number are mutated in less than 5% of samples. This is referred to as the *long tail* of the distribution. The recognition of this large-scale heterogeneity of genes altered in cancer[9] posed a challenge to attempts to explicate the transition of a cell into cancer at the genetic level.

3. Moving Up from Genes to Mechanisms

As they confronted the heterogeneity problem, a number of researchers concluded that in searching for genes responsible for cancer, they had focused at too small a scale. The proteins synthesized from genes work together in larger-scale units that biologists refer to as mechanisms. Although scientists commonly invoke the term *mechanism* without clarifying what they have in mind, the sense seems to correspond to that advanced by the new mechanists in philosophy of science—a set of components that perform different operations and are organized so as to work together in the generation of a phenomenon (Machamer, Darden, & Craver, 2000; Bechtel & Abrahamsen, 2005; Glennan, 2017).

---

[7] An additional motivation for the pan-cancer initiative was that by combining data across cancer types, studies would have increased statistical power and be better able to identify infrequently occurring driver mutations. See Tamborero, Gonzalez-Perez, Perez-Llamas et al. (2013) for some of the new discoveries resulting from this effort.

[8] The endeavor to collect data and genetically characterize various cancers is being continued by the International Cancer Genome Consortium (ICGC), which started in 2008 (the same year as TCGA). ICGC is collaborating with TCGA in the Pan-Cancer Analysis of Whole Genomes.

[9] Yet a further source of heterogeneity is found if one compares cells within the same tissue sample (Fisher, Pusztai, & Swanton, 2013).

Just as there are multiple parts to a mechanism, there are multiple ways in which a mechanism can be incapacitated. From the point of view of the system that depends on what the mechanism as a whole does, which way the mechanism is incapacitated may not matter. A potential reason why alterations to any of a heterogenous set of genes may result in a similar cancer is that each of the resulting proteins figures in the operation of the same mechanism. In whatever way the mechanism is altered, it ceases to function as it normally would. In the case of cancer, many of the mechanisms altered are control mechanisms that in normal cells down-regulate other mechanisms such as the cell cycle. Any mutation that impairs a control mechanism from down-regulating cell division will result in uncontrolled cell division, one of the main hallmarks of cancer.[10]

In the rest of this paper I focus on two strategies through which researchers made the transition from focusing on genes to focusing on mechanisms, one involving the identification of pathways and one involving identification of clusters in networks. Mechanists in philosophy of science tend to count any set of components that causally interact in the generation of a phenomenon as a mechanism. Ross (forthcoming), however, argues for distinguishing pathways and mechanisms as distinct causal concepts. She is correct that there are distinctive features of the way scientists investigate pathways. For instance, those investigating a pathway are more concerned with the sequence of intermediate products than with accounting for how each is generated. The notion of a pathway has its roots in biochemistry. For example, once Buchner (1897) demonstrated that fermentation can occur in a cell-free extract, researchers started identifying intermediates in the generation of alcohol from glucose and trying to link them together in a continuous sequence. This effort culminated in the 1930s in the pathway proposed by Embden and Meyerhof (Bechtel, 1986) that is still accepted today. As molecular biologists turned their attention to signaling processes, they also identified multi-step pathways in which intermediates are generated sequentially until the final signal is produced.

On their own, pathway accounts leave out an important feature emphasized in accounts of mechanistic explanations—the activity or operation involved in generating each subsequent step in the pathway. For example, the mechanism of fermentation involves not just the sequence of reactions but the enzymes that catalyze the various reactions. Nonetheless, researchers often view pathways as an important component of an account of a mechanism, and I will therefore treat pathways as (partial) accounts of mechanisms. There is, however, an important contrast to make: many mechanisms involve multiple parts interacting in the production of the phenomena, not just the sequence of intermediates. Interacting components are often represented in networks, with nodes representing entities and edges the interactions. Large networks, however, often resemble hairballs until they are laid out in an informative manner. A common strategy in network analysis is to identify clusters of highly interconnected components and position these near each other. Researchers often try to identify these highly interactive clusters with mechanisms that have been identified and investigated through more traditional techniques of cell and molecular biology. It should be noted that network accounts of mechanisms, like pathway accounts, are incomplete. In fact, what they often leave out is a specification of the reaction pathway. Thus, pathways and network clusters each offer partial insights into

[10] For the distinction between production and control mechanisms and its relevance in the case of cancer, see (Bechtel, 2018).

mechanisms, but these are often enough to leverage raising the level of inquiry from individual genes or proteins to mechanisms.

The distinction between pathways and network clusters is illustrated in Figure 1, which presents a pathway representation on the left and a cluster in a network representation on the right. Both involve the same proteins, shown in green. As many of the proteins synthesized by early identified oncogenes appeared to figure in signaling processes (which then control mechanisms such as the cell cycle), it was natural to try to organize them into pathways. In some cases, the knowledge needed to construct a pathway was already available in basic biology before the gene alteration leading to cancer was identified. In many cases, however, this knowledge had to be generated by first identifying a gene that is altered in tumors and then investigating the reactions in which the corresponding protein figured. Figure 1a shows the first steps in the epidermal growth factor (EGF) signaling pathway. The small white boxes indicate reactions, green boxes the proteins figuring in the pathway and blue boxes the complexes formed: EGF forms a complex with the EGF receptor (EGFR) and in subsequent reactions is phosphorylated, yielding EGF-p-6Y-EGFR.

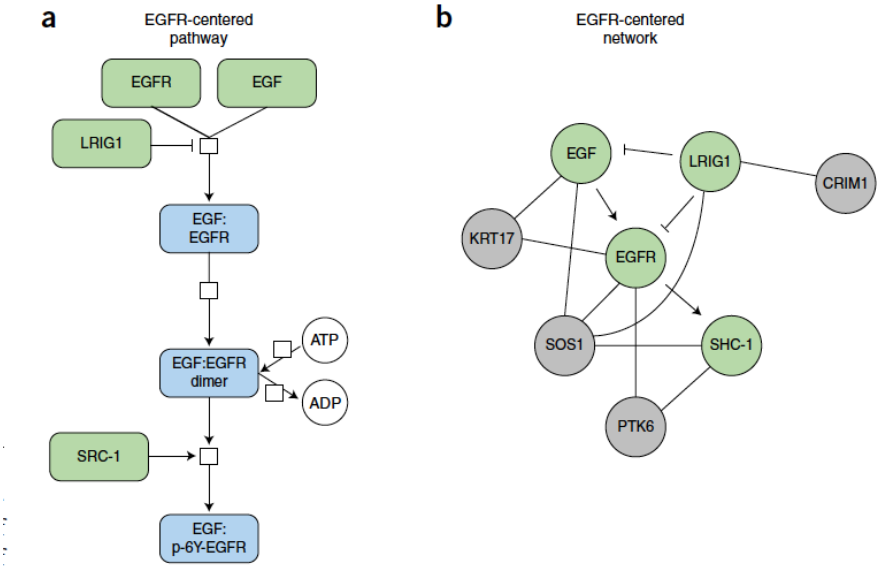


Figure 1. A. A pathway diagram of initial stages in EGF signaling that shows the reactions (white boxes) in a particular reaction pathway. Green boxes represent proteins and blue boxes the resulting complexes. B. A network diagram in which green circles represent the same proteins as in 1A, and the grey circles proteins that interact with them. Reprinted by permission from Springer Nature, *Nature Methods*, Creixell, P., Reimand, J., Haider, S., Wu, G. M., Shibata, T., et al., Pathway and network analysis of cancer genomes, ©2015.

