UNCOVERING THE HIDDEN MECHANISMS OF CANCER TO IMPROVE PERSONALIZED PATIENT CARE

Naif M. Zaman

Department of Anatomy and Cell Biology

McGill University, Montreal

Quebec, Canada

May 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctorate in Philosophy

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am forever thankful to have had the opportunity to pursue my PhD studies from one of the top educational institutions in the world. The experience has engrained in me the ability to think critically in both my personal life and as a scientist. While the journey was anything but easy, in retrospect, I would still choose to do it all over again.

I want to thank my supervisors Dr. Edwin Wang and Dr. Andre Nantel. They have retaught me the valuable lessons of hard-work and finishing what I have started. At times they would hold my hand and other times feed me to the sharks. Both were helpful. During my studies, the most memorable day was one Saturday when Edwin and I were working at the office, a few weeks prior to my first paper submission. Edwin asked me, "Do you want to be successful?!" Of course, I answered "yes." However, so much more was spoken with that one question that it would go on to define who I am as a person. It created a deep yearning for something greater in life.

I want to thank my committee members Dr. Craig Mandato, Dr. Isabelle Rouiller and Dr. Nathalie Lamarche. Dr. Rouiller helped me improve my presentation skills to better explain my research to a non-bioinformatic audience. Dr. Lamarche was always very polite and gave me constructive criticism and showed great enthusiasm towards my research. And, I would really like to thank Dr. Mandato for always rooting for me

and going out of his way to help me, despite his busy schedule as the Chair, during my excruciating times.

I want to thank all of my current and past colleagues at the NRC. However, a special thanks must be given to Chabane Tibiche, M.Sc. He is one of the most patient person I have ever had the pleasure of meeting. I would have never learned to program or pursue my studies if it wasn't for him. He is diligent and meticulous about his work and some of that defiantly rubbed off on me.

I would also like to thank my best friends and core think-tank, Ammar Ahmed M.B.A., Vinh Jutras C.F.A. A.S.A, Senthuran Tharmalingam M.D., Ashwin Dixit M.D., Zubair Sattar, Hamza Bari, Arif Awan M.D., Soham Rej M.D., Charles Taylor C.F.A, Mohamed Saad, Matthieu Afan, and Hesam Azimzadeh M.B.A. Not many people are blessed with so many close friends from such a diverse background that they can meet every night—it's true. I believe the ideas and dreams we exchanged verbally will all come to fruition someday.

Most importantly, I would like to thank my parents Ayesha Zaman and Mukhles Zaman, and my sister Sumaiya Zaman. Without their support and inspirations, I would have never gotten this far in my life. They have left their country to build a better future for me, and with my hard-work, I want to give them a worry-free retirement.

Someday, I hope to make everyone proud.

ABSTRACT

The aim of my work was focused on improving personalized cancer patient care. Using a systems biology approach, I was able to uncover underlying mechanisms of cancer, identify better drug targets, and predict future mutations that may cause cancer. I have acquired skillsets to integrate various types of large datasets, as well as develop algorithms, to build predictive models.

For my first major project, I constructed an integrated network combining genetic screening (RNAi screening) and genetic alteration (copy number variation (CNV) and mutation) data to identify subtype-specific drug target for luminal and basal breast cancer subtypes (Cell Reports, 2013).

I then devised an innovative algorithm that can personalize the prediction of key new mutations that could transform a normal cell into a cancer cell for a given cancer type. The algorithm integrates network propagation to transform discreet mutation data into a continuous form, upon which further modeling is done. This is the first work of its kind, as it not only integrates mutation information in a meaningful way, but is also able to predict new mutations that could cause cancer in the context of a person's existing non-cancer causing mutations. The algorithm is applied to luminal breast cancer and ovarian cancer (manuscript completed, 2016).

RÉSUMÉ

Le but de mon travail a été axé sur l'amélioration des soins personnalisés aux patients souffrant du cancer. En utilisant une approche de biologie des systèmes, je suis en mesure de découvrir les mécanismes sous-jacents du cancer, d'identifier de meilleures cibles de médicaments, et de prédire les mutations futures qui peuvent causer le cancer. J'ai acquis un capacité à intégrer différents types de grands ensembles de données, ainsi que de développer des algorithmes apte à construire des modèles prédictifs.

J'ai construit un réseau intégré combinant le dépistage génétique (dépistage ARNi) et la modification génétique (nombre de copies de variation (CNV) et mutations) des données pour identifier la cible de médicament spécifique aux sous-typex pour les cancers du sein de type luminaux ou basaux (Rapports cellulaires, 2013).

J'ai ensuite dévelopé un algorithme innovant qui peut personnaliser la prédiction de nouvelles mutations clés nécéssaires à la transformation d'une cellule normale en une cellule cancéreuse. L'algorithme intègre la méthode de propagation des réseaux pour transformer les données de mutations discrètes en une forme continue, ce qui permet leur modélisation pas des algorithmes normalement appliqués sur des données transcriptomiques. Ceci est le premier ouvrage de ce genre, car il intègre non seulement des informations de mutations d'une manière significative, mais est aussi capable de prédire de nouvelles mutations qui pourraient causer le cancer dans le cadre

des mutations pré-existantes dans le génome du patient. L'algorithme est appliqué au cancer du sein de type luminal et au cancer de l'ovaire (manuscrit achevé, 2016).

CHAPTER 1

In Biology, like in many other fields, advances in knowledgebase often come in phases, with breakthroughs decades or even centuries apart. In the eighteenth century, Carl Linnaeus formalized a uniform system of naming and classifying organisms[1]. Then, in the nineteenth century, the cell theory was proposed, usually credited to Theodor Schwann and Matthias Schleiden[2], which led to the discovery of the cellular components that make up the cell such as chromosomes, mitochondria, the nucleus, etc. In the twentieth century, the works of Hans Krebs and Carl Cori[3] ushered the biochemistry era by their identification of many key metabolic pathways, enzymes and the series of sequential reactions that produce or degrade the cellular metabolites. Finally, in 1953, the landmark Nature paper[4] by James Watson, Francis Crick on the structure of DNA is seen as the beginning of the molecular biology era and the establishment of its central dogma which defines the relationship of informational transfer between DNA, RNA and proteins. Since then, molecular biology has dominated nearly every aspect of biological research. From the proposition of the cell theory, to the cell being broken down into its molecular and cellular components, and finally, to the explosion of -omics information in the twenty-first century, biological

research has gone through many phases characterized by periods of component identification, determining how these components interact with each other and, finally, understanding the rules that govern these complex interconnected systems.

With every new breakthrough, a little more was discovered and new challenges were met. The mid-twentieth century was focused on the identification of different proteins and their functions. New techniques like gel-electrophoresis, for example, facilitated the ability to separate and identify different proteins based on their molecular weight. Next, the interest shifted to understanding how the different constituent parts of the cell interacted with each other. This led to the construction of simple models such as the MAPK kinase cascade or the EGFR signalling pathway, and how various proteins and molecules in the cell interact in relationship to particular functions and phenotypes. These earlier findings, which mainly rely on linear logics, involve a small number of elements and short chains. Ultimately, proper understanding of these pathways requires existing knowledge to be mathematically formalized.

At the turn of this century, another paradigm shift is taking place in biology, from a reductionist approach to a holistic approach, which proposes that the component parts alone cannot explain all of the inter and that intra-cellular systems that may be at play[5,6]. The reductionist approach that enabled us to understand the many intricate parts of the cell to date, often cannot capture the properties of the biological system as a whole. Traditionally, biologists have tried to break down a complex system into simpler

or fundamental parts that can be more easily understood and validated. However, the cell is a complex dynamic system with tens of thousands of protein interactions and crosstalk between hundreds of dynamic pathways.

1.1 Cancer

The rise in cancer rates is one of the biggest challenges that faces biology in the twenty-first century. It is characterized by abnormal growth and near immortalization of normal cells often followed by their ability to migrate or metastasize to form secondary tumours. It is a leading cause of death in the U.S. and other developed countries, indiscriminately affecting people from a wide range of age, race and economic background[7]. There are over 14.2 million new cases and 8.2 million deaths due to cancer in the United States each year[8]. In Canada there is an estimated 274,000 new cases and 78,000 deaths annually[9]. One of the main reasons why the cancer rate is rising is simply because people are living longer. As life expectancy increases, so do the accumulation of DNA mutations which increase the risks of cell transformation. However, many environmental and societal factors have also been linked to the development of cancer including tobacco use, sun and UV exposure, inadequate diet and physical activity, obesity, alcohol consumption, and environmental carcinogens[10].

The current trend in the U.S., as of 2016, reveals that nearly half of its population could suffer from some form of cancer at least once during their lifetime[7].

While the earliest description of cancer can be dated back to the days of the pharaohs about 3,000 B.C.[11], Hippocrates, a Greek physician, used the term "carcinoma" to describe ulcer-forming tumors. However, it was the Roman physician Celsus, who used the term "cancer," the Latin word for crab, to describe the finger-like spreading of the disease[11]. Hippocrates proposed the "humoral theory," where the body was composed of four fluids: black bile, yellow bile, blood and phlegm and, for the next 1,300 years, it was believed that an excess of one of the humors was the cause of cancer. Around the 1700s, Stahl and Hoffman proposed the "lymph theory," which theorized that cancer was composed of fermenting and degenerating lymph. The theory gained rapid support when John Hunter, the father of surgical oncology, agreed. Hunter reasoned that, because surgically removing tumors did not cure patients, the cancerous fluid is perhaps being replenished by the lymph, like a sap in a tree. In 1838, German pathologist Johannes Muller demonstrated that cancers are derived from normal cells, not lymph.

During the 19th century, although it was becoming clearer that cancer originated from the host's cells, whether they arose spontaneously or as a result of environmental influence remained unclear. In 1890, William Russell identified intracellular particles that were ubiquitously found in cancerous tissues[12]. These particles were later termed

"Russell bodies," which are aggregated immunoglobulins[13]. In the early 20th century, Glover and Livingston identified bacterium in a wide variety of cancers[14]. Although, their work was refuted at their time, in 1984, Warren and Marshall identified a spiral-shaped bacterium, *Helicobacter Pylori*, in gastric ulcers that often transformed into cancers[15]. They were awarded with the Nobel Prize in 2005 for establishing the first causal link between a bacteria and cancer. Others have linked viruses as potential culprits as well[16,17]. For example, human papillomavirus (HPV) has been linked to uterine cancer.

Although, tumors could arise from infectious agents like virus or bacterium, by 1975, scientists have suggested family history and exposure to mutagens could also be associated with cancer development. In 1976, Stehelin, Bishop and Varmus showed that somatic mutations of oncogenes were the dominant force behind cancer[18].

In the 1980s, Vogelstein and colleagues showed that cancer may arise from a single lineage of cells and that mature colon cancers contain more somatic alterations that earlier cancers. This suggested that mature tumors require additional genetic alterations to form benign tumors. Simultaneously, pathologists have long observed and identified tumors in various "stages" during their histological examination of tumors. These observations indicated that, for a normal cell to become cancerous, it required additional somatic alterations. However, which genetic alterations are

required, which pathways are involved and the extent all cancerous cells are, remained elusive.

1.2 Genetic aberrations

One of the more recent focus of cancer research aims at identifying the genetic alterations that could lead to the development of malignant cancerous cells from normal ones. There is a wide variety of alterations. For example, in melanomas and lung cancers, the dominant lesion are point mutations due to the effects of ultraviolet rays and smoking, respectively[19]. Other environmental factors, like exposure to aflatoxin, have also been linked to mutagenesis in liver cancer but, collectively, environmental factors cannot fully explain the diversity of mutations in tumors. When studying families with Lynch syndrome, mutations in DNA mismatch repair enzymes[20-22] were observed. This suggested how somatic mutations could accumulate in cancer cells to such a high degree. A recent survey identified more than 20 different mutational signatures that may be active in tumors[23]. For example, mutations in DNA polymerase result in a high number of mutations in subsets of endometrial and colorectal cancers[24]. A mutational signature describes the location of the mutation not the nucleotide context or how it may effect downstream processes. This is why the impact of the vast majority of mutations remains unexplained.

Another type of genomic lesion that is observed in cancerous cells are structural alterations or genomic rearrangements of large chromosomal regions. Sometimes these alterations are found as an increase in the number of copies of parts of the genome. Recently, larger structural rearrangements resulting from chromosomal shattering were observed in a phenomenon called "chromothripsis"[25]. Cancer genomes also exhibit gain and loss of genetic material (i.e. amplification and deletion, respectively) that are responsible for encoding genes whose protein products may be altered and affect the survival of cancer cells. These copy number alterations (CNA) are observed more frequently in cancer types that typically have low mutation rates[26]. For instance, in ovarian cancer, TP53 have been shown to be mutated across 96% of ovarian cancer patients and no other mutations are as frequently mutated[27]. However, ovarian cancer has an average of 12.3 focal CNAs involving recurrently altered regions, and is thus classified as a 'C class' tumor (i.e. CNA is driving cancer)[26]. Genes frequently involved in CNAs have become effective drug targets[28,29]. However, CNAs affect multiple genes at a time, hence, deciphering the ones that are responsible for driving the cancer process remain an active area of research.

1.3 Cancer hallmarks

Despite all of these innovations, the task to integrate this information in a cohesive manner that can explain cancer phenotypes remains difficult. For example, at first glance, genomic alterations (e.g. mutations) seem random as there is very little overlap when one compares the mutations sites between patients with the same cancer type[30-32]. In addition, aneuploidy, an abnormal number of certain chromosomes, has also been shown to present in tumor cells[33]. These alterations in cancer cell genomes results in the acquisition of particular phenotypical characteristics, called "cancer hallmarks", which distinguish them from normal cells[34]. These cancer hallmarks include the cancer cell's ability to resist cell death, induce angiogenesis, become nearly immortal, evade growth suppressors which allows them to grow indefinitely and sustain proliferative signaling, and, finally, to invade other parts of the body by metastasizing[34].

However, when evaluating cancer patients, it is difficult to measure these hallmarks in a meaningful way that would allow oncologists to quantitate disease progression. There are few gene signatures or mutational patterns that could be used as precise biomarkers to indicate how each of the hallmarks are changing. For example, mutations in the prototypical tumor suppressor TP53 often lead to the breakdown of the apoptotic regulatory mechanisms that can halt further cell-cycle progression of a damaged cell. Alternatively, the combinatorial effect of other genomic alterations could

also induce the cell's loss of proper apoptotic regulation. This is one of the reasons why it is difficult to identify one mutation as a "universal" culprit because there are always other mutations that can accomplish the same result. This also suggest that, while genomic alterations may appear to be random, it is their combined effect on the cancer hallmarks through various cell signaling pathways that allows a cancerous cell to thrive. In addition, there is the presence of passenger mutations which add an additional level of complexity. Passenger mutations do not have a direct effect on carcinogenesis but are present at an elevated rate since the DNA repair mechanism apparatus is often impaired in cancer cells.

In conclusion, recent discoveries have demonstrated that cancer cells exhibit particular phenotypical characteristics, called "cancer hallmarks", which distinguish them from normal cells[34]. The underlying genotypic mechanisms however, remain complex and difficult to fully understand.

1.4 A systems biology approach to the study of cancer

Early cancer research focused on identifying individual genes that are essential for cell transformation[5,31,35]. This approach however was confined to a small pool of pre-defined, well-studied pathways or protein-to-protein interactions. Unfortunately, there are multiple mechanisms that can lead to carcinogenesis and matching the right drug

with the right patient remains difficult. In addition, secondary redundant pathways often allow for the build-up of resistance to initial anti-cancer treatments thus rendering them ineffective[36].

Technological advances over the past three decades have uncovered a vast amount of genomic information about the mechanisms that underlie carcinogenesis. In the mid-1990s the maturation of DNA microarray technology, which allowed for the global quantification of mRNA levels, allowed researchers to simultaneously associate large portions of the genome with pathology[37]. This technology has fundamentally transformed the way research is done and uncovered many noteworthy discoveries[38-40]. In breast cancer, gene expression profiles helped identify three specific subtypes—luminal, HER2, and basal—with distinct clinical characteristics[41]. HER2 subtype expresses HER2 (HER2+), but not the estrogen (ER-) or progesterone receptors (PR-). Basal, or triple negative, tumors do not express estrogen, progesterone or HER2 (ER-, PR-, and HER2-). More recently, the gene expression signature of 50 genes—PAM50[42]—was shown to be capable of predicting prognosis. Oncotype DX is another gene expression derived profile—21 genes—that can help physicians predict the likelihood of recurrence in ER+ breast cancer patients and how they may react to treatment and determine the best course of treatment[43].

More recently, next-generation sequencing (NGS) technologies have added a wide variety of methodologies in which to query the state of the cancer genome. For

example, it is now possible to globally identify point mutations, amplifications, deletions and variant allele frequencies in whole populations or even in individual cells. NGS allowed for the identification of particular mutations linked to specific cancers, even though the majority of these mutations have low frequency in any cancer type. For example, from the TCGA dataset of ovarian patients, TP53 has been shown to be mutated across 96% of its samples[44]. In melanomas, hotspot mutations of BRAF, NRAS and NF1 are present in distinct subgroups and are not generally co-mutated in the TCGA samples[45].

Unfortunately, these clear markers are the exception rather than the rule. Genetic heterogeneity is common between different types of tumors and even between patients with the same tumor types, which makes it extremely difficult to identify common cancer driver genes. For example, mutation frequencies across patients are often less than 2% which makes it difficult to identify most patient sub-groups unless 5,000 samples are sequenced for each cancer type. Currently this is not feasible, but alternatively, it may suggest that cancer could be caused by various combinations of genetic alterations. To further complicate the matter, recent NGS studies have shown that there is also intra-tumor heterogeneity[38,39]. Cutting a tumour sample into multiple parts, and measuring their respective gene expression profiles and mutation sites, have shown fewer similarities than expected[39]. Sequencing analysis also showed that tumours contain multiple subpopulations (clones) and that the greatest tumour volume

often consists of a dominant clone that out competed the others. This is one of the reasons why it is difficult to treat cancer with a single drug. While the drug treatment may eliminate one clone, it provides a competing growth advantage to the other more-resistant clones, and the patient relapses[31].