Identifying clusters through network analyses provides a different strategy for arriving at mechanisms. These approaches begin with information about which cell components interact with each other (e.g., proteins that are capable of forming bonds or actually do form bonds with each other in a given cell type). Another data type involves synthetic lethality in which knocking

out either of two genes individually leaves the organism viable but knocking out both kills it.[11] Increasingly data about such interactions are stored in large, publicly accessible databases such as BIND (Bader, Donaldson, Wolting et al., 2001) or MINT (Zanzoni, Montecchi-Palazzi, Quondam et al., 2002), which researchers can then access. Using tools such as Cytoscape (Shannon, Markiel, Ozier et al., 2003; cytoscape.org), they can identify clusters and layout nodes and edges in an informative fashion. Figure 1b shows such a cluster that corresponds to the pathway in Figure 1a. In one sense, it shows less that Figure 1a since it does not show the intermediates. On the other hand, since it is built from data about all interacting proteins, network representations can include proteins that have not been fitted into pathway accounts. Thus, Figure 1b includes several nodes shown in grey circles whose function are not known. Edges connecting them to other proteins are undirected since even the direction of effect is not known. Those shown as green circles correspond to proteins whose functions are known from other sources (and included in the pathway in Figure 1A). Since what is known includes the direction of causation, the connections between known components are indicated by directed edges.

Researchers identifying pathways and researchers identifying clusters in networks employ different heuristic strategies, but both end up revealing sets of organized components that researchers treat as mechanisms. Each of these ways of identifying mechanisms has proven useful in addressing the heterogeneity problem. Whether they characterize mechanisms as pathways or network clusters, researchers can appeal to these higher-level entities as the entity that operates differently no matter which of its components is altered. The following two sections illustrate the use of pathway and network analysis strategies.

4. Illustrations of Pathway Heuristics for Addressing Heterogeneity

Above I focused on how TCGA sequencing studies identified new genes as altered in various cancer types, thereby increasing the heterogeneity problem. In their analyses, the TCGA researchers often drew upon pathways as a way to address the problem. The first released study, on glioblastoma, identified three signaling pathways that were disrupted in more than three quarters of the glioblastoma samples: the cyclin-dependent kinase/retinoblastoma pathway (RTK/RAS/PI(3)K) that regulates cell division was disrupted in 88%, the TP53 signaling pathway that initiates DNA repair and apoptosis in 87%, and receptor tyrosine kinase pathway involved in controlling cell growth in 78% of samples. The fact that the pathways were much more frequently altered than were individual genes (*CDKN2A* at 52% and *TP53* at 35% were the most frequently mutated genes) pointed to the pathways as the relevant units of analysis for avoiding the heterogeneity problem. Other genes in these pathways were mutated less frequently but were construed as having the same effect in generating glioblastomas. Moreover, the study proposed that the pathway affected might provide insight into the success of treatments:

[11] Until recently, successful synthetic lethal experiments were limited to yeast as RNAi based methods for inhibiting proteins had too many off-target effects. Recently, CRISPR technology has proven effective in identifying synthetic lethal pairs in mammalian cells and offers promise for contributing to the development of new therapeutic approaches that target genes that are synthetically lethal in particular types of cancer (Shen, Zhao, Sasik et al., 2017; Du, Roguev, Gordon et al., 2017).

> It would be reasonable to speculate that patients with deletions or inactivating mutations in CDKN2A or CDKN2C or patients with amplifications of CDK4/CDK6 would be candidates for treatment with CDK inhibitors, a strategy not likely to be effective in patients with RB1 mutation. Similarly, patients with PTEN deletions or activating mutations in PIK3CA or PIK3R1 might be expected to benefit from a PI(3)K or PDK1 inhibitor, whereas tumours in which the PI(3)K pathway is altered by AKT3 amplification might prove refractory to those modalities (Cancer Genome Atlas Research Network, 2008, p. 1066).

The appeal to pathways to explain features of cancer began well before TCGA. Hanahan and Weinberg (2000) identified what they characterized as six hallmarks of cancer: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis.[12] When they turned to explaining how these hallmarks were realized, TCGA researchers sought to arrange individually identified genes or gene products into already known pathways in which one affected another, eventually affecting the cell cycle or other mechanism responsible for a given hallmark. Figure 2 is their diagram showing pathways involved in cell proliferation and programmed cell death (apoptosis).[13] Growth factors, known to promote cell proliferation (by inhibiting components that block proliferation), are shown binding a receptor on the left. Binding to the receptor initiates activity along different pathways, including one involving RAS, RAF, MEK, MAPK, and MYC. Mutations to various components of the pathway result in cancer cells continuing to proliferate. The lower-level details about the operations of individuals genes fit naturally into this pathway analysis. For example, RAS was known to function as a GTPase, and when it hydrolyzes GTP to GDP, it renders itself inactive. Hence, the normal control signal from RAS is of short duration. But when the gene is altered, RAS is unable to hydrolyze GTP. The result is that it remains in the active form and initiates an ongoing proliferation signal. What the focus on the pathway makes clear is that the alteration of RAS as well as alterations to other components of the pathway, such as NF1, RAF, and MYC, all have the effect of sustaining proliferation signaling along the pathway. This explains why mutations to each of them lead to sustained cell proliferation.

---

[12] In a ten year update, Hanahan and Weinberg (2011) added two *emerging hallmarks*: reprogramming of energy metabolism and evading immune destruction.

[13] The Atlas of Cancer Signalling Network provides a more recent, online (https://acsn.curie.fr/), representation of pathways involved in cell regulation that are affected in cancer (Kuperstein, Bonnet, Nguyen et al., 2015). To date it includes separate networks for cell cycle, DNA repair, apoptosis, epithelial-to-mesenchymal transition and motility, and survival that are integrated into a cohesive whole. As with Google Maps, one can zoom in to look at relations of individual genes in detail. One can also click on them for further information. In addition, it is possible to locate mutations in various cancers on the map to assess how they affect cell signaling.

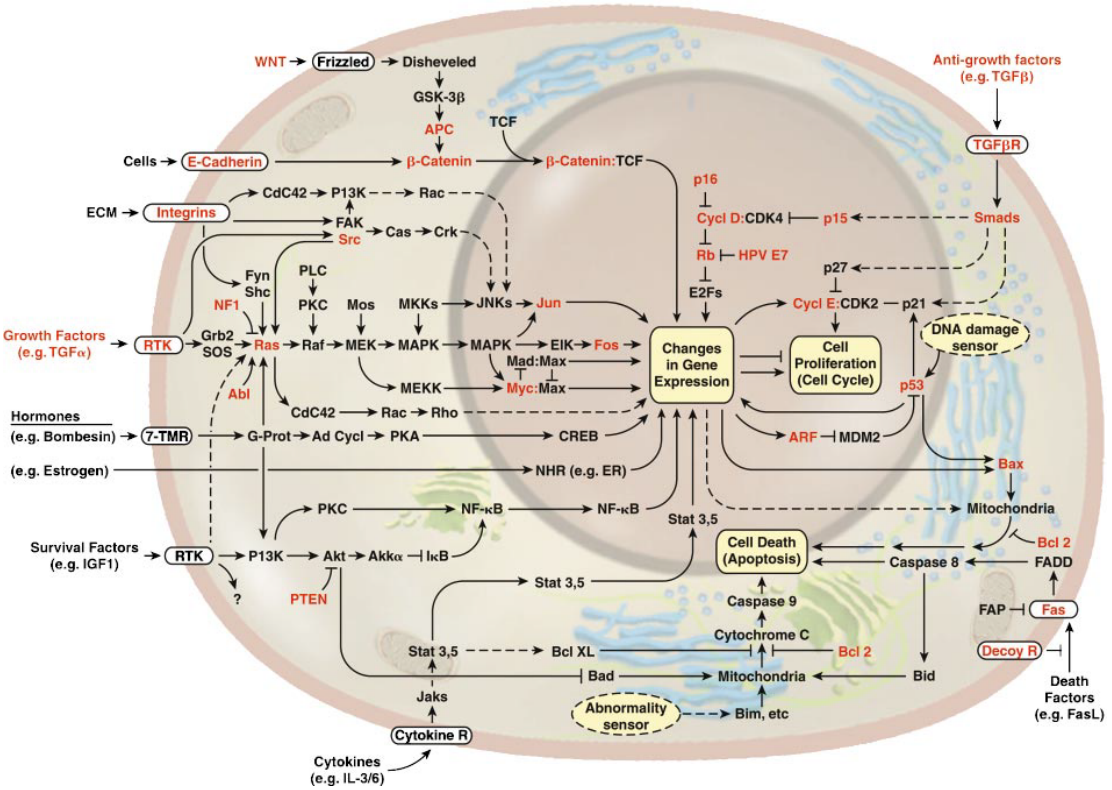


Figure 2. Hanahan and Weinberg's representation of multiple pathways in which mutations can lead to hallmarks of cancer. The proteins coded by the best-known oncogenes (*Ras*, *Myc*) and tumor suppressor genes (*p53* [same as *TP53*] and *PTEN*) are shown in red. Reprinted from *Cell*, Vol 100, Hanahan, D., & Weinberg, R. A, The hallmarks of cancer, Figure 2, ©2000, with permission from Elsevier.

Vogelstein, Papadopoulos, Velculescu et al. (2013) provide a clear, illustrative example of how this sort of pathway analysis can explain heterogeneous mutations generating the same type of cancer and draw out the implication that consequently mutations affecting the same pathway should not occur in the same tumor:

> Recognition of these pathways also has important ramifications for our ability to understand inter-patient heterogeneity. One lung cancer might have an activating mutation in a receptor for a stimulatory growth factor, making it able to grow in low concentrations of epidermal growth factor (EGF). A second lung cancer might have an activating mutation in KRAS, whose protein product normally transmits the signal from the epidermal growth factor receptor (EGFR) to other cell signaling molecules. A third lung cancer might have an inactivating mutation in NF1, a regulatory protein that normally inactivates the KRAS protein. Finally, a fourth lung cancer might have a mutation in BRAF, which transmits the signal from KRAS to downstream kinases. (p. 1555).

A focus on pathways has the potential to radically reduce the heterogeneity problem. Vogelstein et al. (2013) contend that all known cancer driver genes reside in 12 pathways that control 3 processes—cell fate, cell survival, and genome maintenance. This offers great promise for developing accounts of cancer that generalize across specific gene alterations. Enthusiasts for the

pathway perspective, such as Vogelstein and Kinzler (2004), foresee it as bringing order to the heterogeneity of mutations. They propose that even if research reveals a few more pathways, there will be a relatively small number (on the order of 20) of pathways that, when disrupted, result in cancer.

5. Illustrations of the Network Clustering Heuristic for Addressing Heterogeneity

When researchers possess the knowledge, or are able to procure the knowledge, needed to arrange genes altered in cancer into pathways, the pathway become a relevant explanatory unit. However, construction of pathways requires detailed knowledge of the sequence of activities in which proteins engage. Many genes identified as altered in tumors cannot, given current knowledge, be fit into pathways. Hu, Bader, Wigle et al. (2007) observed that of the 291 genes Futreal et al. had identified as cancer genes

> only 28% currently have extensive functional associations in the Kyoto Encyclopedia of Genes and Genomes (KEGG), while only 59%, 58%, 48% and 26% are listed in the IntAct, Biomolecular Interaction Network Database (BIND), Molecular Interaction Network (MINT), and Database of Interacting Proteins (DIP), respectively. This indicates that roughly half of all established cancer genes still lack functional-association information in the main public functional-association databases."

Researchers cannot assign proteins to pathways if they do not know the reactions in which they are involved. Network approaches, which require only more basic information such as which proteins can interact with each other or which genes form synthetic lethals, provide strategies for overcoming this limitation. Above I described the use of cluster analysis to identify clusters of highly interacting genes or proteins that may correspond to mechanisms. To determine what these clusters and their components do, researchers often annotate nodes in networks using Gene Ontology or GO (Ashburner, Ball, Blake et al., 2000). GO draws from the published literature information such as where in the cell a gene is expressed or what cellular function it figures in and organizes this information into hierarchical representations in the form of directed acyclic graphs.[14] To formulate hypotheses about the function of genes or proteins for which there is no current knowledge (e.g., they are not annotated in GO) researchers often employ a heuristic known as *guilt by association*: when an entity without a known function is grouped into a cluster with others that have a known function, assume that it should be assigned the same function (Bechtel, 2017; 2019, presents examples of such inferences in yeast biology).

Hu et al. (2007) illustrates the use of this strategy in cancer research. MLLT2 is mutated in leukaemogenesis in infancy but has no biological process annotation in GO. To develop a hypothesis about its function, the researchers first situated it in a protein-protein interaction network and identified a sub-network of proteins that directly bind to it (GNA11, GNAI3 and NACA, shown in the inner circle in Figure 3). They then added those proteins that bind to these proteins (a sample is shown in the outer circle). What they found noteworthy is that 104 of these proteins, including the immediate neighbors GNA11 and GNAI3, had previously been linked to G-protein coupled receptor (GPCR) signaling. From this they infer that MLLT2 likewise contributes to GPCR signaling. This inference is of course fallible and needs to be evaluated

[14] For a detailed analysis of the construction of GO, see Leonelli (2016)

using more traditional molecular techniques; guilt by association is a heuristic reasoning strategy that can initiate such investigation.

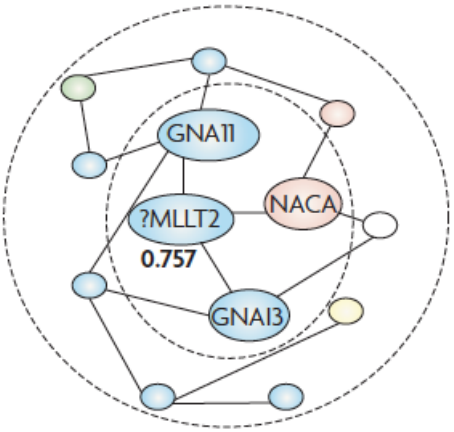


Figure 3. Hu et al. (2007) invoked network connections to propose a G-protein receptor function for MLLT2, which had no annotation in GO. GO annotations are indicated by color: transcription (pink), G-protein coupled receptor (blue), unknown (white) and other functions (green and yellow). The confidence score for the prediction of G-protein coupled receptor is shown. Reprinted by permission from Springer Nature, *Nature Reviews Cancer*, Hu, P. Z., Bader, G., Wigle, D. A., & Emili, A. Computational prediction of cancer-gene function, ©2007.