The details of how genetic alterations impact cancerous phenotypes are still being investigated as we are still in the early days of interpreting our own genetic information. There is a gap that needs to be bridged between what is discovered from sequencing information and how it is interpreted. There is about a decade and a half of data from which to draw conclusions from. So, while there may be only 2% of the population that may benefit from knowledge of the mutations associated with their "unique" cancer, there may be another 20% of the population with a different mutation but that show no sign of disease because it may not be linked to cancer. Fortunately, a systems biology approach that combines empirical, mathematical, and computational techniques is driven by the increase of dataset sizes, not hampered by it. High-throughput technologies have suggested that researchers should consider a "multiple gene model" rather than a "one gene model" since molecules do not work alone but as a system with a collection of interactions.

A systems biology approach to study cancer's inherent complexity could also resolve some of the issues in cancer research and treatment. This holistic approach studies the causal relations and complex interplays between the different components

of the biological system. Its strength lies in its ability to integrate various types of information to explain the emerging properties of a system. In the past decade alone, advances in genome-wide high-throughput technologies have created an exorbitant amount of data[35]. TCGA (the cancer genome atlas), and other consortiums alike, have sequenced over 30 cancer types from thousands of tumor samples to uncover the genetic alterations that drive carcinogenesis. Ideally, their objective was to identify a few somatic mutations per cancer type that could explain the driving force behind the disease in the population. It was hoped that, with the right methodologies, these extensive datasets could be harnessed to better understand the complex systems that control cancer development.

1.5 Integration of high-throughput data

The integration of various high-content datasets is fundamental to a systems biology approach. It allows for the modeling of the cell at a macro-level to be able to elucidate the genetic alterations' and interactions' impact on phenotypes. At the very core of every analysis is comparing and contrasting between a phenotype of interest versus the null. The objective is to find a signal that has a statistically significant difference between two groups. Gene set enrichment analysis (GSEA)[46,47], for example, is a popular computational method which uses predefined functional groups or gene

expression data to identify pathways that are significantly altered between two phenotypes (e.g. cancer versus normal). There are other algorithms that do similar analysis as GSEA. While some of the earlier studies have focused on transcriptional profiling data, new technologies have allowed for new dimensions to be discovered.

The advent of NGS has revolutionized the cancer research field. It allows researchers to address an increasingly diverse range of approaches to understand how genetics changes impacts complex phenotypes. This was possible due to the continuous reductions in sequencing cost of per genome. The human genome project, set up in the early 90s and completed around 2003, took over a decade to complete and cost about $2.7 billion[48]. An individual's genome can now be sequenced to a depth of 30-fold for around $1,000—exceeding Moore's Law, which would expect prices to drop by 50% every two years. Unlike previous methods like Sanger sequencing that relied on bacterial cloning and/or PCR, NGS's ability to sequence complex mixtures of amplified DNA is faster and less expensive. Variations of this technology allows it to decipher complex and repetitive genomes[49], determine the abundance of mRNA and non-coding RNAs and identify epigenetic sites of DNA and histone modifications[50].

The global identification of genes essential for cancer cell survival is one of the most important research topics in the cancer field. Advances made in the regulation of gene expression, mRNA degradation and translation have allowed for the systematic perturbation of cellular processes and measurement of their effects. This enables

scientists to associate genotype with phenotype. To silence the expression of a gene and study the effect, RNA-interference (RNAi) was the first technology to be widely utilized[51]. Libraries of siRNAs or shRNAs are used to supress the expression of a gene and assess its impact on the phenotype of interest. The key difference between siRNAs and shRNAs is that siRNA screening require multiwell tissue culture plates for reverse transfection into cells, whereas shRNAs use plasmids that integrate into the cell's genome which can then constitutively express shRNAs for gene silencing[52]. shRNAs can be used to infect cells, which are then split into two groups, one subject to treatment and the other serves as control.

Recently, RNAi screening has been performed on the entire genome to produce a new kind of data[53] that can be harnessed for functional relationships. In this exhaustive study, half of the genome's known genes have been knocked-down across 72 cancer cell lines to determine their respective contributions to cell survival[53]. This new kind of dataset allows us to infer what makes certain genes essential versus others in a cancer cell and which genes collaborate with each other—key questions in cancer research.

The CRISPR-CAS9 endonuclease technology is the latest "craze" in whole genome screening methodologies and addresses some of the disadvantages of shRNA screens. CRISPR is a repeat structure of 29-nucleotide repeats in Escherichia coli, discovered in 1987 by Ishino et al[54]. Coupled with the CAS9 endonuclease, CRISPR guide RNAs can cleave dsDNA complementary to the guide sequence. Repair by the

non-homologous end-joining system often leaves indels that inactivate the gene. Earlier genome editing techniques like Zinc-finger-nucleases[55] and TALENs[56,57] required cloning and optimization, whereas CRISPR guide RNAs is much simpler. CRISPR screens in human cells were first performed in 2014[58,59]. Unlike shRNAs screens that could suffer from off-target effects, CRISPR exhibit greater reproducibility between replicates[58]. More importantly, whereas shRNAs supress gene expression, CRISPR-based methods can knockout a gene, activate its expression or even change the epigenetic environment. The impact of these methods on functional genomics and biology as a whole will be extremely profound.

Finally, it is important to consider that cell lines may not always be the best representation of a cancerous cell from a patient's tumor. Cell lines have accumulated additional genetic alterations over time, they're cultured in liquid media and plastic containers, an environment that is very different from the human body, especially since they will have no interactions with the immune system and the stroma.

1.6 Protein-protein interactome

Protein-protein and protein-DNA interactions are the building blocks of signalling pathways and the regulation of gene expression[60,61]. These interactions have been used to gain biological insights at the level of the individual gene and up to the

global properties of the cell[62,63]. To identify protein interactions, yeast two-hybrid and affinity purification mass spectrometry are often used[64]. Many publicly available databases of protein-protein interactions are available. The IntAct molecular interaction database[65], the Database of Interacting Proteins (DIP)[66], the Molecular Interactions database (MINT)[67] and the Biological General Repository for Interaction Database (BioGRID)[68] are some examples of ongoing efforts by independent groups to curate interactions. The International Molecular Exchange Consortium (IMEx) was recently formed to unify curation rules and so to avoid redundancy[69]. There are additional databases with focuses on signaling and metabolic pathways, like KEGG PATHWAY[70,71] and Reactome[72]. Other databases include Agile Protein Interaction DataAnalyzer (APID)[73], the Michigan Molecular Interactions database (MiMI)[74] and the United Human Interactome database (UniHI)[75]. While these experimental efforts have compiled a large number of interactions, we are still far from a complete mapping of all possible protein-protein interactions.

1.7 Graph theory

With different types of high-throughput data, integrating them together to capture the big picture is at the core of systems biology. No one data type is sufficient to uncover the intricacies involved in cancer development. A gene may be mutated or

amplified, but if it is not expressed then perhaps its impact on disease progression is not significant. Also, if a gene product is a drug target, resistance can become an issue if an alternative gene or pathway can be activated to replace its role. Identifying these potentially alternative pathways is a challenge.

An effective strategy to integrate data from these inter-disciplinary methods is through the application of graph theory and networks. It is a mathematical representation of all of the protein-to-protein or functional interactions in the cell. The interactions (i.e. edges) between gene and gene products (i.e. nodes) that make up the networks are built from literature mining, manual curation from literature, and high-throughput protein-protein interaction studies and datasets (e.g. KEGG)[70,71]. Sub-networks can be constructed focusing on hallmark genes, modulated genes, cancer essential genes, time-course information, and so on, which can identify relationships between genes and proteins under specific conditions.

These networks illustrate the plasticity and redundancy that is inherent in signalling pathways. As an illustration when driving from home to work there may be different routes that you might choose from. However, people tend to choose the paths that take the least amount of time—the most efficient. If there is heavy congestion because of an accident, traffic will automatically be rerouted to the next quickest route based on available knowledge and information. The interactions among genes and proteins can be viewed in a similar fashion. To signal from one gene to another, the cell

can use the most efficient paths (i.e. shortest paths) and, if there is a blockage (e.g. a gene product is knocked-down by a drug), it will reroute the signal to the next most efficient path. The cell has different redundant paths that can serve as backups, whose activation may, for example, result in drug resistance. One of the strengths of a network approach is its ability to help identify novel pathways and cross-talks between known pathways.

A more detailed examination of a network may help us identify important genes through the analysis of its topology—hubs and clusters. Hubs are nodes with a high number of interactions, and clusters are a group of genes that have a higher number of interactions between them than a group of genes would be expected to have by random selection. Hubs aid in identifying important genes because they exhibit a high number of interactions with other genes. Genetic or functional alterations in hub genes (or their products) are often more detrimental to the cell's function than a more secluded gene. Many different pathways can use a hub gene as a mean of cross-talk between other genes or pathways thus making these useful as drug targets. Clusters help to group genes that have high interactions, thus indicating that they may be functionally relevant. In Chapter 2 we use network hub genes to identify breast cancer subtype specific drug targets.

1.8 Network propagation

Network propagation[76] through a protein interaction network is one possible method to transform discreet mutational data into a continuous form, which greatly increases the power of analytical algorithms that can be used to evaluate the functional impact of point mutations. This is similar to the random-walk algorithm, where, given an initial position on a network, it calculates the probability of landing on another network node randomly and spreading part of its score every node in the network. While mutations may be sparse, network propagation helps transform discreet mutation data into a continuous form, and this could help us identify regions or nodes of the network that have a higher score. Google's PageRank, Amazon, Facebook, Netflix and other companies alike, use these kinds of algorithms to suggest webpages, items to purchase, friends to add, and movies to watch based on a user's previous activities. In our second publication, we have shown how the implementation of this method can help to improve prognostic patient care on a personal level.

1.9 Application to cancer therapies

Medical treatment for cancer mainly relies on nonselective cytotoxicity, namely radiotherapy and chemotherapy. Radiotherapy emerged during the late 1800s when x-rays were discovered and had shown the ability to kill cancer cells. Toxic chemicals (i.e.

chemotherapy) were the next modality used to treat cancer patients. These toxic alkylating agents spawned from the use of mustard gas during World War I. Although, radiotherapy and chemotherapy have been widely used by modern science for nearly a century, its nonselective nature has hastened the deaths of many. They reduce the quality of life for those who survive and sometime leave them irreparably harmed. More importantly, a very small subset of patients respond, and when they do, the contribution to 5-year survival in adults is estimated to be less than 3%[77]. For this reason, current drug-discovery strategies are focused on therapies that are capable of differentiating between healthy and cancerous cells.

Therapies that have specificity have been the most successful. These newer kinds of therapies can attack cancer cells while doing little damage to normal cells. Many of these are monoclonal antibodies and small molecules often recognizable by names ending with "-mab" and "-ib", respectively. Breast and prostate cancer, which are influenced by the signaling of sex hormones, respond well to these kinds of therapies. For example, Tamoxifen, which binds and inhibits the transcriptional activity of estrogen receptors are used for patients with breast cancer, and androgen deprivation therapy is used for prostate cancer. Herceptin, also known as trastuzumab, is a monoclonal antibody that interferes with the HER2/neu receptor, and is used as a targeted therapy for breast cancer patients who are HER2-positive. Another new form of therapy that is showing some promise is immune therapy. This type of treatment

takes advantage of cell surface markers that are present in cancerous cells but not in normal cells. One such therapy is Chimeric Antigen Receptor T-cell therapy (CAR-T), where the autologous T-cells are infected with an immune-stimulating receptors that bind to CD19 of B-cells[78-80]. These kinds of targeted therapies are at the heart of personalized medicine.

THESIS ORGANIZATION

This thesis will explore how a systems biology approach is used in cancer research. The work in Chapter 2, deals with identifying subtype-specific drug targets for breast cancer. In silico, we developed a methodology that can help pinpoint drug targets for luminal and basal subtypes with >80% accuracy. This is the first work to integrate RNAi screening data in predictive analysis; included also are sequencing data, gene expression profiles, copy number variations, and protein interaction networks. These findings have been validated in both TCGA patient samples and drug screening data. Chapter 3 represents the first work of its kind. It takes on the daunting task of using NGS data to predict somatic mutations required to develop cancer, given an individual's germline variants. This algorithm is personalized, as patients do not have to fit into predefined subgroups of any kind. For proof of concept, we have applied it to ovarian and luminal breast cancer, and have had similar results. In Chapter 4, we discuss future directions that should be taken to improve cancer patient care.

SCIENTIFIC CONTRIBUTIONS

Highlighted here are the eight publications that I co-authored during my PhD, with one more manuscript that will be submitted. My specific contributions to each projects are indicated.

1. Zaman N, Li L, Jaramillo ML, Sun Z, Tibiche C, Banville M, Collins C, Trifiro M, Paliouras M, Nantel A, O'Connor-McCourt M, Wang E. Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. Cell Rep. 2013 Oct 17;5(1):216-23. doi:10.1016/j.celrep.2013.08.028. Epub 2013 Sep 26. PubMed PMID: 24075989.
    - Zaman N: results and figures
    - Li L: data generation and figures
    - Tibiche C, Banville M, Collins C, Trifiro M, Paliouras M, Nantel A, O'Connor-McCourt M, and Wang E: edit the manuscript.
        a. Integration of networks with multiomic and shRNA data identifies cancer genes
        b. Genes switch roles between cancer causing and essential among cancer subtypes

c. Evolutionary convergence and deterministic paths of cancer genomic alterations

d. Subtype-specific networks successfully predicted subtype-specific drug targets

2. Zaman N, Sebestian J, Tibiche C, Nantel A, Wang E. Predicting Key Personalized Cancer Driver-Mutating Genes in Advance Based on Healthy Individuals' Genetic Makeup (Ready for submission)

- Zaman N: results and analysis
- Sebestian J: data generation
- Nantel A and Wang E: edit manuscript

a. Implemented a new pipeline for network propagation.

b. Developed a new innovative algorithm for predicting somatic mutations from germline variants.

3. Zaman N, Giannopoulos PN, Chowdhury S, Bonneil E, Thibault P, Wang E, Trifiro M, Paliouras M. Proteomic-coupled-network analysis of T877A-androgen receptor interactomes can predict clinical prostate cancer outcomes between White (non-Hispanic) and African-American groups. PLoS One. 2014 Nov

19;9(11):e113190. doi: 10.1371/journal.pone.0113190. eCollection 2014. PubMed PMID: 25409505; PubMed Central PMCID: PMC4237393.

a. Used a network approach and integration of different types of datasets.
b. Different outcomes between white and African-American people.
c. Although I am the first author, this publication was not chosen to be one of the chapters of this thesis because the work was off tangent to the focus of the thesis.

4. Paliouras M, Zaman N, Lumbroso R, Kapogeorgakis L, Beitel LK, Wang E, Trifiro M. Dynamic rewiring of the androgen receptor protein interaction network correlates with prostate cancer clinical outcomes. Integr Biol (Camb). 2011 Oct;3(10):1020-32. doi: 10.1039/c1ib00038a. Epub 2011 Sep 7. PubMed PMID:21901193.

    a. Performed all of the bioinformatics analysis.

5. Wang E, Zou J, Zaman N, Beitel LK, Trifiro M, Paliouras M. Cancer systems biology in the genome sequencing era: part 1, dissecting and modeling of tumor clones and their networks. Semin Cancer Biol. 2013 Aug;23(4):279-85. doi: 10.1016/j.semcancer.2013.06.002. Epub 2013 Jun 19. Review. PubMed PMID: 23791722.

a. Created all of the figures and helped to write the manuscript.

6. Wang E, Zou J, Zaman N, Beitel LK, Trifiro M, Paliouras M. Cancer systems biology in the genome sequencing era: part 2, evolutionary dynamics of tumor clonal networks and drug resistance. Semin Cancer Biol. 2013 Aug;23(4):286-92. doi: 10.1016/j.semcancer.2013.06.001. Epub 2013 Jun 18. Review. PubMed PMID: 23792107.

    a. Created all of the figures and helped to write the manuscript.

7. Wang E, Zaman N, Mcgee S, Milanese JS, Masoudi-Nejad A, O'Connor-McCourt M. Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. Semin Cancer Biol. 2015 Feb;30:4-12. doi: 10.1016/j.semcancer.2014.04.002. Epub 2014 Apr 18. Review. PubMed PMID: 24747696.

    a. Created all of the figures and helped to write the manuscript.

8. Schweitzer M, Makhoul S, Paliouras M, Beitel LK, Gottlieb B, Trifiro M, Chowdhury SF, Zaman NM, Wang E, Davis H, Chalifour LE. Characterization of the NPC1L1 gene and proteome from an exceptional responder to ezetimibe.

Atherosclerosis. 2016 Mar;246:78-86. doi:10.1016/j.atherosclerosis.2015.12.032. Epub 2015 Dec 24. PubMed PMID: 26761771.

a. Performed the bioinformatics analysis required for a figure.

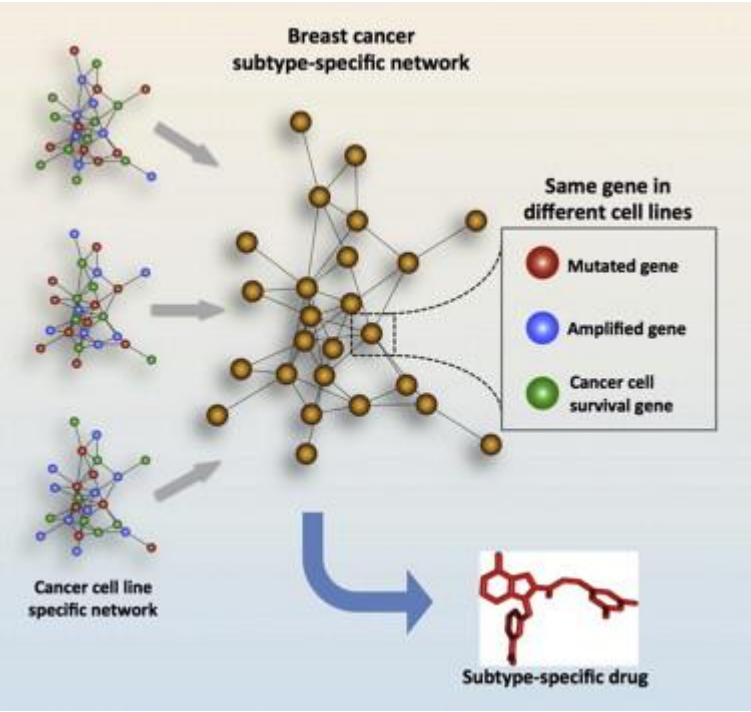
9. Gao S, Tibiche C, Zou J, Zaman N, Trifiro M, O'Connor-McCourt M, Wang E. Identification and Construction of Combinatory Cancer Hallmark-Based Gene Signature Sets to Predict Recurrence and Chemotherapy Benefit in Stage II Colorectal Cancer. JAMA Oncol. 2016 Jan 1;2(1):37-45. doi:10.1001/jamaoncol.2015.3413. PubMed PMID: 26502222.

a. Created all of the figures and helped to write the manuscript.

CHAPTER 2

Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets

Zaman N, Li L, Jaramillo ML, Sun Z, Tibiche C, Banville M, Collins C, Trifiro M, Paliouras M, Nantel A, O'Connor-McCourt M, Wang E. Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. Cell Rep. 2013 Oct 17;5(1):216-23. doi:10.1016/j.celrep.2013.08.028. Epub 2013 Sep 26. PubMed PMID: 24075989.



Graphical abstract

Abstract

Individual cancer cells carry a bewildering number of distinct genomic alterations (e.g., copy number variations and mutations), making it a challenge to uncover genomic-driven mechanisms governing tumorigenesis. Here, we performed exome sequencing on several breast cancer cell lines that represent two subtypes, luminal and basal. We integrated these sequencing data and functional RNAi screening data (for the identification of genes that are essential for cell proliferation and survival) onto a human signaling network. Two subtype-specific networks that potentially represent core-signaling mechanisms underlying tumorigenesis were identified. Within both networks, we found that genes were differentially affected in different cell lines; i.e., in some cell lines a gene was identified through RNAi screening, whereas in others it was genomically altered. Interestingly, we found that highly connected network genes could be used to correctly classify breast tumors into subtypes on the basis of genomic alterations. Further, the networks effectively predicted subtype-specific drug targets, which were experimentally validated.

2.1 Introduction

Thus far, several thousands of tumors representing more than 20 cancer types have been sequenced. These efforts have identified thousands of genomic alterations such as somatic mutations, amplifications, deletions, chromosomal translocations, and gene fusions in each individual cancer genome[44,81,82]. With so many genomic alterations in each tumor genome, it is a big challenge to dissect, prioritize, and uncover the functional importance of the genomic alterations and the underlying mechanisms that drive cancer development, progression, and metastasis[83].

During cancer cell evolution, some genomic alterations become the underlying cause for tumor cell proliferation, fitness, and clonal selection. Cell survival, proliferation, and apoptosis are the most primitive and fundamental cancer hallmarks[34]. Systematic identification of genes that are essential for cell proliferation and survival or cancer-essential genes (i.e., functional screens in which gene knockdown results in cancer cell growth inhibition) by genome-wide RNAi screening has shown that indeed there exist distinct subsets of genes that are selectively required by different cancer cells[84,85]. The mixture of mutations in a tumor prevents linking genes to functions. It could be interesting to dissect and sequence the major clones of tumors or conduct single-cell genome sequencing so that each mutation could be functionally investigated in the cell/clone bearing that mutation. Toward this end, in this study we performed

genome-wide exome sequencing for a panel of breast cancer cell lines and matched their corresponding genome-wide RNAi screening data[53] to perform an integrated network analysis to gain insight into the underlying mechanisms of cancer cell survival and proliferation driven by genomic alterations.

Breast cancers have been classified into three molecular subtypes—luminal, HER2, and basal (basal A and basal B)[41]—using a 50-gene expression signature (PAM50)[42]. The HER2 subtype often has mutated or amplified ERBB2 and has had some degree of clinical success because of the effective therapeutics that can target ERBB2[86]. The luminal subtype is often characterized by the expression of estrogen receptor (ER+), which is not expressed in the basal subtype. The ER+ group (known as luminal breast cancer) has some degree of varying drug response, while triple-negative breast cancers (known as basal-like breast cancer) lack the expression of ER, progesterone receptor, and HER2 and have very limited chemotherapy or other molecularly targeted drug treatment options available. Therefore, we focused on developing integrated networks, composed of both genetic screening (RNAi screening) and genomic alteration (mutation and copy number variations) data, to further characterize the luminal and basal breast cancer subtypes. This approach is likely to generate more insight into the fundamental network wiring in cancer, with the more focused aim of identifying subtype-specific breast cancer genes that may lead to better treatment options in the near future.

2.2 Results and discussion

Genome sequencing of breast cancer cell lines

A genome-wide cell survival RNAi screen has previously been conducted for a panel of luminal and basal breast cancer cell lines[53]. Furthermore, since five lines in the panel have already been exome sequenced[87], we performed exome sequencing of the remaining 11 lines (see Extended Experimental Procedures; Table S1). After removing naturally occurring genetic polymorphisms using the data from the dbSNP database and 1000 Genomes Project (see Extended Experimental Procedures), we identified 3,817 somatic point mutations. Of these tumor-associated genetic alterations, 2,548 were predicted to generate missense mutations (annotated as nonsynonymous mutation by Annovar; http://www.openbioinformatics.org/annovar/), 192 produced nonsense (or stopgain) mutations, 111 mutations were shown to contain an essential splice site, 4 mutations resulted in stop codon readthrough (or stopless) mutations, and 1,073 were synonymous substitutions that would result in silent changes in protein sequence. We also identified, 164 small insertions or deletions (79 and 85, respectively), of which 94 introduced translational frameshifts while 50 were in-frame single-nucleotide variants (SNVs), 5 were stopgain SNVs, and 1 was a stoploss SNV (Table S2). Based on the Annovar program, which predicts potential functional mutations, we obtained 1,630 potential driver-mutating genes (i.e., cancer-causing genes) for all 11 cell lines (Table

S3). Mutants of MAP kinase family were found across all of the lines. As expected, mutant TP53 (80%) was associated with basal subtypes. These results are in agreement of the results of genome sequencing of nearly 1,000 breast cancer samples[81,88]. We also compared the driver-mutating genes in this study to those derived from COSMIC database and ~1,000 breast cancer samples mentioned above and found 45 novel driver-mutating genes in at least one cell line. Three genes among them (ZBTB18, TENM4, and TMEM178A; Table S3) are found in two cell lines.

Subtype-specific survival signaling networks highlight the evolutionary convergence of selective genomic alterations

Cells employ signaling pathways and networks to drive biological processes. Genomic alterations in signaling pathways and networks might result in malignant signaling, which then leads to cancer phenotypes. Genome-wide RNAi screening experiments not only uncover cancer essential genes, but also pinpoint genes that are involved in influencing cell proliferation. Knocking down a proliferation-influencing gene will not necessarily lead to cell death, but it will greatly reduce cancer cell growth (see Experimental Procedures; Figure S1). If a gene that is involved in the regulation of proliferation genes is also subject to nonsynonymous genetic alterations (mutations) or amplified, we defined that gene as a "cell-survival-related driving regulator" (called

"driving regulator" in this study) (see Experimental Procedures and Figure S1). Previously, we showed that modeling and perturbing of signaling networks[89-91] and cancer hallmarks[92] provided insight into cancer gene mutations and identifying high-quality cancer biomarkers. To obtain further insight into the underlying mechanism of cancer cell proliferation trigged by cancer genomic alterations, for each cell line we mapped driving regulators and cancer-essential and proliferation-influencing genes onto a manually curated human signaling network (containing ~6,000 genes and 50,000 relations)[90,93,94] to generate integrated cell-line-specific survival networks (Figure 1; see Experimental Procedures). Such a network represents the signaling mechanism for cancer cell survival and proliferation. The gene amplification data processed using GISTIC[95] were obtained from the Cancer Cell Line Encyclopedia (CCLE; http://www.broadinstitute.org/ccle/home). Detailed information for defining cancer-essential and proliferation-influencing genes and driving regulators can be found in Figure S1 and Experimental Procedures.

Highly connected network genes, called hubs, act as global signal integrators or global regulators for multiple signaling pathways[5,96]. To find out whether the driving regulators and essential and proliferation-influencing genes shape the survival networks, we conducted both fuzzy k-mean clustering and hierarchical clustering analyses of the cell lines using the hubs of cancer-essential genes, driving regulators, or both, respectively. In this study, we defined the top 10% of highly connected genes in a

network as hubs. In general, we also tested the hubs using the top 15% as a cutoff in all the analyses and found that both cutoffs generated similar results. As seen in Figures 2A and 2B, the hubs of either driving regulators ($p = 0.12$, fuzzy k-mean clustering and Fisher's test) or essential genes ($p = 1.0$, fuzzy k-mean clustering and Fisher's test) alone were unable to classify the individual cell lines to the luminal and basal subtypes. However, when we combined the hubs of the driving regulators and essential genes, the cell lines were better classified and distinguished into the luminal and basal subtypes (Figure 2C; $p = 0.03$, fuzzy k-mean clustering and Fisher's test). Permutation tests (see Extended Experimental Procedures) showed that significant classification of luminal and basal subtypes by the network hubs couldn't be obtained at random ($p = 9.0 \times 10^{-4}$). These results suggest that although both driving regulators and essential genes are profoundly different between cancer cells (Figure S2), they are complementary and converge to similar survival signaling mechanisms within their respective subtypes. To further explore these observations in detail, we constructed subtype-specific survival networks (see Extended Experimental Procedures). A subtype-specific network contains ~200 genes that appear across ≥50% of a subtype's cell lines. Nearly all the genes (>95%) in a subtype-specific network act as cancer-essential genes in one cell line, but act as driving-regulators in another line (Figure 3). Randomization tests (see Extended Experimental Procedures) showed that the recurrent usage of the genes in luminal and basal subtypes, respectively, is not random ($p < 1.0 \times$

10^{-4}). These network genes are recurrently used by the subtype's lines, suggesting that cancer cells are "addicted" to their respective subtype-specific network for survival and proliferation. A subtype-specific network represents core survival signaling mechanisms that shed light on convergent evolutionary events and provide functional constraints for selecting genomic alterations that could offer a competitive growth advantage for cancer cells. The selective pressure led to the emergence of distinct network hubs in the luminal and basal subtypes (Figures 3A–3C). For example, AKT1, PIK3, and ESR1 are dominantly selected in luminal subtypes, whereas TP53 and SRC are genetically dominant in the basal subtypes.

We explored network modules for the subtype-specific networks using the Gene Ontology-guided Markov cluster (MCL) algorithm (see Extended Experimental Procedures). Three functional modules, where one is centered by CDK1 for cell cycle, one is centered by P53 for apoptosis and genome instability, and another one is centered by growth factors such as EGFR and MAPK pathway components for cell proliferation, were found in the basal-specific networks (Figures 3A and 3B). Two network modules, where one is centered by CDK1/MYC for cell cycle and the other is centered by AKT/PIK3CA growth factors such as MET and MAPK pathway components for cell proliferation and growth, were found in the luminal-specific network (Figure 3C). To further interpret these findings, we conducted pathway enrichment analysis (see Extended Experimental Procedures) for each subtype-specific network using the cancer-

essential genes, proliferation-influencing genes, and driving regulators. Signaling pathways of cell-cycle, apoptosis, MAPK/growth factors (i.e., MET), and transcription processes were found in both luminal and basal lines, highlighting the fact that these cancer subtypes share core survival pathways commonly used by breast cancers (Table S4). In addition, cancer cells of the basal subtype (basal A and basal B) share the signaling pathways for genome instability such as P53, DNA repair, and telomere extension and maintenance (Table S4), which were not commonly used by luminal cells. Most of the essential genes affecting genome instability pathways are relatively unique for the basal subtype, which highlights the signature of basal subtype and provides unique drug targets for the aggressive groups such as triple-negative groups.

Subtype-specific survival signaling networks provide predictive power

The convergence of the cancer essential genes and driving regulators into their respective subtype-specific survival networks suggests that in each subtype there is a "deterministic" path for cancer cell proliferation driven by genomic alterations, and the networks could therefore provide "predictable" power for selective genomic alterations. Consequently, we tested whether the integrated subtype-specific networks could have predictive power in order to accurately identify breast cancer tumor subtypes. To demonstrate this, we used the hub genes of the subtype-specific networks to classify the

16 cell lines. To do so, we first identified differential hubs between the subtype-specific networks (see Extended Experimental Procedures) and then classified the 16 cell lines. Indeed, hub genes were able to distinguish between luminal and basal subtypes (Figure 4A; $p = 1.2 \times 10^{-4}$, fuzzy k-mean clustering and Fisher's test). These results suggest that amplification or mutation of a few top hub genes could activate the entire network for cancer cell survival and proliferation. Therefore, we extended this analysis to demonstrate that these hub genes' genomic alteration profiles (amplification and drive-mutating status) were able to significantly classify 402 breast tumor samples (see Extended Experimental Procedures) into the basal and luminal subtypes (Figure 4B; $p = 2.2 \times 10^{-16}$, fuzzy k-mean clustering and Fisher's test). These results highlight the convergent and deterministic properties of selective genomic alterations, which exploit distinct core survival signaling networks (i.e., subtype-specific networks) for cancer cell proliferation. These genomic alterations could be gradually or suddenly (i.e., through genome duplication) accumulated and then selected during cancer evolution. Detection of the genomic alterations of a fraction or all of the genes in this hub gene set could help in the early diagnosis of breast tumors. Recently, a plasma genome sequencing approach has shown that copy number variations and mutations of plasma DNA are detectable and comparable between cancer patients and healthy individuals[97,98]. As sequencing costs continue to decline, these genes could be used to develop noninvasive

tests (e.g., using plasma genome sequencing[99]) for screening very early stage breast cancer patients or distinguish breast cancer subtypes.

To further demonstrate their predictive power of the subtype-specific networks, we sought to predict subtype-specifically therapeutic interventions. If a hub gene specifically appears in either a luminal or basal subtype-specific network, we expected that this gene could be a drug target specifically for its subtype. Based on this criterion, AKT1, mTOR, MET, MDM2, HSP90AA1, RAF1, SFN, FYN, CHEK1, and ESR1 were predicted as potentially luminal-specific drug targets, while TGF-β, IGF1R, MAPK3, GRB2, SRC, TUBB, JAK2, and EGFR were predicted as potentially basal-specific drug targets (undruggable differential hubs between subtypes such as transcription factors like P53 were not considered). To validate these predictions, we obtained the data from systematic drug screenings of cancer cell lines, including over 40 breast cancer cell lines[100,101], and statistically evaluated the sensitivity of these drugs for luminal and basal subtypes (see Experimental Procedures). The predicted targets, which have been included in the drug screenings (except for MAP2K1 and CHEK1), were in agreement with the experimental screening results (Table 1).

In summary, using an integrative network analysis of the data derived from exome sequencing and genome-wide RNAi screening of a breast cancer cell line panel, we have shown that a set of primitive core signaling pathways such as cell cycle, apoptosis, growth factors/MAPK, and transcription are commonly exploited by

genomic alterations for cancer cell survival in all the breast cancer cells, while the signaling pathways of P53 and genome instability such as telomere maintenance are specifically exploited by genomic alterations in the basal subtype. The essential genes in these pathways are unique drug targets for the aggressive breast (i.e., basal subtype) cancer groups. The functional convergence of the essential genes and driving regulators in a limited number of signaling pathways leads to the emerging of subtype-specific survival signaling networks in which genes recurrently switch roles between cancer-essential genes and cancer-driving regulators in cancer cells. These networks elucidate underlying signaling mechanisms governing cancer cell survival and proliferation and imply selective pressures for evolutionary convergence of cancer genomic alterations. However, it is clear that signaling mechanisms of the two subtypes are different. This is evident by the existence of a set of network genes (i.e., genes that are differentially different between the two subtype-specific networks) whose genomic alteration profiles (amplification and mutating status) can significantly distinguish breast tumor samples into luminal and basal subtypes. Furthermore, these networks predicted subtype-specific drug targets. Importantly, most (~80%) of the predicted drug targets have been experimentally validated. Together with the finding that more amplified genes could act as cancer drivers, these results may have profound clinical implications in the personalized treatment of cancer patients[30,31] and the screening of early stage breast cancer patients by plasma DNA sequencing using this set of network genes.

Experimental procedures

Samples for exome sequencing

Eleven breast cancer cell lines (BT549, MDAMB436, BT20, MDAMB231, MDAMB468, SKBR3, ZR751, HCC1500, MDAMB453, MCF7, and T47D) were obtained from ATCC for exome sequencing.

Data sets

Exome-sequencing data for five breast cancer cell lines (Table S1) were obtained from[87]. Microarray and copy number data of the 16 breast cancer cell lines were obtained from the CCLE (http://www.broadinstitute.org/ccle/home). Data for genome-wide RNAi screening of cell survival and proliferation of the 16 breast cancer cell lines were obtained from the COLT-Cancer database (http://colt.ccbr.utoronto.ca/cancer/). The human signaling network (Version 4, containing more than 6,000 genes and more than 50,000 relations) includes our previous data obtained from manually curated signaling networks[90,93,94] and by PID (http://pid.nci.nih.gov/) and our recent manual curations using the iHOP database (http://www.ihop-net.org/UniPub/iHOP/). Pathway gene lists were obtained from the GSEA Molecular Signatures Database

(http://www.broadinstitute.org/gsea/msigdb/index.jsp). Data of systematic drug screenings of breast cancer cell lines were obtained from these studies[100,101].