A second example illustrates the power of this approach to identify mechanisms in which genes altered in tumors participate. Chuang, Lee, Liu et al. (2007) sought to distinguish among breast cancer patients those whose tumors metastasized from those whose tumors did not metastasize. They began with expression profiles in patients whose tumors metastasized and those that did not and identified 8,141 genes that showed differences. They overlaid these on a protein-protein interaction network and searched for subnetworks in which expression discriminated patients that metastasized. From one of their datasets, which they took from TCGA, they identified 149 subnetworks, some of which are shown in Figure 4.

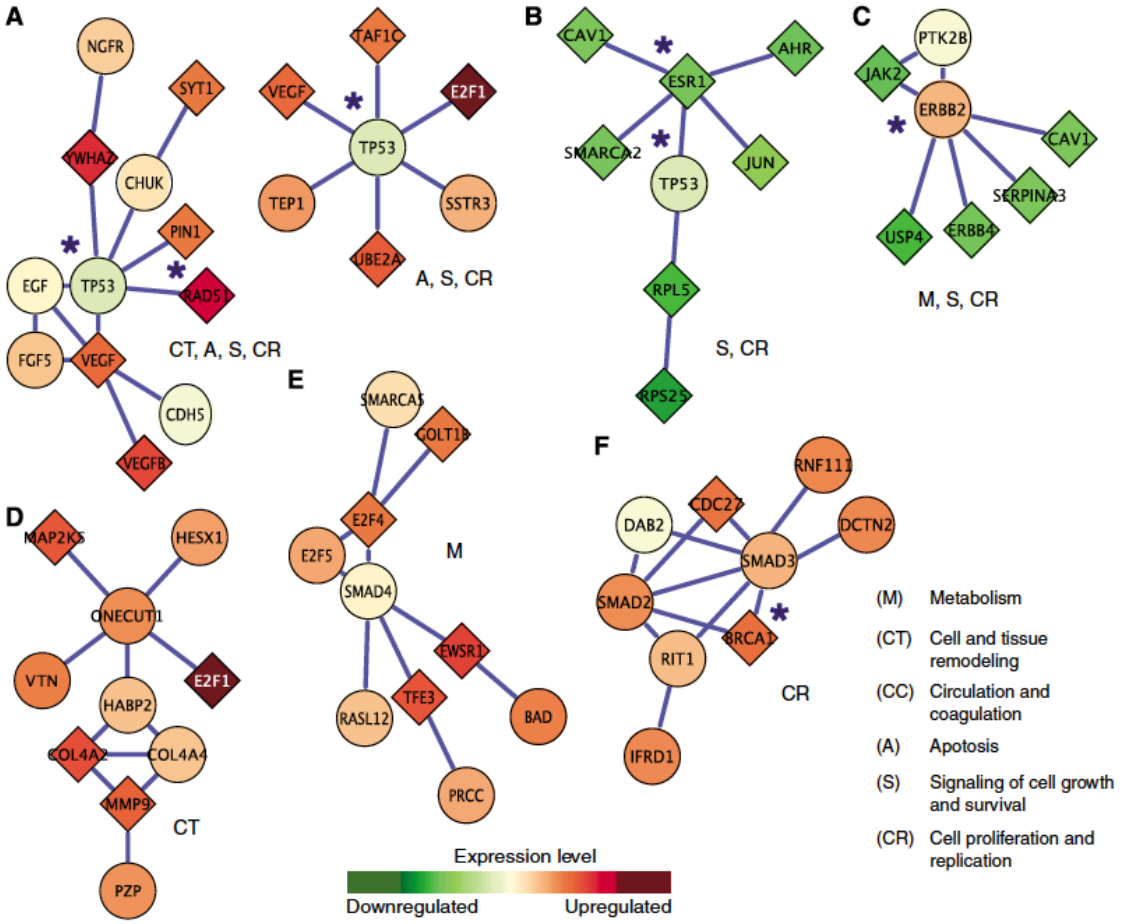
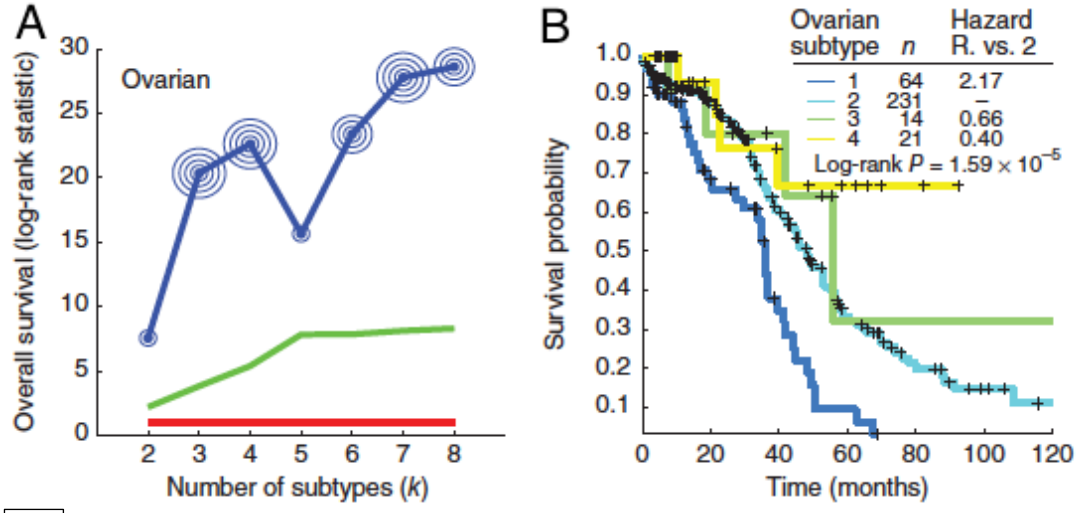


Figure 4. Subnetworks identified by Chuang et al. (2007) of proteins that were differentially up- or down-regulated (indicated by color) in breast cancer patients whose tumors metastasized. Letters next to each subnetwork indicates cell processes in which the subnetwork is involved. Reprinted with permission from John Wiley and Sons, *Molecular Systems Biology*, Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T., Network-based classification of breast cancer metastasis, ©2007.

When they annotated the proteins using GO, somewhat more than half the subnetworks were enriched for proteins assigned to at least one biological process. (In Figure 4 the biological process is identified by the letters next to the subnetworks, which are interpreted in the legend.) Many of these are processes that figure in Hanahan and Weinberg's hallmarks of cancer: proliferation and replication, apoptosis, circulation, and metabolism. The color of nodes indicates whether the expression is up-regulated or down-regulated in tumors that later metastasized, and diamonds indicate that the change in expression is statistically significant. Chuang et al. showed that after scoring subnetworks in terms of average increased or decreased expression of proteins in the network, they could train a classifier based on logistic regression to predict metastasis with ~70% accuracy, which is much higher than models based on individual genes. They take this result to indicate that the subnetworks they identify are mechanisms that differentially determine whether the tumor will metastasize. By focusing on these mechanisms, heterogeneity is significantly reduced.