Cancer essential genes, proliferation-influencing genes, and driving regulators

The following descriptions of driving regulators and essential and proliferation-influencing genes are summarized in Figure S1. Genome-wide RNAi screening results of the 16 breast cancer cell lines was collected in the COLT-Cancer database (http://colt.ccbr.utoronto.ca/cancer/). In the database, the essentiality of each gene for a given cell line has been scored based on GARP (Gene Activity Rank Profile) scores and p values, which were computed in each experiment of the genome-wide RNAi screening[53]. A lower p value depicts higher significance for the "higher gene essentiality" (e.g., higher degrees of influencing cell survival). Details for calculating of GARP scores and p values were described previously[53]. Housekeeping genes were also annotated in the database. If a gene in a given cell line has a RNAi-screening p value < 0.05 and does not belong to the housekeeping genes, that gene was defined as a "cancer-essential gene" in that cell line[53]. Validation experiments supported this p value cutoff (i.e., 0.05) for defining the cancer-essential genes[53]. If a gene in a given cell line has an RNAi-screening p value less than 0.1 but greater than 0.05, we defined that gene as a

"proliferation-influencing gene" in that cell line. We assumed that knocking down a proliferation-influencing gene will not lead to cell death, but will significantly reduce cell growth and survival. We asked that an essential gene, proliferation-influencing gene in a given cell line should be among the top 75% of the expressed genes for that cell line as described previously[44]. Amplification genes are considered if they have a GISTIC score > 0.3 and are among the top 50% of the expressed genes for that cell line. The cutoff 0.3 is widely used to define gene amplifications[102,103]. Details of setting the cutoff of 50% for gene expression are explained in Extended Experimental Procedures. If an amplified gene in a given cell line has a RNAi-screening p value < 0.4, we defined this gene as a cell-survival-related driving regulator in that cell line, assuming that knocking down the driving regulators will affect cell growth and survival. It should be noted that the definitions of these terms are based on certain cutoffs. We changed the cutoffs of RNAi-screening p values for these genes (i.e., $p < 0.03$, $0.03 < p < 0.1$, and $p < 0.5$ for cancer-essential genes, proliferation-influencing genes, and driving regulators, respectively) and reran all the analyses in this study. We found that the results are similar to those obtained using the original cutoffs. However, when interpreting the results, one should take into consideration the definitions and the cutoffs used in this study.

Drug sensitivity analysis

For a given drug, we compared the IC_{50} values between luminal and basal lines. Kruskal-Wallis ANOVAs were used to evaluate the statistically significant differences in IC_{50} values between the subtypes. Heiser et al. (2012) performed drug screening on more breast cancer cell lines (~50 cell lines) than Garnett et al. (2012). Therefore, we mainly used the data from Heiser et al.

2.4 Extended experimental procedures

Exome capture and sequencing

Each qualified genomic DNA sample was randomly fragmented into fragments with a base pair peak of 150 to 200 bp, and then adapters were ligated to both ends of the resulting fragments. The adaptor-ligated templates were purified by the Agencourt AMPure SPRI beads and fragments with insert size about 200bp were excised. Extracted DNA was amplified by ligation-mediated polymerase chain reaction (LM-PCR), purified, and hybridized to the SureSelect Biotinylated RNA Library (BAITS) for enrichment. Hybridized fragments were bound to the streptavidin beads whereas non-hybridized fragments were washed out after 24h. Captured LM-PCR products were subjected to Agilent 2100 Bioanalyzer to estimate the magnitude of enrichment. Each

captured library was then loaded on Hiseq2000 platform. We performed high-throughput sequencing for each captured library independently to ensure that each sample meets 30x coverage. Raw image files were processed by Illumina base calling Software 1.7 for base calling with default parameters and the sequences of each sample were generated as 90bp paired-end reads. The reads have reached ~60% coverage at 20x depth and ~80% coverage at 10x depth of the Agilent exome-array defined CDS targets. The Exome capture and sequencing were performed by BGI (Beijing Genome Institute).

Sequence alignment and genome mapping

For the sequencing data (raw data) generated from the Illumina pipeline, the adaptor sequences in the raw data were removed, and low quality reads which have too many Ns and low base quality bases were discarded. This step produced the "clean data." We applied Burrows-Wheeler Aligner (BWA) to conduct the alignment. BWA gives the result in BAM format files. The BAM format files were used to perform other processes, such as fixing mate information of the alignments, adding read group information, marking duplicate reads caused by polymerase chain reaction (PCR). After completing these processes, the final BAM files for variant calling were generated. Quality Control was applied in the whole pipeline for cleaning data, sequence

alignment, and calling variants etc. Sequencing reads were aligned to the reference genome sequence using BWA. For mapping we used the human genome build37 (hg19) as the reference genome.

Mutation detection and identification of potential cancer driver-mutating genes

We performed the sequencing data-processing pipeline as the same as that described previously[103] with the following modifications: (1) Hg19 was used as the Human Reference Genome; (2) dbSNP137 and the data of 1000 Genome projects (version February 2012) were used for variant filtration. Variants were considered to be somatic mutations only if they were absent in the 1000 Genomes database and the dbSNP database. For mutation data interpretation, it should be aware of that the cancer cell lines sequenced do not have matched normal genomes. Although we filtered somatic mutations by removing sites in dbSNP and 1000 genomes, which likely represent miscalled germline variants, the variants called from the cell lines still contain rare germline variants; (3) Nucleotide substitution were detected with MuTect (version 1.1.4)[104] by applying default parameters. Short indels were called with UnifiedGenotyper in GATK (version 2.5-2)[105]. These somatic mutation/variant data were generated using Annovar (version 2012, May25) (Wang et al., 2010) with the refSeq annotation. We validated the variants called in this study using Cancer Cell Line

Encyclopedia (CCLE, http://www.broadinstitute.org/ccle/home) cell lines, which have been targeted sequenced for pre-selected 1,600 genes. The eleven cell lines sequenced in this study are also included in CCLE. By comparing the variants of the 1,600 genes in both studies, we found that 85% of the variants called in this study have been supported by CCLE.

Cell-line-specific survival network construction

After collecting the essential genes, driving-regulators and proliferation influencing genes for a given cell line, we mapped these genes onto the human signaling network, and then extracted the mapped nodes and their links as the survival network for that cell line.

Determination of top 50% of the expressed genes for amplified genes

If a gene is amplified as a potential cancer driver across a set of cell lines, it is possible to examine whether it significantly regulates other genes. Therefore, we examined whether an amplified gene as a regulator significantly regulates other genes' expression. To do so, we obtained the gene expression data and the SNP6.0 data (for

copy number variations) of the 51 breast cancer cell lines from CCLE. For an amplified gene defined in the 16 cell line panel, we assigned the 51 cell lines into two groups where one group consisted of the cell lines in which that gene was amplified (GISTIC score > 0.3 and among the top 75% or 50% of the expressed genes in that cell lines), the other group consisted of the rest of the cell lines. If both groups defined by an amplified gene contained at least 10 cell lines, we further conducted a t test to examine whether it significantly regulated other genes between the two groups. We found that ~90% and ~70% of the amplified genes do significantly regulate other genes when used Top 50% and 75% of the expressed genes as cutoffs, respectively. Therefore, we used Top 50% of the expressed genes for the amplified genes.

Subtype-specific survival signaling network construction

Among the cell line-specific survival networks within a subtype, if a gene plays a role as an essential gene, a driving-regulator or a proliferation influencing gene in at least three cell lines, we defined it as subtype-specific network genes. For a given subtype, the subtype network genes and their links in the original human signaling network formed a subtype-specific network.

Permutation and randomization tests of networks

To test the significance of subtype classification using the cell line-specific networks, we conducted permutation tests. Briefly, we randomly shuffled the networks between the subtypes and used the 'network hubs' to classify the 16 cell lines (using fuzzy k-mean clustering) to test (using Fisher's test) whether the network hubs are able to significantly classify the 16 cell lines into luminal and basal subtypes. We conducted 10,000 times of random shuffling. Among the 10,000 times of permutations, we counted the number (N) of the permutations which generated a small p value ($p < 0.05$) for Fisher's tests. N/10,000 is the p value for the permutation tests. By doing so, we obtained a low p value ($p = 0.0009$), suggesting that by random the networks can't significantly classify the subtypes.

To demonstrate the statistically significant recurrent usage of network genes by their subtype-specific networks, we conducted randomization tests. Briefly, we first constructed a cell line-specific 'random network' by randomly assigning the genes for essential, proliferation influencing genes and driver-regulators onto the human signaling network to extract the 'random network' for that cell line. The number of the genes assigned for each category (i.e., essential, proliferation influencing genes and driver-regulators) is the same as the original number of these genes in that cell line. Using the cell line-specific 'random networks', we constructed 'subtype-specific random

networks' followed the same procedure for constructing the original subtype-specific networks. The original subtype-specific networks contain ~200 nodes. In the randomization tests, we tested whether each 'subtype-specific random network' contains at least 10 nodes. None of the 'subtype-specific random networks which contain at least 10 nodes was observed among 10,000 times of randomizations ($p < 1.0x10-4$). Furthermore, we tested whether the hubs of the 'subtype-specific random networks' are able to classify the 16 cell lines into luminal and basal subtypes using k-mean clustering and Fisher's test ($p < 0.05$ is regarded as significant). It is not surprise that these 'subtype-specific random networks' can't do so ($p < 1.0x10-4$).

Pathway enrichment analysis for subtypes

We first downloaded the pathway information from Gene Set Enrichment Analysis (GSEA) database (http://www.broadinstitute.org/gsea/msigdb/index.jsp). After collecting the essential genes (E), driving-regulators (R) and proliferation influencing genes (PI) for a given subtype-specific network, we counted the number of E (Ne), R (Nr) and PI (Npi), and then mapped each gene to its gene family based on Ensembl genome database (http://www.ensemblgenomes.org/). For a given pathway, we determined the number of genes (families) for E (PNe), R (PNr) and PI (PNpi) by mapping E, R and PI genes onto that pathway, respectively. We conducted

randomization tests by randomly assigning the same number of gene (families) of E, R and PI genes (Ne, Nr and Npi), respectively, to the pathway gene (family) pool. Multiple test p values were corrected by FDR. Pathways, which are significantly enriched with all three-category genes, E, R and PI (q value < 0.1), were regarded as statistically enriched pathways for the subtype.

Identification of functional network modules using gene ontology-guided markov cluster algorithm

To identify network modules, we used luminal subtype-specific network as an example. We first obtained network clusters using Markov Cluster (MCL) Algorithm, which has been implemented in GraphWeb[106] on the luminal subtype-specific network. MCL defines a set of nodes as a cluster if random walk in the network is likely to remain in that set. For the luminal subtype-specific network, MCL gave several big and small network clusters (for example, small clusters contained only 5-8 nodes). Furthermore, not all the nodes (called isolated nodes here) in the network had been assigned to network clusters by MCL. To merge these network clusters and the isolated nodes into functional modules, we performed Gene Ontology (GO) enrichment analysis on each cluster using the GO term analysis tool, DAVID[107]. For the clusters which have significant GO terms, we ranked the GO terms for each cluster, and picked up the top-

ranked GO terms which are among cancer hallmarks such as cell cycle, cell proliferation and apoptosis (we are more interested in these GO terms because we are dealing with cell survival networks) as described previously[92]. Network clusters representing more than 70% of the subtype-specific network genes were significantly enriched with one of the cancer hallmark GO terms. For two network clusters, which shared a picked GO term, if they shared higher linkage between them (i.e., their linkage is significantly higher than the average linkage of all the network clusters), we merged them into a functional module. By doing so, all the network clusters which shared a same picked GO term will be merged into a functional module. For the remaining clusters (for example, small clusters that had no significant GO terms) and the isolated nodes which did not yet belong to any of the functional modules, we assigned them based on function and linkage. For example, if a cluster or node that was not assigned to one of the modules, we manually checked their function and assigned them to the right functional module, if they have links with that module. Otherwise, we calculated the linkages from a gene/cluster to all the functional modules. A gene/cluster was assigned into the functional module which has the highest linkage with that gene/cluster. The same procedure was applied to basal subtype-specific networks.

Classification of cell lines and breast tumor samples using hub genes of the subtype-specific networks

We collected the hub genes which specifically appear in either luminal or basal subtype-specific networks for further clustering analysis. For clustering of the 16 cell lines, if a hub gene is an essential gene, a driving regulator or a proliferation influencing gene in a given cell line, it was marked as 1, otherwise 0 for that cell line. Data of all breast tumor samples were obtained from The Cancer Genome Atlas Network[88]. From this data set, we extracted the PAM50[42] defined luminal and basal tumor samples, which also have the data of genome sequencing, SNP6.0 array and microarray. By doing so, 402 tumor samples were obtained and then clustering analysis was conducted. For a given sample, if a network hub gene is amplified and also among Top 50% of the expressed genes, or functionally mutated, we marked it as 1, otherwise 0 in that tumor.

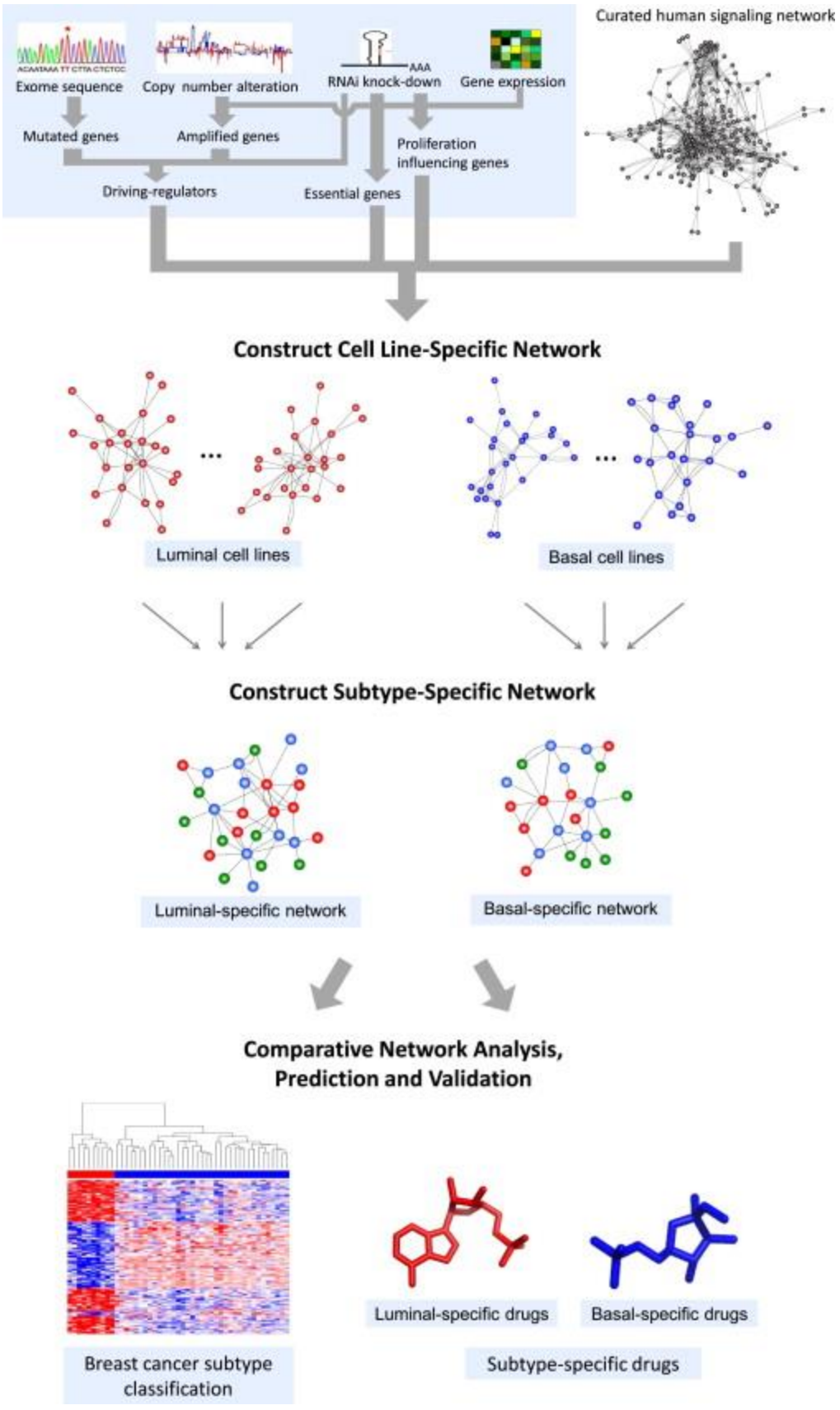

Curated human signaling network
Exome sequence
Copy number alteration
RNAi knock-down
Gene expression
Mutated genes
Amplified genes
Proliferation influencing genes
Driving-regulators
Essential genes
Construct Cell Line-Specific Network
Luminal cell lines
Basal cell lines
Construct Subtype-Specific Network
Luminal-specific network
Basal-specific network
Comparative Network Analysis, Prediction and Validation
Luminal-specific drugs
Basal-specific drugs
Breast cancer subtype classification
Subtype-specific drugs

Figure 1.

Analysis of Integrated Networks for Breast Cancer Cell Survival and Proliferation

The data of genome sequencing, genome-wide RNAi screening, copy number variations, and gene expression profiles of individual lines were used for constructing an integrated network for each individual cell line. Cell-line-specific networks across each of the breast cancer subtypes were used for constructing subtype-specific networks for cancer cell survival and proliferation. Comparative and differential analysis of the subtype-specific networks allowed us to predict subtype-specific treatments and significantly classify breast tumor samples. See also Figure S1 and Table S1.