A relatively recent promising network analysis strategy for identifying mechanisms altered in tumors treats the genes that are modified as sources of heat and applies a diffusion algorithm to distribute the heat to nodes nearby in the network. (Since the whole network is connected, the duration of diffusion must be limited; otherwise heat will disperse and reach equilibrium over the whole network.) This strategy is particularly effective when heat from multiple nodes diffuses into the same cluster, which can then be identified as the relevant mechanism that, when disrupted by alteration of any of the various genes, results in cancer.

Hofree, Shen, Carter et al. (2013) illustrate use of diffusion to stratify patients with ovarian, uterine, and lung cancer into patient groups that exhibit similar outcomes (measured in terms of survival, response to drugs, etc.). Their hypothesis was that the similar outcomes might result from mutations affecting a common underlying mechanism. Their Network Based Stratification (NBS) approach first locates altered genes in a network. They then apply a network propagation algorithm developed by Vanunu, Magger, Ruppin et al. (2010) to spread activity over the neighborhoods around these genes.[15] Based on the resulting values, they cluster nodes into a varying number of clusters that they viewed as potentially corresponding to subtypes of these cancers. Finally, they evaluate how well membership in a cluster predicted patient outcome. Figure 5A compares the performance of NBS (blue) in predicting ovarian cancer patient outcome when patients were clustered into various numbers of subtypes compared to standard clustering (red) or a permuted version of NBS (green). The number of concentric circles around a data point indicates significance (*p*-value). When divided into 3 or 4 subtypes, NBS's improvement in predicting patient outcome was highly significant ($p < 0.0001$). Figure 5B presents a Kaplan-Meier analysis showing duration before relapse after treatment with platinum chemotherapy when NBS identified four subtypes. Pluses indicate time of relapse for individual patients, with their location with respect to the y-axis indicating the percentage of patients that have still not relapsed at that point. The colored lines connect these points for each group. When it creates four clusters, NBS differentiates four subtypes of ovarian cancer with different periods to relapse.



[15] There are several additional network diffusion algorithms that researchers have applied to cancer data such as HotNet (Vandin, Clay, Upfal et al., 2012; Vandin, Upfal, & Raphael, 2011) and HotNet2 (Leiserson, Vandin, Wu et al., 2015).

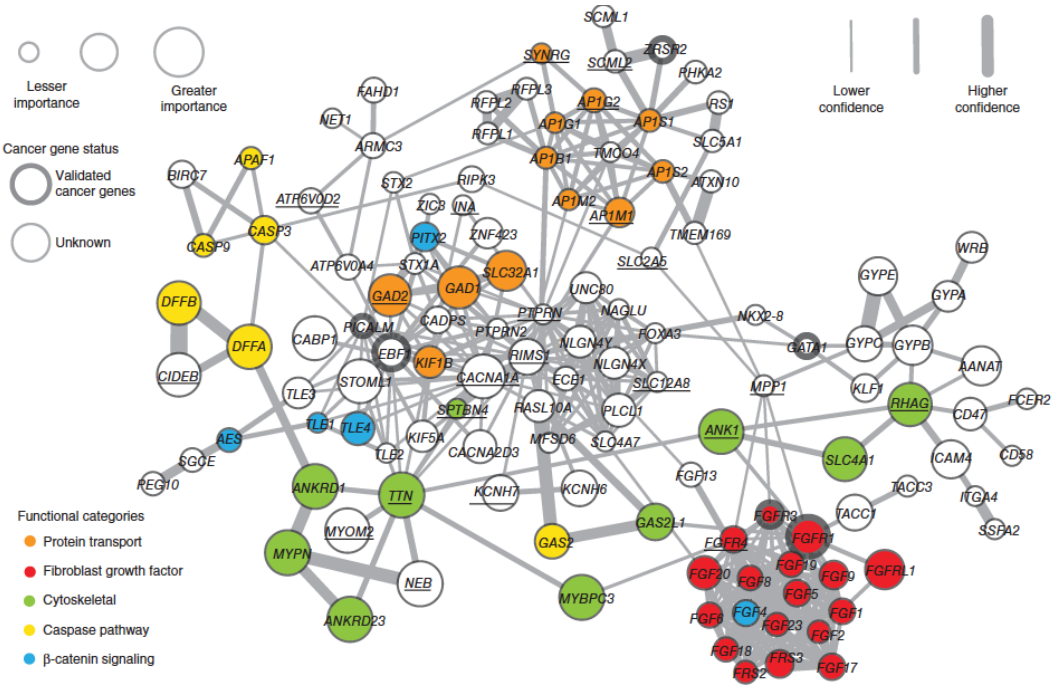


Figure 5. A. Comparison of Network Based Stratification (NBS; blue) versus standard clustering (red) or a permuted version of NBS (green) in predicting patient survival when clustered into different number of groups. Number of blue circles indicates *p*-value. B. Kaplan-Meier analysis of survivability when NBS stratified samples into four groups. C. Identification of most active subnetworks in the subtype of ovarian cancer with the poorest prognosis. Actually mutated genes are indicated by underlining their names. These were treated as hot spots and node size indicates score after diffusion. Color indicates annotation to specific cell functions, as indicated in the legend. Reprinted by permission from Springer Nature, *Nature Methods*, Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. Network-based stratification of tumor mutations, ©2013.

Hofree et al. then examined the subnetworks that were most active in the four different subtypes. Figure 5C shows the subnetwork most involved in the first (poorest prognosis) subtype of ovarian cancer. Mutated genes (indicated by underlining their names) were plotted on an interaction network. Edge width indicates degree of confidence that there is an interaction between the gene products while the size of the circle for a gene indicates the mutation score after diffusion. The researchers used GeneMANIA to annotate the genes in terms of cell function. Genes already assigned a role in cancer in COSMIC are shown with thickened borders. The network reveals clusters of genes associated with the mutated genes in this subclass of ovarian patients. The genes in these clusters are hypothesized to function together in mechanisms contributing to the designated cell function. For example, the genes indicated in red are involved in the fibroblast growth factor signaling pathway. The one gene mutated in the cluster, *FGFR4*, was not a known cancer gene, but activity spread through the interconnections to other genes, including two known cancer genes. The authors hypothesize that *FGFR4* drives cancer by altering the same mechanism as these other genes. Such hypotheses must be tested experimentally; the objective of network analysis is only to generate plausible hypotheses for further testing.

The network analysis strategies presented in this section each reveal clusters of nodes that can be interpreted as cellular mechanisms. By identifying those clusters in which mutated genes reside or that become targets of activity using diffusion, researches target those mechanisms that are affected and whose altered operation may explain cancer. As with the pathway strategy, these higher-level units become the relevant explanatory units, significantly reducing the heterogeneity problem.

6. An Illustration Combining Pathway and Network Heuristics to Address Heterogeneity

In the previous two sections I have presented examples in which pathway and network heuristics have been applied separately. In a study of glioblastoma Wu, Feng, and Stein (2010) showed how they can be productively combined. They began with a network approach. Drawing upon multiple sources, the researchers generated what they termed a Functional Interaction (FI) network of 10,956 proteins and 209,988 interactions. Wu et al. then integrated FI with a pathway approach. They identified 73 proteins in TCGA's glioblastoma pathways. They then use FI to add proteins that interacted with one or more of these proteins. This effectively selected a subnetwork out of FI whose components are plausibly linked to glioblastoma. Two segments are shown in Figure 6. The nodes in grey were included in the TCGA pathway, those in blue are added from FI (mostly connected with undirected edges since pathway information is lacking). From this network, the authors generated hypotheses about how mutations lead to cancer. One hypothesis involves NUP50, shown in the left panel. It has a reduced copy number in three TCGA samples. Since it is connected to CDKN1B in the network, the authors propose that it is required for degradation of CDKN1B and its altered copy number contributes to glioblastoma by causing increased activity of CDKN1B in the cell cycle. In the right panel, tenascin-C (TNC), mutated in three TCGA samples, is shown as a ligand for epidermal growth factor receptor (EGFR). Since EGFR is upstream of the RAS complex, the authors propose that mutation of TNC could contribute to cancer by up-regulating RAS, resulting in uncontrolled proliferation.

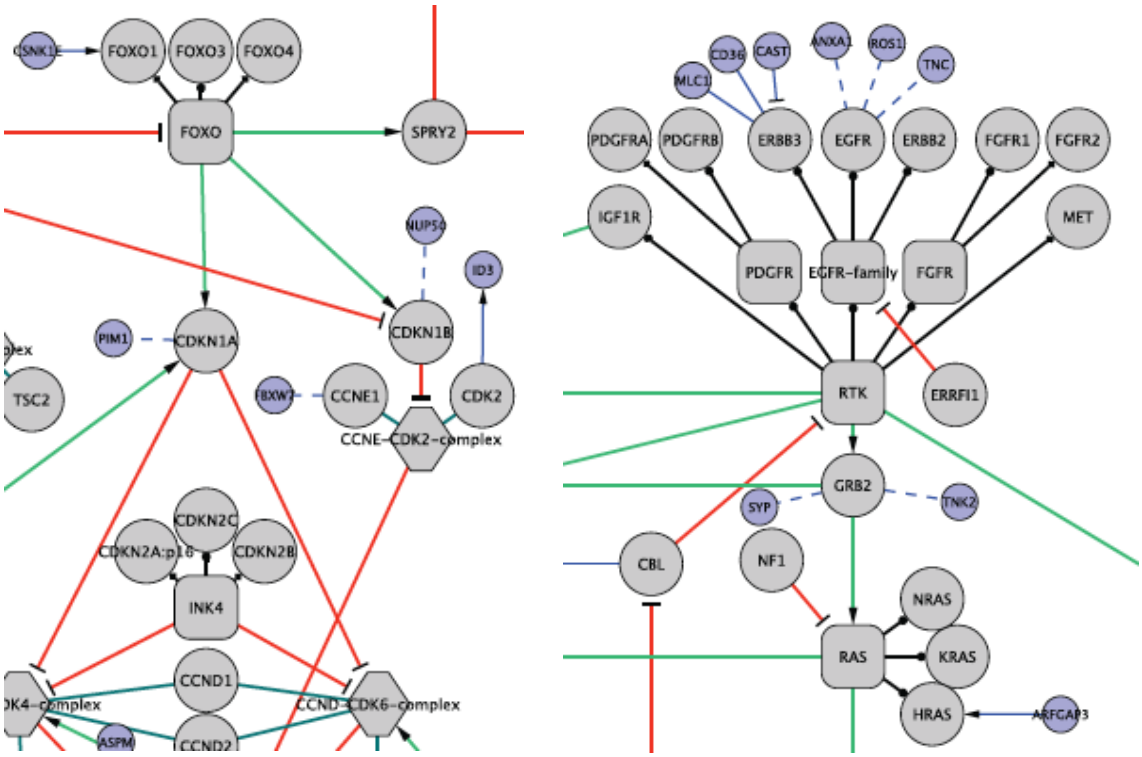


Figure 6. Two portions of the Wu et al.'s FI network. Large circles identify genes mutated in TCGA. The small nodes in blue represent proteins in FI not included in TCGA's pathway analysis. See text for hypotheses concerning possible roles of NUP50 (left panel) and TNC (right panel). Reprinted under Creative Commons Attribution (CC-BY) license from Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biology, 11*, R53.

Wu et al. then applied cluster analysis techniques to the FI subnetwork, which revealed 17 modules, of which six had four or more nodes (shown in Figure 7, with shading identifying the two largest modules). Module 0 contains proteins found in the cytoplasm and plasma membrane that are mostly involved in signal transduction, whereas Module 1 contains nuclear proteins that are mostly involved in cell cycle, DNA repair, and chromosome maintenance. From "[t]he fact that most of the [glioblatoma] samples have altered genes in both modules" the researchers advance a mechanistic hypothesis: "these two major modules are acting cooperatively in establishing and/or maintaining the [glioblastoma] phenotype, and . . . the development of [glioblastoma] cancers involve malfunctions in both signaling transduction and cell-cycle regulation" (p. 10).