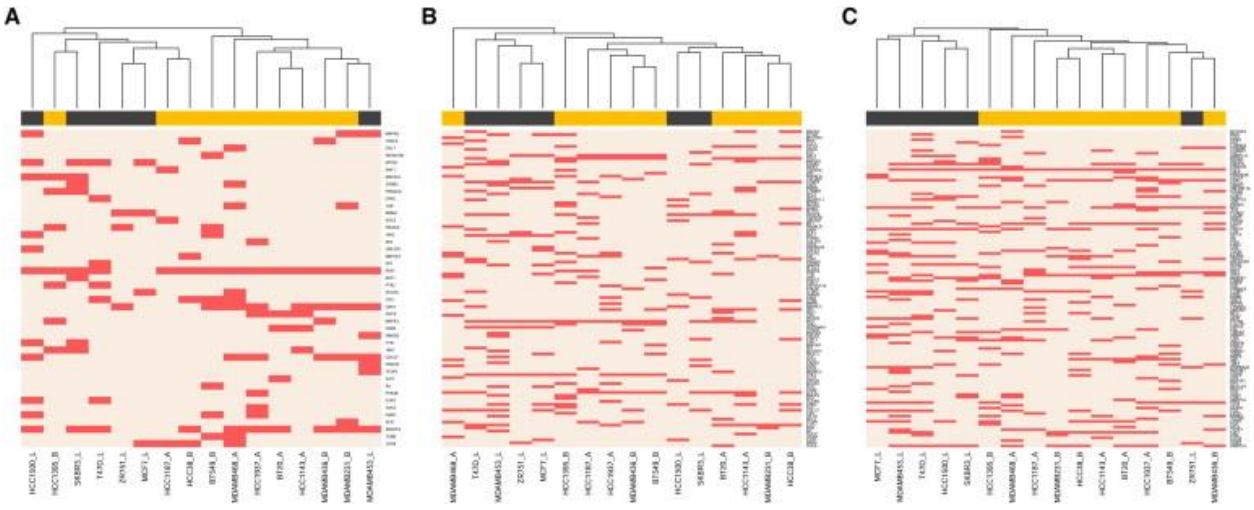


Figure 2.

Hierarchical Clustering of the 16 Breast Cancer Cell Lines

Hierarchical clustering of the cell lines using cell-line-specific network hubs: (A) driving-regulator hubs, (B) essential gene hubs, and (C) the hubs of essential genes and driving regulators combined. Red and beige in the heatmaps indicate whether the hub genes are present or absent, respectively, in a cell line. See also Table S3[108].

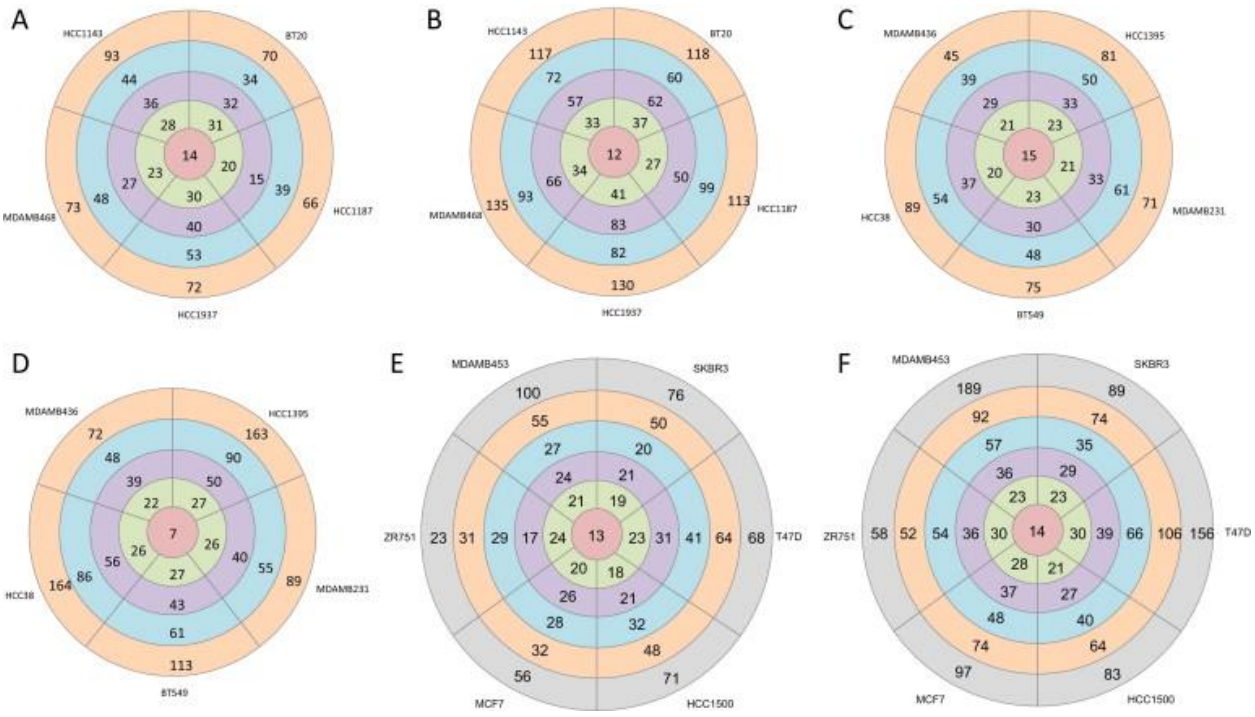


Figure S2.

Dartboard Showing Overlap of Essential Genes and Driving Regulators for Cell-Line-Specific Networks of the Subtypes, Related to Figure 3

Overlapping of essential genes for basal A (A), basal B (C) and luminal (E) and overlapping of driving-regulators for basal A (B), basal B (D) and luminal (F). The outer most circle colored with peach (basal A and basal B) or gray (luminal) represents the number of genes that do not overlap with any other cell line within their respective subtype, and are therefore unique to that cell line. Going toward the center, the bull's eye contains the number of genes shared by all the cell line-specific networks in that subtype.

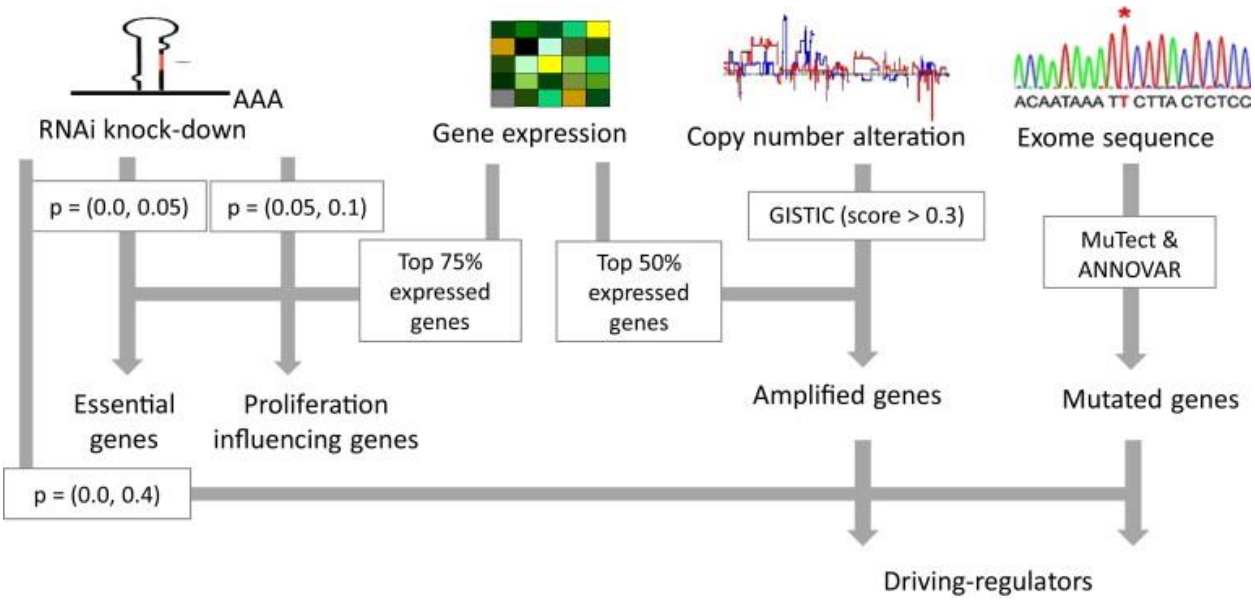


Figure S1

Data Sources and Cutoff Values for Defining Driving Regulators and Essential and Proliferating-Influencing Genes, Related to Figure 1

For a given cell line, if a gene is among Top 75% of the expressed genes and its RNAi-screening p value < 0.05, it is defined as an essential gene in that cell line; if a gene is among Top 75% of the expressed genes and its RNAi-screening 0.05 > p value < 0.1, it is defined as a proliferation influencing gene in that cell line; if a gene is driver-mutating gene, or has a GISTIC score > 0.3, plus among Top 50% of the expressed genes, and if its RNAi-screening p value < 0.4, it is defined as a driving-regulator for cell survival in that cell line.

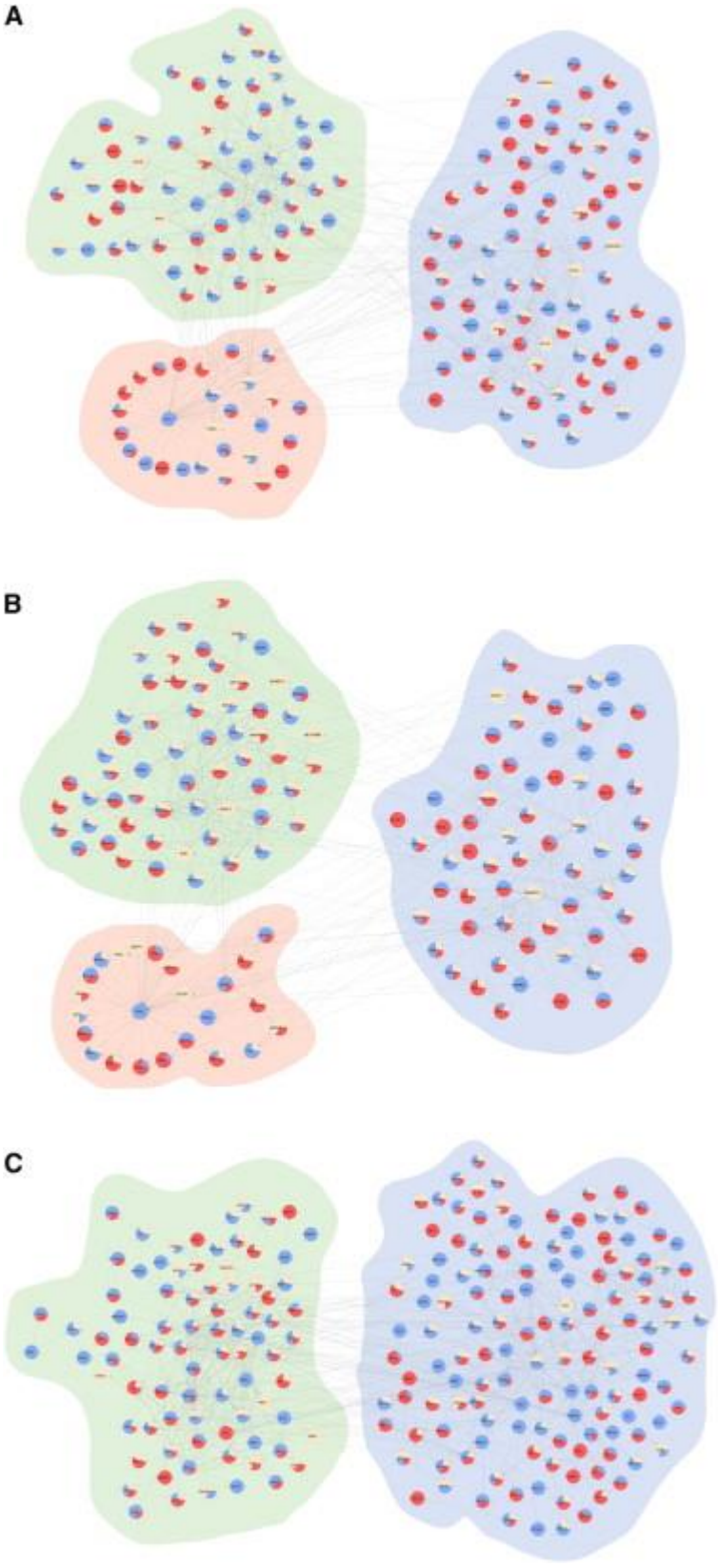

A
B
C

Figure 3.

Subtype-Specific Survival Signaling Networks

Subtype-specific survival signaling networks for basal A (A), basal B (B), and luminal (C) subtypes. Nodes represent genes while links represent regulation (directed links) or interaction (neutral links) between genes. A node is represented by a pie chart that shows each gene's distribution as essential gene (red), a driving-regulator (blue), or a proliferation-influencing gene (cream) in its subtype. The background color behind the clusters represents a cluster's function in relation to one of the cancer hallmarks: apoptosis (pink), cell proliferation (green), and cell cycle (blue). Cytoscape[109] was used to present and visualize the networks. See also Figure S2 and Table S4[108].

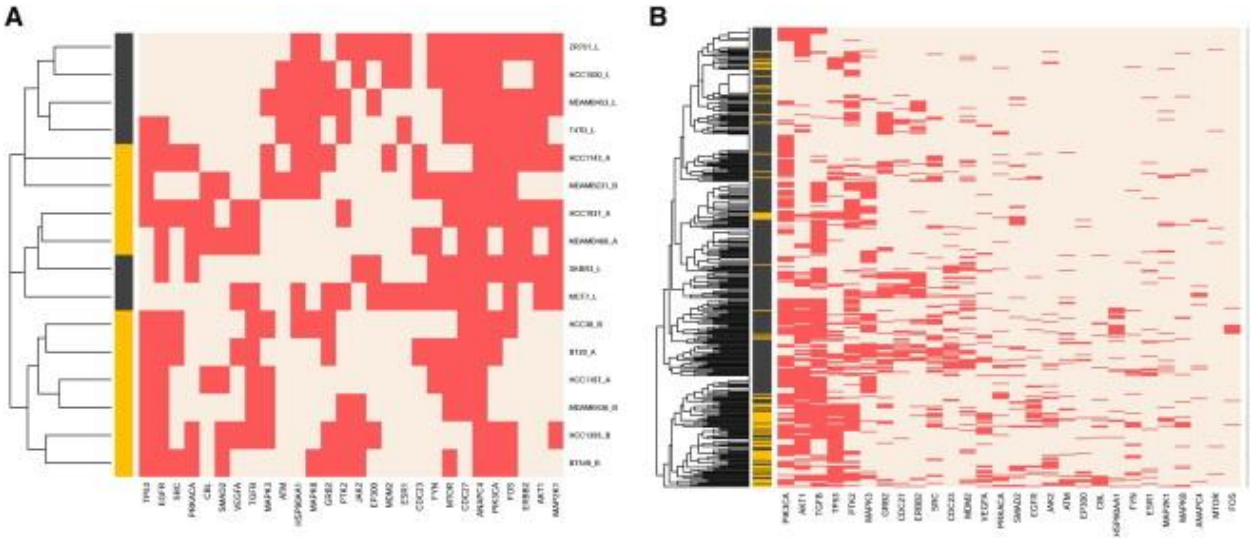


Figure 4.

Clustering of 16 Breast Cancer Cell Lines and 402 Breast Tumor Samples Using the Hubs from Subtype-Specific Networks

(A) Hierarchical clustering of the 16 cell lines using the differential hubs from the subtype-specific networks of luminal and basal subtypes. In the heatmap, for a given cell line, a hub gene appears in red if it is an essential gene, a driving regulator, or a proliferation-influencing gene; otherwise, it appears in beige. On the side bar, gray and yellow represent luminal and basal cell lines, respectively.

(B) The same differential hubs from (A) were used to classify 402 breast tumor samples. In the heatmap, red represents mutated genes or amplified genes that are among the top 50% of the expressed genes for tumor samples.

Table 1.

Validation of the Predicted Subtype-Specific Drug Targets

Compound	Predicted Subtype-Specific Drug Target	Basal versus Luminal p Value	Subtype Specificity
Sigma AKT12 inhibitor	AKT1, AKT2 (luminal)	5.04×10^{-4}	luminal
Tamoxifen	ESR1 (luminal)	3.92×10^{-2}	luminal
Nutlin-3a	MDM2 (luminal)	3.13×10^{-2}	luminal
Rapamycin	mTOR (luminal)	1.78×10^{-3}	luminal
17-AAG	HSP90 (luminal)	3.98×10^{-2}	luminal
Bosutinib	SRC (basal)	1.08×10^{-2}	basal
Docetaxel	TUBB1 (basal)	1.27×10^{-2}	basal
BMS.536924	IGF1R (basal)	4.95×10^{-2}	basal
VX-680	JAK2 (basal)	4.95×10^{-2}	basal
Erlotinib	EGFR (basal)	2.33×10^{-2}	basal
RDEA119	MAP2K1/MEK12 (luminal)	2.04×10^{-2}	basal
TCS 2312 dihydrochloride	CHEK1 (luminal)	1.46×10^{-1}	not significant

Transition from Chapter 2 to Chapter 3

In Chapter 2 we saw how a network approach can be used to identify subtype specific drug targets for breast cancer. This was the first work to integrate RNAi screening data.

Highlights of Chapter 2:

- Integration of networks with multiomic and shRNA data identifies cancer genes
- Genes switch roles between cancer causing and essential among cancer subtypes
- Evolutionary convergence and deterministic paths of cancer genomic alterations
- Subtype-specific networks successfully predicted subtype-specific drug targets

Drug prediction has come a long way and continues to be improved. However, in the cancer research field, one of the fundamental questions is how do mutations build up in a healthy individual to transform a normal cell into a cancerous one. Two patients who may have the same cancer type but have very different mutational profile. In addition, there is intra-tumor heterogeneity because of the different sub-populations. In Chapter 3, we aim to identify new somatic mutations required to develop cancer, given a patient's germline variants. This is patient-specific and does not rely on predefined subgroups.

CHAPTER 3

Predicting Key Personalized Cancer Driver-Mutating Genes in Advance Based on Healthy Individuals' Genetic Makeup (Ready for submission)

Zaman N, Milanese JS, Tibiche C, Nantel A, Wang E.

3.1 Introduction

A majority of cancers are diagnosed at the middle- or late-stage (i.e., advanced cancer) at which time most tumors have spread and become incurable. Consequently, with some notable exceptions, improvements in overall survival and morbidity over the past few decades have been modest. Historical cancer data in the Surveillance, Epidemiology, and End Results program (http://seer.cancer.gov/) illustrate that advanced cancer has poor survival, whereas, cancer diagnosed at early stages (Stage I and II) has relatively good survival rates. These data suggest that, other than lifestyle changes, early detection could be one of the most effective approaches to reducing the growing cancer burden. For example, early detection of colorectal cancer through colonoscopy saves many lives each year. A substantial reduction in colorectal cancer death rate in the USA can be attributed to early detection, where ~95% of Stage I colon cancer patients can be cured through surgery alone. At present, most cancers are

diagnosed too late and become incurable. Therefore, one of the best hopes for reducing mortality from cancer is clearly the development of a sensitive and specific screening test to detect early curable disease. Detecting cancers when they are at their earliest stages allows for treatment strategies which have a good chance of truly curing the disease instead of simply adding a few extra months or years of life.

While a substantial effort for cancer early detection has been made, relatively few approaches have proven sufficiently effective. Recent advances in genome sequencing provide tremendous potential for the development of tools to aid cancer early detection. However, to date, no methods have been developed to construct clinically-useful predictive models from genome sequencing data, mostly because of excessive variability in the identity of mutated genes even within tumors of a same type. For example, two breast tumors of the same type and stage will rarely share the same cancer-driver genes. This issue has become a crucial bottleneck in the translation of sequencing technology to the clinic. Here we developed an algorithm, eTumorMonitor, that predicts key personalized cancer-driver mutated genes whose mutations are required for the first steps of malignant transformation (i.e. the formation of the cancer founder cells) based on the patient's original genetic makeup. Validation tests of the algorithm in several hundreds of breast and ovarian cancer patients showed that the predicted key somatic driver-mutating genes (~30-50) based on individuals' genetic makeup have been enriched by 16-25-fold compared to a random set. Along with

improvements in liquid biopsy methodologies, this algorithm could be used for the early detection of tumors. For example, for a given healthy individual who is at high-risk of developing breast cancer, eTumorMonitor is able to predict ~30-50 key personalized somatic mutating genes which drive malignant transformation based on her germline genomic landscape. A breast tumor formed at its early stage will release cell-free DNA (cfDNA) in the blood. Therefore, prospectively targeted sequencing of blood cfDNA samples could detect predicted mutating genes associated with the early stage of cancer. Thus, eTumorMonitor could help in personalized early-cancer diagnostics by prospectively targeted sequencing of blood cfDNAs.

3.2 Results

An overview of eTumorMonitor

Cancer is a process of asexual evolution driven by genomic alterations. A single normal cell randomly acquires a series of mutations that allows it to proliferate and to be transformed into a cancer cell (i.e., founding clone, or FClone) thus initiating tumor progression. This process is called malignant transformation. Our underlying hypothesis is that mutagenic processes are essentially blind or non-purposeful. However, to drive a malignant transformation, new mutations will be selected if they integrate into the pre-existing genomic landscape (i.e., germline mutations) to trigger or

activate a cancer survival and proliferation network (i.e., cancer survival network) which promotes clonal expansion[32]. Therefore, eTumorMonitor uses cancer type-specific survival networks to predict key somatic-mutated genes that are required to work together with the pre-existing germline mutations to drive a malignant transformation in an individual who has high-risk of developing cancer. eTumorMonitor has 3 components: (1) Discriminating and -Aligning of Network-derived Profiles (DANP); (2) Driving Targeted-tumor profiles via Network Perturbation (DTNP); (3) prioritization of predicted candidates.

Because each tumor has an individually unique genomic profile, it is very hard to construct predictive models using only gene mutations. Therefore, we first transform a sample's functionally mutated genes (i.e., all the mutations mentioned here are functional mutations, which have been determined by a few tools, see Methods) into a Network-derived Profile (NetProfile), which is similar to a gene expression profile, by projecting mutated genes on a cancer survival network using a network propagation approach[110]. This approach works by projecting the mutated genes of a sample onto the network where each mutated gene is represented as a heat source. The heat source diffuses to neighboring genes along the edges of the network. At a certain time point, the diffusion stabilizes, until and finally each gene on the network receives a certain amount of 'energy', which is represented by a 'heating-score' (Fig 5). By converting mutations into NetProfiles, we can overcome the challenge of individually unique

mutations and construct predictive models from sequencing data.

In TCGA, data from each breast cancer patient includes whole-exome sequencing data for their normal blood and paired tumor, which means that, for each patient, we can generate a germline-NetProfile and its paired FClone-NetProfile. There are sets of genes (i.e., gene signatures) whose differences in heating-scores between the germline-NetProfile and paired FClone-NetProfile represent the transition from a normal cell to a cancer cell. Therefore, for a given germline sample, DTNP will model the network in silico to examine which genes, beyond pre-existing germline mutations, need to acquire mutations to drive the gene signature's profile (i.e., heating scores for all the genes of the gene signature) of the germline into a pattern similar to those observed in luminal breast tumors (Fig 6).

Both germline-NetProfiles and FClone-NetProfiles are heterogeneous, it is thus possible to classify either germlines or FClones into 'subypes' (i.e., subgroups) based on their NetProfiles. Therefore, we proposed a super-gene signature (SGS), which is composed of a set of cancer hallmark-based genes, that can classify FClones and germlines into highly consistent subgroups. For example, a SGS can cluster FClones and germlines, respectively, into 2 subgroups, such that germline subgroup 1 and FClone subgroup 1 share over 85% of the patients, and so do germline subgroup 2 and FClone subgroup 2. DANP is designed to identify cancer hallmark-based SGSs based on NetProfiles.

In summary, to improve the accuracy of mutated genes prediction, DANP identifies a SGS which discriminates subgroups consistently for germlines and FClones (Fig 6). For a given new germline sample, based on a SGS, DTNP assigns it to germline subgroup 1 or germline subgroup 2. If it is assigned to germline subgroup 1, its expected FClone will be assigned to FClone subgroup 1. Further, DTNP examines which network gene would have to be mutated mutated in the context of the pre-existing germline mutations for the updated profile of that SGS to move closer to the centroid (eg, average heating scores of SGS's genes of the subgroup samples) of that SGS for FClone subgroup 1 (Fig 6). Finally, we prioritize the predicted candidates using the network features of cancer driving-mutated genes.

As a proof-of-concept, we used the sequencing data of breast tumors and their paired normal samples to implement eTumorMonitor. Breast cancer has two major molecular subtypes: luminal and basal. Luminal tumors represent a large proportion of the breast tumors. TCGA collected several hundreds of luminal tumors but only ~100 basal tumors. Unfortunately, TCGA does not include sufficient basal-breast tumors for both the construction of predictive models and and their validation. Therefore, we decided to limit this study to luminal tumors. We respectively used 150 and 130 luminal breast cancer samples as a training set and a validation set.

Identify super-gene signatures using DANP

Tumor genome sequencing studies showed that each tumor has an individually unique genomic profile even for a same cancer type or subtype. This feature makes very difficult to construct predictive models using only gene mutations. For example, in a recent Dialogue for Reverse Engineering Assessment and Methods (DREAM) effort that benchmarked ~50 prediction algorithms, the authors have shown that genome sequencing data alone cannot be used to build predictive models[111]. To develop eTumorMonitor, we have to overcome this hurdle. Our earlier analysis of cancer mutations on a human signaling network showed that mutations of different patients form network modules and clusters on networks[90]. Similar results have been obtained when using genome- sequencing data. Figure 5 and Figure 7 shows how discreet mutation data can be transformed into a continuous form. This allows for better comparison between patients with different mutations. Therefore, by converting mutations into network-derived profiles, network profiling overcomes its limitations in constructing predictive models using sequencing data. Several algorithms such as random walk and network propagation could be used for network profiling. As a first step in allowing network profiling, we constructed a luminal-breast cancer survival network as described previously[108]. The uniqueness of this survival network is that it uses experimentally determined luminal breast cancer-specific specific survival and proliferation genes were used. By doing so, we ensure that the network is related to

cancer cell proliferation and reduce the impact of noisy data so that network-based prediction produces better results[108]. For each tumor sample and its paired germline, we seeded the survival network with functionally-mutated genes and used the network propagation algorithm to generate tumor FClone- and germline-Netprofiles. Thesed netProfiles were then transformed and normalized as described in the Methods section.

To improve prediction accuracy, we also classified germlines and their paired FClones between 2 subgroups so that new samples could be assigned into a particular subgroup prior to predicting somatic-mutated genes. We used DANP (see Methods) to identify 7 SGSs, each of which containing 30 genes, to represent different cancer hallmarks. Each SGS classifies germlines and FClones, respectively, into 2 consistent subgroups. We further validated (see Methods) the SGSs using the validation set (130 samples) and showed again that FClone subgroup 1 and germline subgroup 1 share over 85% of the patients as do the FClone subgroup 2 and germline subgroup 2.

Predict somatic-mutated genes base on germline mutations using DTNP

To predict the key genes whose mutations would feed into the pre-existing gremline mutations to drive luminal-breast cancer malignant transformation, we applied DNTP to the validation samples. For each sample, based on its germline mutations, DTNP uses a SGS to identify 50 somatic-mutated genes. The predicted mutated genes in that sample were benchmarked and compared with the genes that

were actually mutated in the paired FClones. On average, 6% of the predicted genes based on each SGS needed to be corrected. We extended the predictions using DTNP to the validation samples and obtained similar results.

Prioritize predicted candidates using features of cancer mutated genes

To improve the prediction accuracy of eTumorMonitor, we prioritized the predicted candidates in each samples based on multiple SGSs and the network features of cancer-mutated genes. To get the network features of the somatic cancer mutation genes, we analyzed the germline mutated genes and FClones' somatic-mutated genes in the training samples. On average, a breast cancer patient has 230 and 150 mutated genes in it's germline and founding clone, respectively. The mutated genes in founding clones are almost always a subset (87.0%) of those mutated in germline samples. It is well-known that somatic mutated genes are rarely shared between tumors even of a same cancer type These results thus suggest that both germlines' and FClones' mutations are convergent to a set of genes that could drive malignant transformation. Furthermore, the frequency of mutated genes (i.e., a fraction of the samples where a gene gets mutated) in FClones is positively correlated with the frequency observed in germline samples (correlation coefficient, cc= 0.68, $P=2.2x10^{-16}$ and cc= 0.75, $P=2.2x10^{-16}$, respectively, for luminal-breast and ovarian cancer). These results indicate that a subset

of the mutated genes is more critical for malignant transformation. To further explore the co-mutation relationships of these mutations, we constructed an association network of co-mutated genes by looking for statistically co-mutated genes in FClones (i.e., genes from both germline and somatic mutations have been considered) of tumor samples (Fig 8, Methods). We found that the probability of somatic-mutated genes in a FClone has some relations with the co-mutation patterns of the pre-existing germline mutations. For example, if genes A, B, C and D have been mutated in germlines, gene E has high chance to be somatic-mutated in the paired FClones.

We further analyzed the mutated genes on the survival network sample by sample. For each sample, we mapped its mutating genes from germline and founding clone onto the luminal-breast cancer survival network (Method) to analyze the network features of the mutating genes. Mutated genes from the same patient are not direct neighbors (i.e. they do not have any interactions) of each other in the network (P=0.020 and P=0.13 respectively, for luminal-breast and ovarian cancer, randomization tests). Somatic mutating genes are significantly not hubs (i.e., top 10% of high-link genes, P=0.017 and P=0.13, respectively, for luminal-breast and ovarian cancer, randomization tests).

We extended all the analyses in this section to the validation samples of the luminal-beast and ovarian tumors. Similar results were observed suggesting that these features are reproducible and robust. Since our purpose is to predict key, but not all,

mutated genes, based on these results we developed a procedure for prioritizing predicted candidates which contains 3 components: multiple SGSs-based, co-mutated gene association-based and, finally, network feature-based gene prioritizations (Methods, Fig 8). For a sample, each SGS can be used to predict a set of potentially mutated genes. Multiple SGSs-based gene prioritization relies on an assumption that if a predicted gene appears in multiple SGS-derived predicted gene sets, then there is a higher chance that its mutation is necessary for carcinogenesis. Co-mutated gene association-based gene prioritization is based our finding that somatic-mutations in FClones rely on the co-mutation patterns of germline mutations of an individual. Network feature-based gene prioritization relies on our finding that FClone somatic-mutations are unlikely to be the network hubs and are rarely network neighbors to other mutated genes in the same FClone.

Validate eTumorMonitor in breast cancer

For each validation sample, we ran DTNP and the procedure for prioritizing predicted candidates and predicted 50 genes whose mutations are required for malignant transformation based on its germline mutations. We then compared the predicted genes with the actual somatic-mutated genes in their paired FClone. In total 130 luminal-breast cancer samples from TCGA were used for the validation. There are

10, 8 and 6 genes, respectively, with a recall rate (i.e., the fraction of the testing samples have been predicted) of 38.1%, 85.2% and 98.0%, respectively. In other words, compared to random, predictions, these have been enriched by 19, 15.2 and 11.3 fold, respectively, with a recall rate of 38.1%, 85.2% and 98.0%, respectively. In Figure 9 shows the accuracy and power for our validation set. We assume that the probability of 6-10 genes to all have somatic-mutations in the same patient to be very low, therefore, the detection of 6-10 mutated genes (i.e., beyond the germline mutations) among the 50 predicted genes in the blood ctDNAs of an individual would a strong marker that an early-stage luminal-breast tumor is present. ctDNAs levels are very low in the blood, especially when tumors are at early stages. We expect that high-coverage sequencing (10,000X) of prospectively ultra-targeted exomes of the predicted genes in blood ctDNAs could detect early tumors.

Develop and validate eTumorMonitor in ovarian cancer

To demonstrate that eTumorMonitor can be extended to other cancer types, we randomly selected 100 ovarian cancer samples from TCGA as a training set and constructed an ovarian cancer version of eTumorMonitor (i.e., eTumorMonitor-ov) using the same procedure which has been used for constructing eTumorMonitor-lbc. To

validate this algorithm, we used 120 samples which have not been included in the training set. The results were similar to those obtained in luminal-breast cancer.

3.3 Discussion

Worldwide, there were 14.9 million new cancer cases and 32 million people living with cancer (within 5 years of diagnosis) in 2013 worldwide[112]. Over the past 40 years, cancer-related deaths have increased by 38%, accounting for 8.2 million deaths globally each year[112]. Furthermore, it is estimated that cancer deaths will increase to 13 million by 2030[113]. In many cases, cancer is not diagnosed until cancer cells have already invaded surrounding tissues and metastasized distant organs. For those patients, most conventional therapeutics are limited in their success. In fact, therapy-based cancer outcome improvements have been modest over the past few decades. Because a majority of cancers are diagnosed at the middle- or late-stages (i.e., advanced cancer) whereby cancers have spread and become incurable, the overall survival of patients has not been significantly improved. In fact, survival rates for people diagnosed with advanced cancer have changed little over the past four decades (http://seer.cancer.gov/). On the other hand, data suggest that the early detection of cancer is crucial for its ultimate control and prevention[112].

Tumor circuiting cell-free DNA (cfDNA) in the blood has been recognized for several decades. ctDNA has been used as a "liquid biopsy" to monitor the response to treatment and relapse. Furthermore, highly-sensitive, targeted deep sequencing of plasma has been developed to detect early-stage cancer: a recent investigation with various cancer types showed that ctDNA can be detected in more than half of the early-stage tumors that had not spread beyond the initial site[114]. Combining with targeted deep sequencing of plasma, eTumorMonitor could provide an important step towards the application of targeted genome sequencing technology to monitor healthy individuals identified with a high-risk of developing cancer. For example, for a given healthy individual who is at high-risk of developing breast cancer because of her familial history, eTumorMonitor could predict ~30 key somatic mutating genes which would be required to drive malignant transformation in the context of her germline mutations. We could prospectively sequence selected genes from her plasma samples every two years to monitor the appearance of mutations in these predicted 30 genes. If 8-9 genes out of the predicted 30 genes are shown to be functionally mutated (i.e., prediction rate is 30%), this could be indicative of early breast cancer which would warrant deeper evaluations. The chance that 8-9 genes among the 30 genes would all be mutated in the plasma of a single individual is very low. Therefore, this approach would be expected to have a low false-positive rate. Furthermore, the predicted mutating genes are expected to be founding clone mutations, consequently they will be

present in every cells of the tumor, and thus would be easier to detect than subclone-specific mutations. To improve the detection of early cancer, we could combine eTumorMonitor with other screening methods (i.e, clinical data such as a detailed family history, imaging modalities, some existing and newly developed screening approaches for breast, lung, liver, colon, cervical and pancreatic cancers[112]. This improved sensitivity and specificity could overcome the potential problem of overdiagnosis.

eTumorMonitor overcomes the difficulty in constructing predictive models from genome sequencing data by transforming mutations into network-derived profiles. The second important concept in eTumorMonitor is the super-gene signature which can represent the profiles of subgroups of both tumor founding clones and normal samples. To identify super-gene signatures, we developed a novel deep-mining approach, DANP which uses 1 million re-sampled gene sets from a cancer hallmark gene group to classify the network-derived profiles of tumor founding clones and normal samples into subgroups which are then compared. This approach allows super-gene signatures to be more accurate and robust. The third important concept is simulating the addition of extra mutations in the pre-exiting germline mutations on the network so that the super-gene signature profile of a germline sample becomes more similar to the profile of that super-gene signature of a subgroup of tumor founding clones. We developed a novel network perturbation approach, DPNP which links network perturbation and network-

derived profiles. These key novel concepts and methods form the core components of eTumorMonitor, so that it is able to predict cancer evolutionary mutations in advance, an effort that has been rarely attempted in the past. eTumorMonitor could provide an efficient tool for personalized early-cancer diagnostics for healthy individuals who are at high risk of developing cancer. At moment, several dozens of cancer predisposition genes linked to high or moderate cancer risk have been documented[115]. Furthermore, germline genomic landscapes are able to more accurately predict who has a high risk of developing which types of cancer (Zou et al., unpublished data).