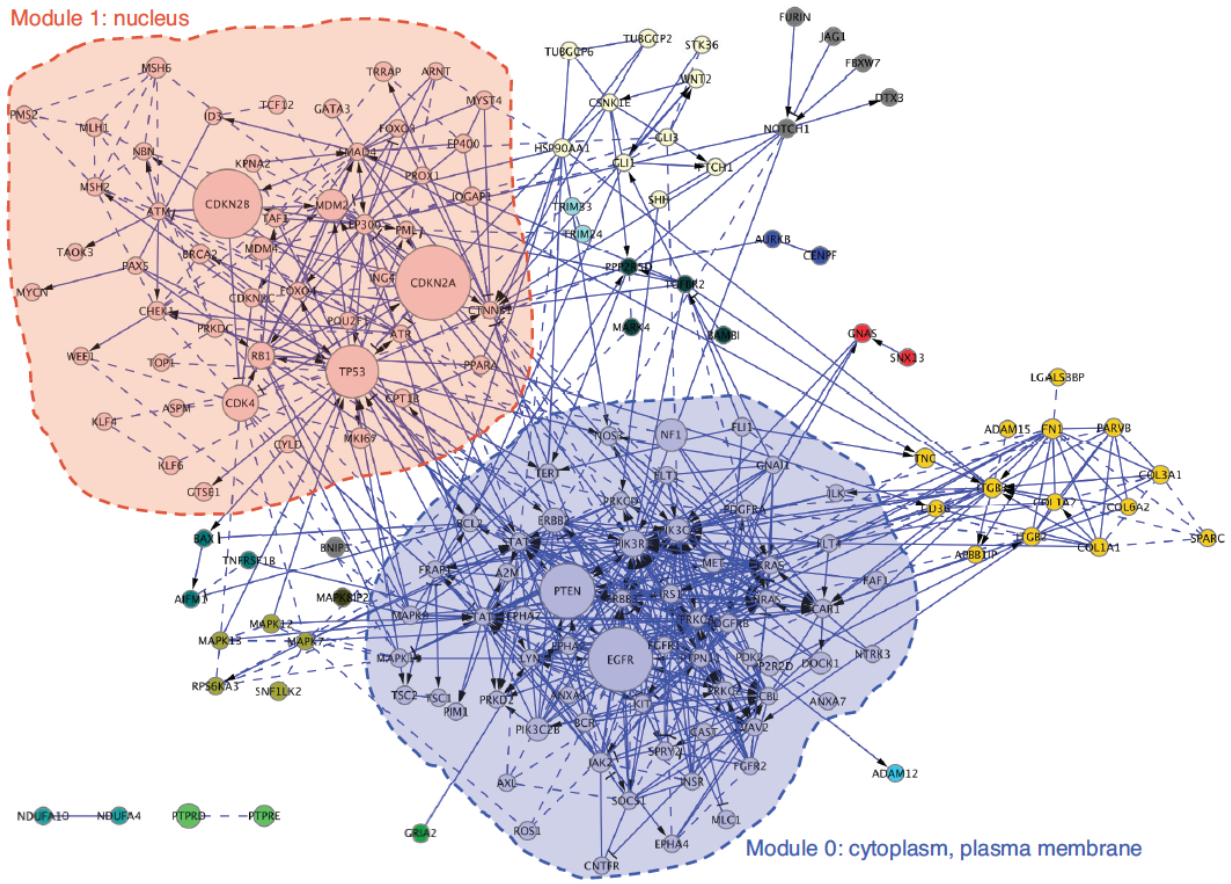


Figure 7. Application of betweenness algorithm reveals six modules in portion of FI network containing mutated or copy-number altered genes in TCGA glioblastoma samples. GO annotations shown for the two largest modules, shown in color. Size of nodes indicates frequency of a given gene being altered in TCGA data. Dashed edges are interactions predicted from model organisms, not empirically confirmed. Reprinted under Creative Commons Attribution (CC-BY) license from Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biology, 11*, R53.

In another approach, Wu et al. started with genes mutated in at least two TCGA samples. By adding the minimum number of genes from FI needed to generate a connected subnetwork containing > 70% of altered genes, they built a network of 77 genes and 5 linker genes. These genes turned out to be far more interconnected, with a much shorter path length between them, than random sets of genes. As shown in Figure 8, when they projected pathway information back onto the core subnetwork, they found four pathways—TP53, focal adhesion, signaling by PDGF, and cell cycle—highly represented in this core subnetwork. Moreover, as the figure indicates, they are highly intertwined, with overlap and cross talk between the pathways. By revealing this, the network research enriched the understanding provided by the pathways alone.

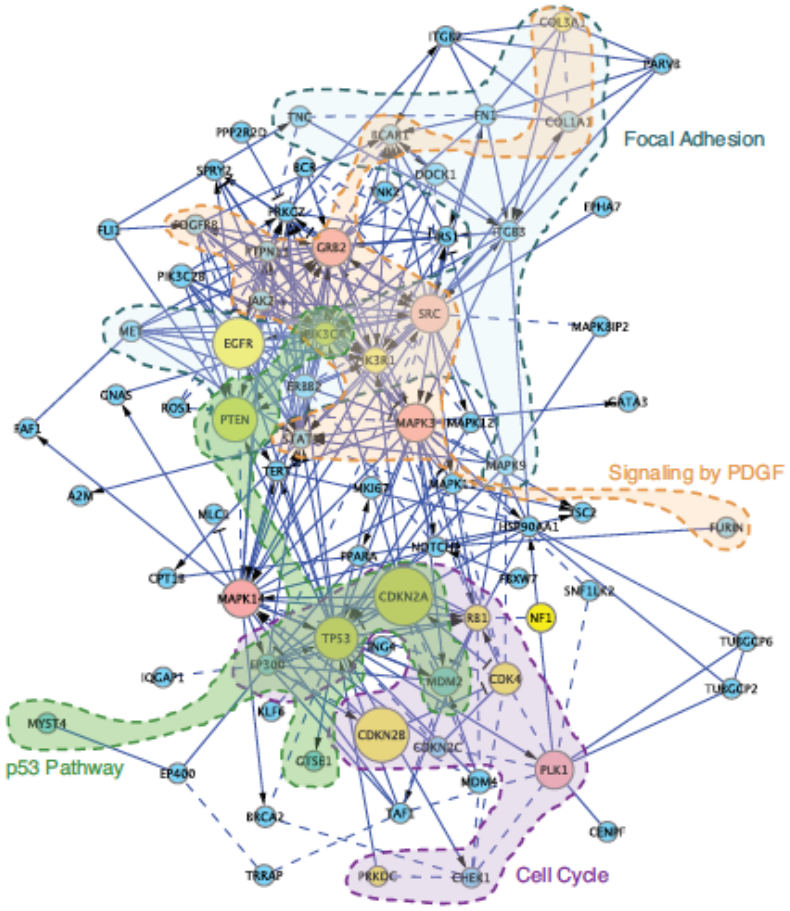


Figure 8. Core subnetwork extracted from TCGA glioblastoma data with identification of genes in four pathways shown in shaded regions. The color of the nodes indicates genes shared in another sample of glioblastoma tumors (yellow) or not shared (blue). Red indicates those nodes added to connect the network. Node size indicates frequency of mutation in TCGA sample. Reprinted under Creative Commons Attribution (CC-BY) license from Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biology, 11*, R53.

Wu et al.'s study illustrates how one can draw upon both pathway strategies and network clustering strategies in developing mechanistic hypotheses about how genes altered in cancer result in hallmarks of cancer. Above I noted that the two heuristic strategies each offered partial but complementary perspectives on a mechanism—identification of clusters in pathways left out specification of pathways, while pathway strategies lack the capacity to include proteins whose specific contribution is unknown. Wu et al.'s success in integrating them offers promise that these approaches will converge and produce robust accounts of possible mechanisms that explain how cancer hallmarks are generated. The ability to link multiple genes altered in tumors with these mechanisms further serves to reduce the heterogeneity problem.

7. Conclusions

Philosophers concerned with mechanistic explanations have focused on heuristic strategies for taking mechanisms apart to identify their components and determine what they do. In this paper I have described two heuristic strategies that work in the opposite direction: they start with components and hypothesize mechanisms. I have shown how cancer researchers are employing these heuristic strategies to address the enormous heterogeneity among genes that are found to be altered in cancer patients. By relating multiple altered genes to the same mechanism, researchers are seeking to explain why any of these alterations results in cancer.

More specifically, I have differentiated two heuristic strategies for advancing from altered genes to higher-level mechanisms in which the proteins coded by these genes function. The first identifies pathways of connected proteins, viewing those pathways as constituting the relevant higher-level control mechanism. This approach requires knowledge of how proteins affect each other—by, for example, transferring phosphate groups from one protein to the next in a signaling pathway. The set of proteins organized into a pathway constitute a mechanism and, when sufficient knowledge is available to generate a pathway, one can view the mechanism as the entity affected by alterations to any genes coding for components of the pathway. The second heuristic strategy starts with data about protein or gene interactions and constructs a network from this data. Clustering algorithms are then invoked to identify groups of genes or proteins that are highly interactive. These are treated as constituting a higher-level mechanism. Unlike the first approach, this strategy identifies proteins as parts of a mechanism without knowing in which specific activities they figure. To apply this strategy, researchers only need evidence that the genes or proteins interact in some way. Once these clusters are identified, researchers can use techniques such as diffusion to identify the mechanism that is likely affected by the alteration of the gene.