The uniqueness of this survival network construction method is that experimentally determined luminal breast cancer specific survival and proliferation genes were used. These genes are highly related to luminal-breast cancer cell proliferation. By doing so, we removed a lot of noise from the network which improved the accuracy and robustness of our network-based predictions. For example, our use of luminal- and basal-specific breast cancer survival networks allowed us to predict luminal- and basal-specific drug targets with 80% accuracy based on experimental validations[108].

It is increasingly obvious that tumor evolution is not random but convergent. This point has been demonstrated in an early network study[90] which showed that mutated genes form signaling subnetworks and recent tumor sequencing studies showed that mutations are enriched in a handful of signaling pathways[24,88]. Therefore,

we proposed that germline genomic landscape constrain tumor evolution so that we should be able to predict the next evolutionary move required for malignant transformation. Similarly, we can hypothesize that somatic mutations are also not mutated blindly, but are selected to act together in a complementary manner with the pre-existing mutations of a cancer survival network to trigger malignant transformation and promote clonal expansion[32]. Furthermore, that multiple cancer-driving factors could have thousands of combinations where new driver-mutating genes work with the pre-existing germline mutations and finally trigger a limited number of cancer hallmark networks which transform normal cells into cancer cells[32]. Germline genomic landscapes from individuals are very different from person to person, therefore, it is expected that the complementary somatic mutations that drive malignant transformation would also be very different between distinct tumors. These results are in agreement with genome sequencing efforts which revealed that each tumor has a unique mutation landscape. Of note, although germline genomic landscapes constrain tumor evolution, other factors such as epigenetic and environmental regulators also exert constrains on tumor evolution. It is expected that the prediction performance of eTumorMonitor could be significantly improved by the addition of such factors in future models.

In summary, our efforts have been to predict in advance a tumor's next evolutionary move during cancer evolution by seeking probability patterns which are

somewhere between "random" and "deterministic". This is a something which has not been explored and used in the community. Predicting a tumor's next action allows for less invasive and less morbid therapeutic options. For example, eTumorMonitor could be used to detect early tumors so that the patients could be cured by surgery. Similar algorithms could be developed for foreseeing the acquisition of drug resistant mutations in advance so that new therapeutic approaches could be applied to forestall the next evolutionary move. By doing so, a paradigm shift would shift clinical practice to react to tumor before it changes. This is very different from current practice of reacting to a tumor's change after it has occurred. Predictions derived from cancer hallmark network-based modeling could ultimately be used in diagnosis, optimized patient management and prevention of cancer.

The network propagation algorithm works by projecting the functionally mutated genes of a sample onto a cancer type-specific metastasis network in which each functionally mutated gene is represented as a heat source. The heat source diffuses to neighboring genes along the edges of the network. This process is analogous to heat diffusion. At a certain time point, the diffusion stabilizes, and finally each gene on the network receives a certain amount of 'energy', which is represented by a 'heating' score. Here, instead of using gene heating-scores in networks for topological analysis, we used them differently: Once we obtained gene heating-scores, we didn't consider the networks anymore and extracted only the genes and their heating-scores to form a GR-

profile (i.e., similar to a gene expression profile). We called this process as network profiling which transforms network data into a profile-based data. Network profiling provides an advantage: Alternatively, deep learning algorithms could be applied to profile-based data rather than network-based data. A GR-profile represents the collective effect of the functionally mutated genes from a sample on the cancer metastasis network. Because the heating score is generated in the context of a functional mutations and gene relations of the metastasis network, these scores could represent the underlying molecular regulatory effects of mutations on the metastasis process.

3.4 Experimental procedures

Variant calling and identification of functional mutations

We applied the GATK pipeline for data pre-processing as described in our previous work[108]. All the variants were identified using Varscan2. Samples with a purity greater than 70%, as determined with absCNseq[116], were retained for downstream analyses. Based on tumor purity, sequencing reads from each variant were adjusted accordingly, and then the VAF (Variant Allele Frequency) was recalculated. To determine the mutations in FClones, only mutations in 2n regions of chromosomes (i.e., not in amplified and deleted regions), which were obtained from the segmental files of tumors from TCGA, were considered. Germline mutations included (1) a homozygous

mutation whose VAF is greater than 90 in both normal and tumor samples; (2) a heterozygous mutation whose VAF is 55 >= N >= 45 in normal samples. FClone somatic mutations included (1) a homozygous mutation whose VAF is >=90 in tumor but not in its paired normal sample; (2) a heterozygous mutation whose VAF is 55 >= N >= 45 in the tumor sample but not in its paired normal sample. Finally, functionally mutated genes were determined using VEP[117], CRAVAT[118] and MutationTaster2[119].

Construction of luminal breast cancer-specific survival network

We constructed a luminal breast cancer-specific survival network using our procedure developed previously[108]. Briefly, we collected cancer survival and proliferation genes (i.e., Set 1) from the genome-wide shRNA knock-down data in luminal cancer cell lines. We further identified potential cancer gene regulators (Set2) by combining gene expression values and copy number data (SNP 6.0 data) from the 150 training samples of luminal tumors in TCGA (see Method[108]). These regulators and FClone somatic-mutated genes of the 150 luminal tumors were defined as Set 2. We mapped all the genes from Sets 1 and 2 onto the protein interaction network[120] and extracted their links to obtain a luminal-breast cancer specific survival network.

The uniqueness of this survival network construction method is that experimentally determined luminal breast cancer specific survival and proliferation

genes (Set 1) were used. These genes are highly related to luminal-breast cancer cell survival and proliferation. By focusing on these genes, we removed a lot of noise in the network so that network-based prediction yields better results. For example, by constructing luminal- and basal-specific breast cancer survival networks, we predicted luminal- and basal-specific drug targets, which have 80% prediction accuracy based on experimental validations[108].

Generation of NetProfiles via network profiling

To generate a NetProfile for a sample, we projected its mutated genes as seeds on the cancer survival network and then applied a network propagation approach[110] to obtain heating scores of the network genes. For each patient, we generated NetProfiles for both germline (i.e., germline mutated genes as seeds) and paired tumor FClone (i.e., seed genes include germline and somatic mutated genes of the FClone). We combined all the NetProfiles and then conducted data transformation using the median centering and z-score between sample approach[92]. The resulting data is called the training set.

Identification of super-gene signatures via discriminating and -aligning of NetProfiles (DANP)

In the training set, there are 150 FClones' NetProfiles and their paired germlines' NetProfiles. We first conducted t-tests between germline-and FClone-NetProfiles to identify the genes whose heating scores are significantly higher in FClones than in germlines ($P<0.05$). From these genes, we extracted cancer hallmark-associated genes based on GO annotation[92]. For each cancer hallmark gene group, if it contains more than 40 genes, we will used that gene group to identify a SGS by applying DANP. From the training set, we then generated 200 random datasets (RDSs) of germline-NetProfiles, each containing 60% of randomly picked germline-NetProfiles from the training set. For the germline-NetProfiles in each RDS, we will added their paired FClone-NetProfiles. Meanwhile we will generated 5 (or 2, 8 and 10) million random gene sets (ie., RGS, each contains 30 randomly-picked genes from the GO-defined cancer hallmark gene group). Finally, we conducted fuzzy clustering (k=2) for the germline- and FClone-NetProfiles, respectively, of each RDS using each RGS.

For a given RDS, if a germline subgroup defined by a RGS has more than 85% (or 75%) patient overlap with a FClone's subgroup defined by the same RGS, meanwhile, and the Fisher's test of the patient overlap between germline- and FClone- subgroups reaches $P<.05$, then we assumed that that RGS significantly classifies the RDS into 2

consistent subgroups between the germline and FClone groups. If it happens onmore than 90% of the 200 RDSs, we collected that RGS and called it a passed gene-set. For a given cancer hallmark-based GO term, if the number of the passed gene-sets is greater than 1,000 but less than 5,000 (for 5 million RGSs, P <.001), we ranked the genes based on their frequency among the passed gene-sets in a descending order. The top 30 ranked genes are considered as the SGS representing that cancer hallmark. Previously we have conducted a simulation study to show that the top 30 genes are sufficient to produce a robust gene signature[92]. The SGS was validated using the validation set including 250 luminal-breast samples.

To determine an optimal N for subgroups of both germlines and FClones, we applied DANP for identifying SGSs by classifying germlines and FClones into 3, 4, or 5 subgroups. The optimal subgroup number (N) was determined such that each germline subgroup shares the highest percentage of its patients to the corresponding FClone subgroup. For example, if optimal subgroup number N is 2, more than 85% of the patients are shared between FClone subgroup 1 (or 2) and germline subgroup 1 (or 2), respectively. When N is 3, 4, or 5, patients in common between FClone subgroup and its corresponding germline subgroup is lower than 85%. Once an optimal N is determined, the SGSs which classified the germlines/FClones to N subgroups are the SGSs which was used for DTNP.

Prediction of personalized mutated genes using the driving targeted-tumor profiles via network perturbation (DTNP)

For a healthy woman who is at high risk of developing luminal-breast cancer, we will first used an SGS to assign her to germline subgroup 1 or 2, based on the correlations between her germline SGS's profile (i.e. by generating her germline NetProfile and extracting heating scores for the SGS's genes) and the centroids of SGS for the 2 germline subgroups (see Method[92]). Assume that she has been assigned into germline subgroup 1, its expected FClone's GS profile will be represented by the centroid of the GS for FClone-NetProfiles subgroup 1. For a gene (M) in the network which is not germline mutated, we let it mutated *in silico* and then generated a new profile for the GS (i.e., projecting M and germline mutated genes together as seeds on the network to get a NetProfile and extracting heating scores for the GS genes from it). We calculated the correlation between the resulted GS profile and the GS centroid for FClones. We extended this analysis for every network gene that was originally germline mutated in that sample. These were then ranked based on their correlation coefficients. Our purpose was to predict key mutated genes but not all mutated genes, therefore, we only selected highly ranked genes as potential candidate genes for mutation. To identify such genes, we ordered gene list $\mathcal{M}$=[M_1, M_2, …, M_n] be in a descending order according to the calculated correlation coefficients Cor=[Cor_{NS+Mi}, i = 1 … n], where n is the total

number of candidates. Then, the key mutated genes (KGs) were defined by two ways depending on whether a turning point (Cor_T) for dramatically falling of the correlation on Cor exists or not:

$$KG = \begin{cases} \{M \in \mathcal{M}: Cor_{NS+M} > Cor_T\} & if\ Cor_T\ exists \\ \{M_1, M_2, \ldots, M_K\}\ where\ K = \text{minimum}(\{k \in \mathbb{Z}_+: \sum_{i=1}^{k} D_{NS+M_i} > \sum_{j=k+1}^{n} D_{NS+M_j}\}) & otherwise \end{cases}$$

Top 50 genes were extracted from this list.

Predicted gene prioritization

Multiple SGSs-based gene prioritization: For a given sample, based on a SGS, we could predict 50 mutated genes. We have identified 8 SGSs for modeling the malignant transformation for luminal-breast cancer. Therefore, we could predict 8 sets of mutated genes (50 each). We assume that genes, which are common in two or more predicted gene lists, could have a high probability to be mutated for driving malignant transformation. Therefore, for a given sample, we scored the predicted genes based on their frequency in the predicted gene lists derived from 8 SGSs. For example, we assigned score of 2 or 4 if a gene appeared in 2 or 4 predicted gene lists.

Co-mutated gene association-based gene prioritization: To apply this approach, we first constructed a co-mutated gene association network. To do so, we listed all the mutated

genes (germline- and somatic-mutated genes) of each FClone of the training set. For each mutated gene pair regardless of germline- or somtic-mutated, we conducted Fisher's test to examine if they were co-mutated among these samples ($P<0.05$). Significantly co-mutated genes were linked to form a co-mutated gene association network. For predicted candidate gene by DTNP, if it interacted with more than N germeline mutated genes, that gene would be removed from the candidate list. N= 5 + 2*(number of germline mutated genes in the sample)/1000.

Network feature-based gene prioritization: Finally, we will integrate network features of somatic-mutated genes to prioritize the predicted genes. By analyzing germline and somatic mutations in each FClone of the training samples on the cancer survival network, we found that mutated genes in a FClone are unlikely to be network neighbors each other (P=.02). Somatic-mutated genes are significantly depleted from network hubs (P=.01, top 10% of the high-linked nodes in the network). Of note, these network patterns of mutated genes from a single FClone are different from those of the mutated genes collected from multiple samples. Finally, we mapped all the germline mutated genes of a sample onto the survival network. We removed a predicted gene if it was a network hub or a network neighbor of a germline mutated gene. Finally, If two predicted genes are network neighbors, the one which has a lower score was removed.

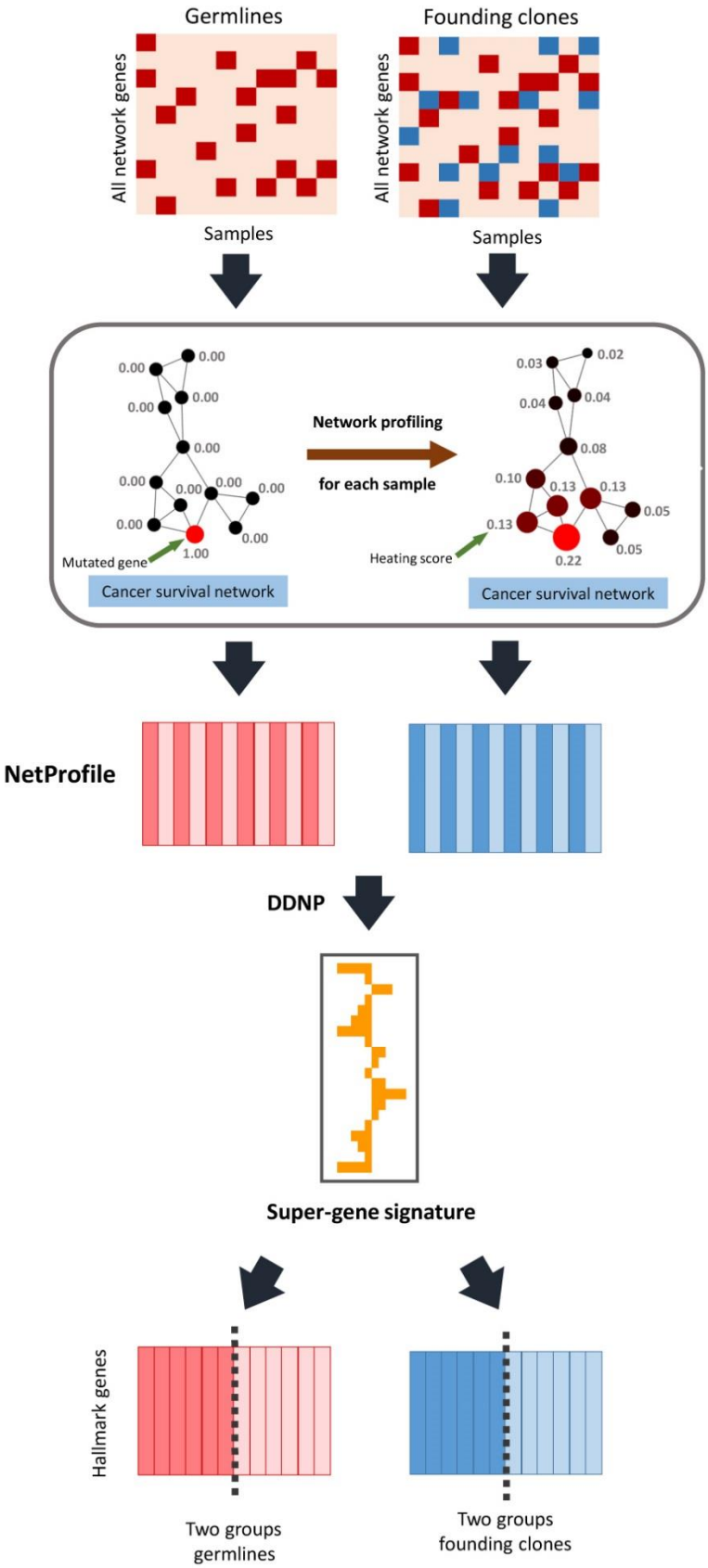
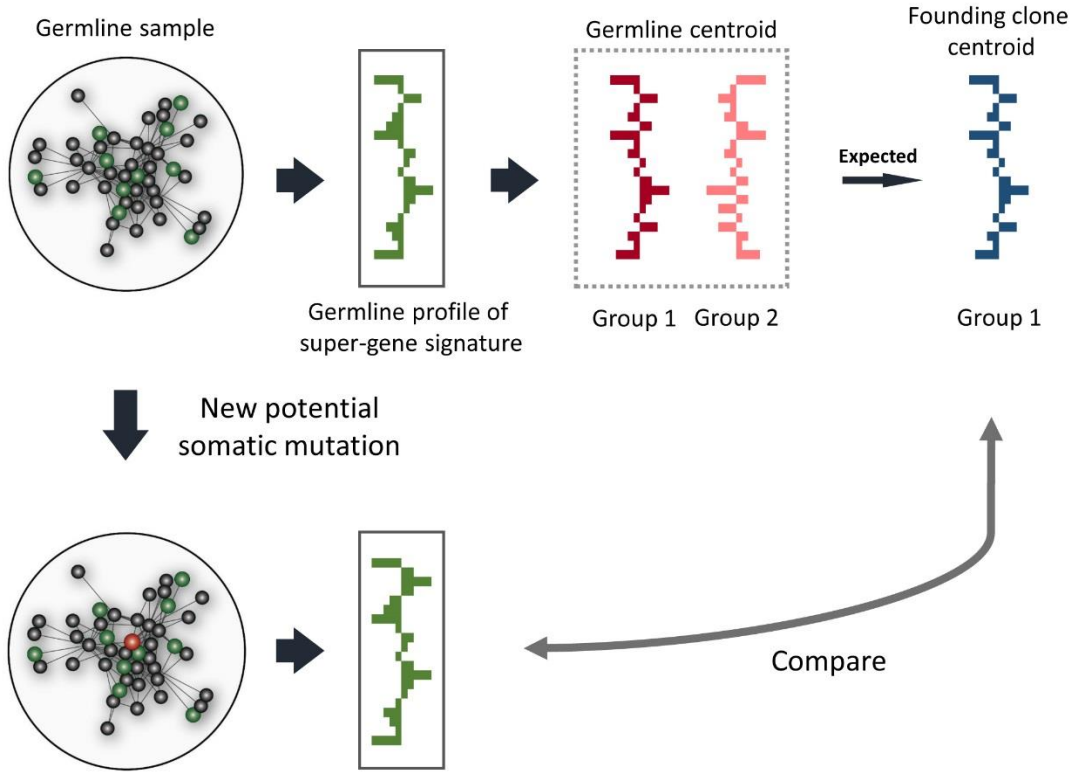

Germlines
Founding clones
All network genes
All network genes
Samples
Samples
Network profiling
for each sample
Mutated gene
Heating score
Cancer survival network
Cancer survival network
NetProfile
DDNP
Super-gene signature
Hallmark genes
Two groups
germlines
Two groups
founding clones

Figure 5:

From top down, the germline mutations (red squares) and somatic mutations (blue squares) for each sample was first propagated and then DDNP extracted the SGS which can then separate the germline and founding samples into two corresponding groups.