Like the heuristic strategies identified by Bechtel and Richardson (1993/2010) and Craver and Darden (2013), the strategies of appealing to pathways and network clusters to identify mechanisms are discovery strategies. They are used to help researchers formulate reasonable hypotheses for further inquiry; they do not show that the hypotheses arrived at are true. These strategies, however, are different from those of more traditional mechanistic research since the goal (to determine which components work together as mechanisms) is different. Along the way, though, they also serve some of the same goals as the traditional heuristics—identifying new parts and operations of mechanisms and how they are organized together to produce specific phenomena. In the case of cancer, the main focus is on how these mechanisms generate the hallmarks of cancer when they are altered so that the mechanism no longer operates in its normal fashion. In this context of the heterogeneity problem, by turning to mechanisms and not just their parts, researchers acquire a way of understanding how multiple different alterations all produce the same cancer hallmarks.

Acknowledgments

I thank Sara Green, Anya Plutynski, Ingo Brigandt, and two anonymous reviewers for this journal for valuable comments and suggestions on earlier versions of this paper. I also thank

Trey Ideker for allowing me to participate in his laboratory meetings in which strategies for using networks to identify mechanisms are often discussed.

References

Annunziata, C. M., Davis, R. E., Demchenko, Y., Bellamy, W., Gabrea, A., et al. (2007). Frequent engagement of the classical and alternative NF-κB pathways by diverse genetic abnormalities in multiple myeloma. *Cancer Cell, 12*, 115-130.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., et al. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics, 25*, 25-29.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T., et al. (2001). BIND - The Biomolecular INTeraction network database. *Nucleic Acids Research, 29*, 242-245.

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer, 91*, 355-358.

Bechtel, W. (1986). Building interlevel pathways: The discovery of the Embden-Meyerhof pathway and the phosphate cycle. In J. Dorn & P. Weingartner (Eds.), *Foundations of Biology* (pp. 65-97). Vienna: Holder-Pichlert-Tempsky.

Bechtel, W. (2017). Using the hierarchy of biological ontologies to identify mechanisms in flat networks. *Biology & Philosophy, 32*, 627-649.

Bechtel, W. (2018). The importance of constraints and control in biological mechanisms: Insights from cancer research. *Philosophy of Science, 85*, 573-593.

Bechtel, W. (2019). Analyzing network models to make discoveries about biological mechanisms. *British Journal for the Philosophy of Science, 70*, 459-484.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421-441.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Bertolaso, M. (2016). *Philosophy of cancer: A dynamic and relational view*. New York, NY: Springer Berlin Heidelberg.

Buchner, E. (1897). Alkoholische Gärung ohne Hefezellen (Vorläufige Mittheilung). *Berichte der deutschen chemischen Gesellschaft, 30*, 117-124.

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature, 455*, 1061-1068.

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

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

Cancer Genome Atlas Research Network. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature, 497*, 67-73.

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics, 45*, 1113-1120.

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology, 3*, 140.

Craver, C. F., & Darden, L. (2001). Discovering mechanisms in neurobiology: The case of spatial memory. In P. K. Machamer, R. Grush & P. McLaughlin (Eds.), *Theory and method in neuroscience* (pp. 112-137). Pittsburgh, PA: University of Pittsburgh Press.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. Chicago: University of Chicago Press.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution*. Cambridge: Cambridge University Press.

Du, D., Roguev, A., Gordon, D. E., Chen, M., Chen, S.-H., et al. (2017). Genetic interaction mapping in mammalian cells using CRISPR interference. *Nature methods, 14*, 577-580.

Ellis, R. W., Defeo, D., Shih, T. Y., Gonda, M. A., Young, H. A., et al. (1981). The p21 *src* genes of Harvey and Kirsten sarcoma viruses originate from divergent members of a family of normal vertebrate genes. *Nature, 292*, 506-511.

Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *British Journal of Cancer, 108*, 479-485.

Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., et al. (2006). Cosmic 2005. *British Journal of Cancer, 94*, 318-322.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., et al. (2004). A census of human cancer genes. *Nature Reviews Cancer, 4*, 177-183.

Garraway, L. A., & Lander, E. S. (2013). Lessons from the cancer genome. *Cell, 153*, 17-37.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford: Oxford University Press.

Green, S. (forthcoming). Is there a "right" level or scale of analysis? Some lessons from cancer research.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*, 153-158.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. *Cell, 100*, 57-70.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*, 646-674.

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based stratification of tumor mutations. *Nature Methods, 10*, 1108-1115.

Hu, P. Z., Bader, G., Wigle, D. A., & Emili, A. (2007). Computational prediction of cancer-gene function. *Nature Reviews Cancer, 7*, 23-34.

Ideker, T. (2016). The real value of an atlas. https://ncip.nci.nih.gov/blog/real-value-atlas/

Keats, J. J., Fonseca, R., Chesi, M., Schop, R., Baker, A., et al. (2007). Promiscuous mutations activate the noncanonical NF-κB pathway in multiple myeloma. *Cancer Cell, 12*, 131-144.

Knudson, A. G., Jr. (1971). Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences, USA, 68*, 820-823.

Kuperstein, I., Bonnet, E., Nguyen, H. A., Cohen, D., Viara, E., et al. (2015). Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis, 4*, e160.

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics, 47*, 106-114.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago: University of Chicago Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Malumbres, M., & Barbacid, M. (2003). *RAS* oncogenes: the first 30 years. *Nature Reviews Cancer, 3*, 459-465.

Morange, M. (1993). The discovery of cellular oncogenes. *History and Philosophy of the Life Sciences, 15*, 45-58.

Morange, M. (1997). From the regulatory vision of cancer to the oncogene paradigm, 1975-1985. *Journal of the History of Biology, 30*, 1-29.

Morange, M. (2001). History of cancer research. *eLS*: John Wiley & Sons, Ltd.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Plutynski, A. (2018). *Explaining cancer: Finding order in disorder*. New York, NY, United States of America: Oxford University Press.

Ross, L. N. (forthcoming). Causal concepts in biology: How pathways differ from mechanisms and why it matters.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*, 2498-2504.

Shen, J. P., Zhao, D., Sasik, R., Luebeck, J., Birmingham, A., et al. (2017). Combinatorial CRISPR-Cas9 screens for *de novo* mapping of genetic interactions. *Nature Methods, 14*, 573-576.

Soto, A. M., & Sonnenschein, C. (2011). The tissue organization field theory of cancer: a testable replacement for the somatic mutation theory. *BioEssays, 33*, 332-340.

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*, 719-724.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., et al. (2013). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep, 3*, 2650.

Thomas, R. K., Baker, A. C., DeBiasi, R. M., Winckler, W., LaFramboise, T., et al. (2007). High-throughput oncogene mutation profiling in human cancer. *Nature Genetics, 39*, 347-351.

Vandin, F., Clay, P., Upfal, E., & Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Symposium on Biocomputing*, 55-66.

Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology, 18*, 507-522.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., & Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *Plos Computational Biology, 6*.

Vogelstein, B., & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine, 10*, 789-799.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr., et al. (2013). Cancer genome landscapes. *Science, 339*, 1546-1558.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., et al. (2007). The genomic landscapes of human breast and colorectal cancers. *Science, 318*, 1108-1113.

Wu, G., Feng, X., & Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. *Genome Biology, 11*, R53.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., et al. (2002). MINT: a Molecular INTeraction database. *FEBS Letters, 513*, 135-140.