A


Germline sample
Germline profile of super-gene signature
Germline centroid
Group 1
Group 2
Expected
Founding clone centroid
Group 1
New potential somatic mutation
Compare

B

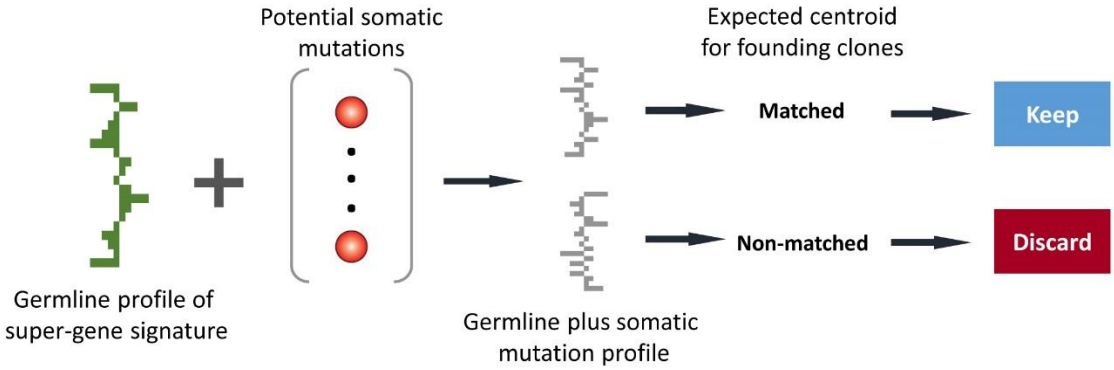

Potential somatic mutations
Germline profile of super-gene signature
Germline plus somatic mutation profile
Expected centroid for founding clones
Matched
Keep
Non-matched
Discard

Figure 6:

(A) A new germline sample's mutations (green balls) are used to create a germline profile which is then matched to a germline centroid. Next we can determine which new somatic mutations can bring the new sample's germline profile closer to the Founding clone centroid.

(B) One by one, each potential somatic mutations are added to the new sample's germline profile and it is accepted if it matches to the corresponding Founding centroid or discarded if does not match.

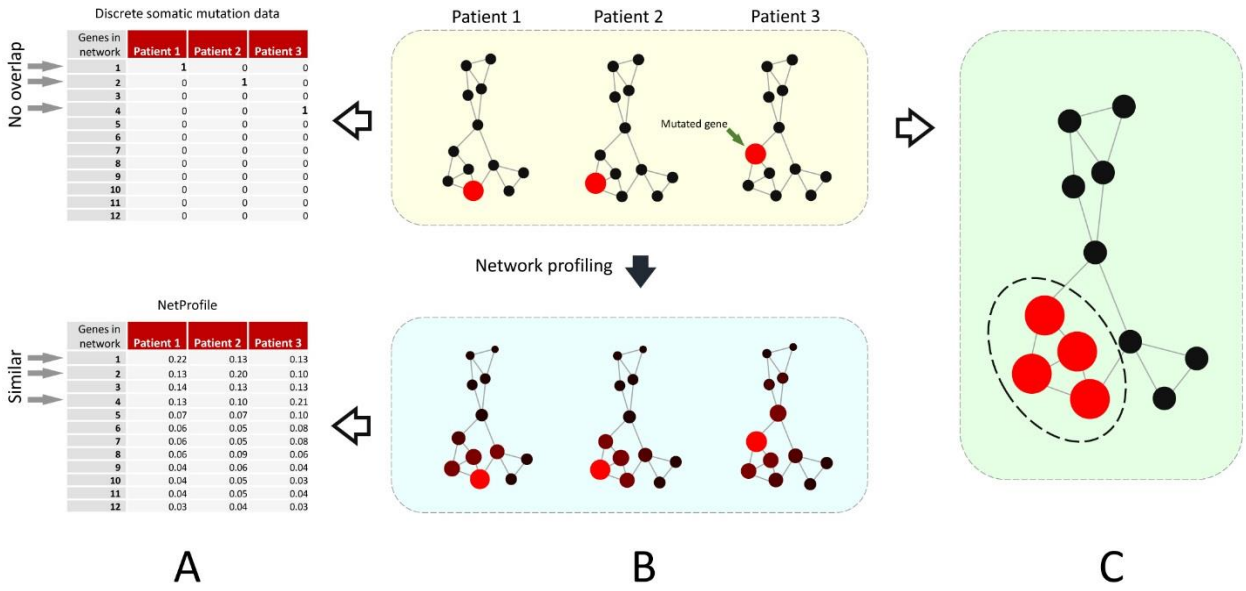
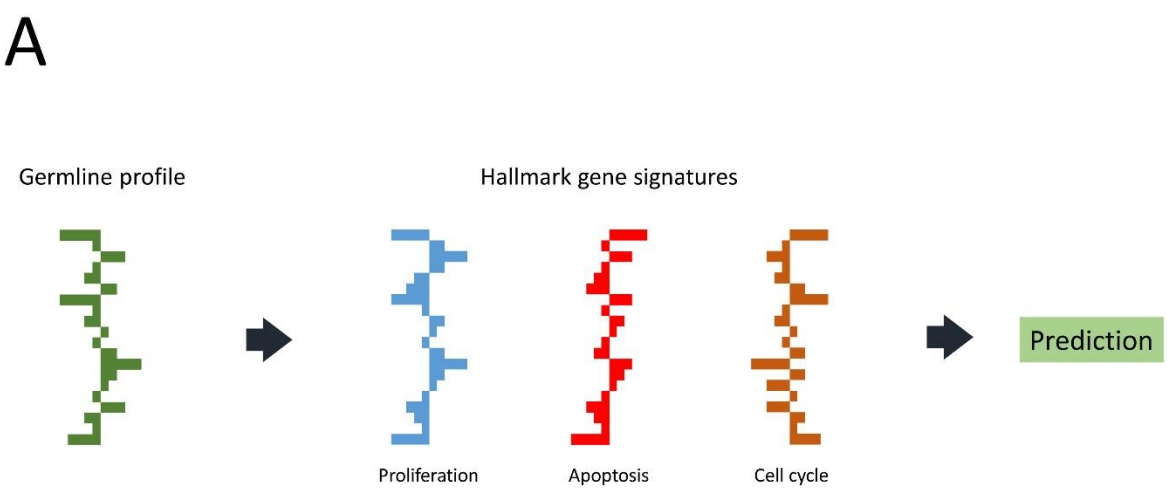


Figure 7:

(A) Discreet somatic mutations across samples have no overlap. After network profiling there is similarity among groups of genes which are mutated across samples.

(B) Graphical representation of (A)

(C) Different genes may be mutated across samples, but they target a similar group of genes that may have similar function or role.


A
Germline profile
Hallmark gene signatures
Proliferation
Apoptosis
Cell cycle
Prediction

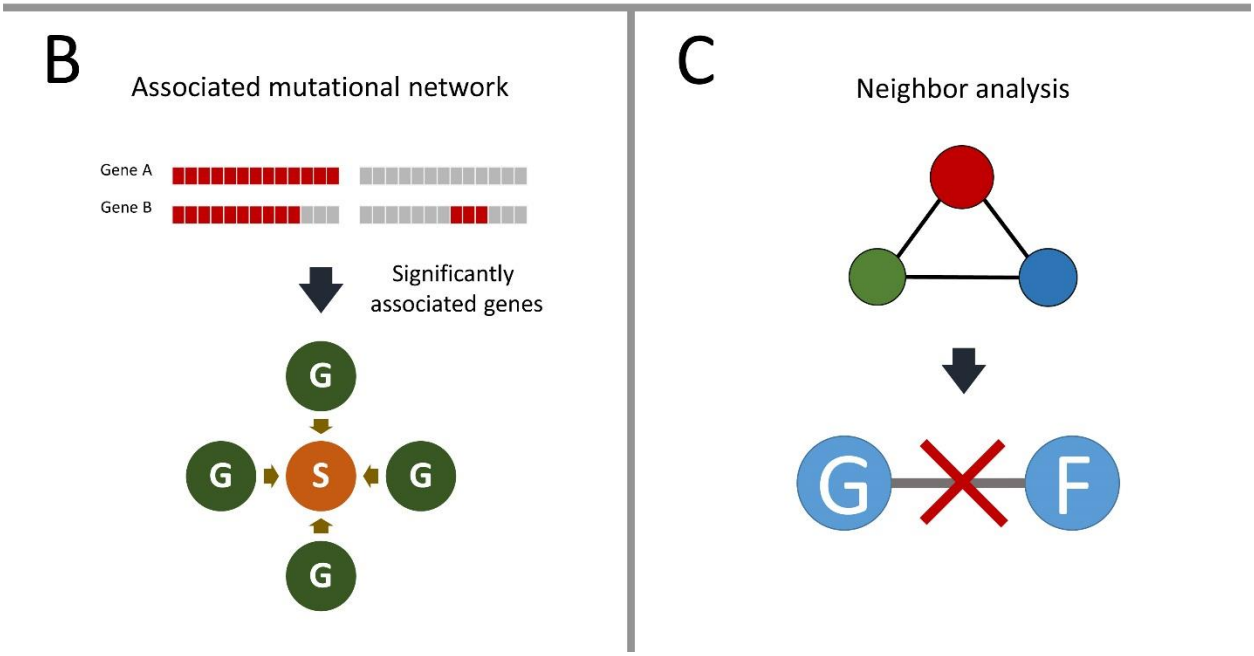

B
Associated mutational network
Gene A
Gene B
Significantly associated genes
G
G
S
G
G
C
Neighbor analysis
G
F

Figure 8:

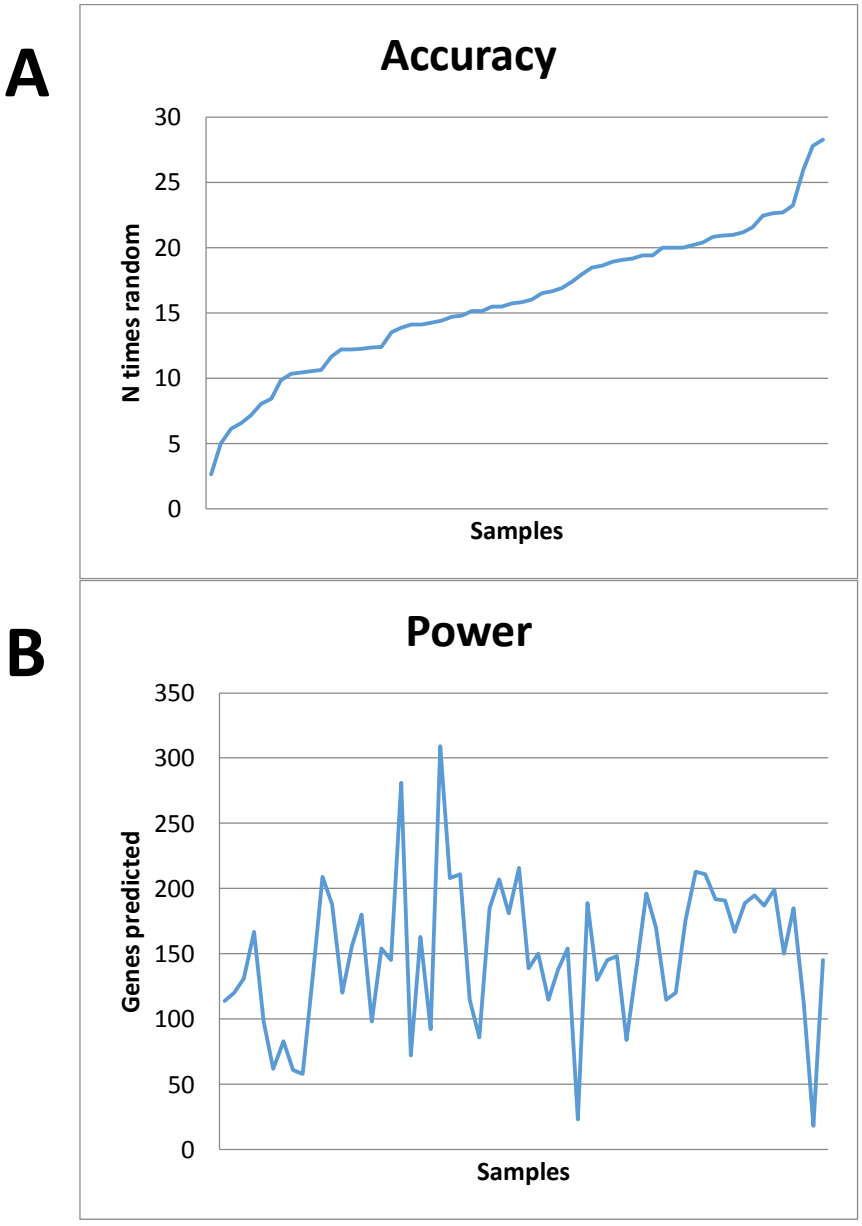
eTumor Monitor gene prioritization

(A) Different cancer hallmark SGS are used to predict a set of potentially mutated genes. The more SGS a gene overlaps with the higher the priority.

(B) Potentially mutated genes that are significantly ($P < 0.05$) co-mutated with germline mutations are considered otherwise they are discarded.

(C) Potentially mutated genes that share a high number of interactions with germline mutations are discarded

Figure 9:



Accuracy and power of our model

(A) The percentage of genes we predicted accurately. Randomly we can identify only 1% of genes accurately.

(B) The number of genes we can predict per sample for our validation set.

CONCLUSION

With new technologies and algorithms arising, it is an exciting time to be doing cancer research. With the integration of various datasets and computational modeling, it is now possible to step out of the paradigm of looking at one gene or one pathway and to look at cancer as a dysregulated system. The works highlighted in this thesis have shown how a systems biology approach can be used to identify subtype-specific breast cancer drug targets[108] and key somatic mutations that represent the first steps needed to transform a normal cell into a cancerous cell.

Precision medicine is the holy grail of the cancer genomics systems biology field. The high heterogeneity of cancer cells within a tumor and sparse overlapping mutation frequencies across patients limits our ability to fully translate our knowledge into improvements in cancer treatment. Currently, there is a high demand for the development of methodologies that can tackle these problems. In Chapter 2 we identified subtype specific drug targets for breast cancer. We were able to recapture well known luminal and basal subtype-specific genes as well as identify new genes. We observed the convergence of different types of genetic alterations (e.g. copy number variation and point mutations) and how they elude researchers that are only looking at

point mutations or CNV. Nevertheless, it is important to note that there are limitations to our approach. Our protein interaction network, which is the backbone of this work, is not complete since it accounts for less than half of the proteins encoded in the human genome. The case is similar for the RNAi screening data which is based on a methodology that is already obsolete. Most notably, RNAi do not fully inactivate its target genes rather reduces its protein levels. In addition, they have significant off-target effects. In the future, CRISPR/CAS9 technologies will be used to carry out these types of functional screens. These technologies are much more precise and versatile, capable of not only inactivating specific genes, but also of modulating their expression. Incorporation of CRISPR/CAS9 datasets should greatly improve the accuracy of our networks and to overcome the shortcomings of RNAi.

The fewer the types of data required for a method as input, the better. Gene expression, copy number variation and other genomic alteration measurements of a cancerous cell are not easy to measure in a clinical setting. It becomes even more difficult when the genetic heterogeneity of the tumor is taken into account. This is why, in Chapter 3, we only used exome sequencing data as the main source of data for our modeling. This type of modeling is where future research should be headed.

The objective in Chapter 3 was to identify a patient-specific geneset that could be used by clinicians to monitor cancer development and the acquisition of new relevant gene mutations. While whole-genome and exome-sequencing costs are still dropping, it

is still perhaps a decade away, or more, before the majority of the hospitals in North America can afford to sequence samples for every patient. However, targeted sequencing of a small geneset (~30-50) is much more affordable and could be adopted at a much faster pace. Currently, this geneset can be monitored from the circulating DNA of dead cancer cells isolated from the blood samples of patients. However, in the future, sequencing may be something that will be carried out from the comfort of your own home, away from the hospital setting. A simple software that processes and analyzes this information, could help patients monitor themselves and suggest when a visit the doctors might be useful.

In both Chapter 2 and 3, using a wide array of dataset and computational techniques we have modeled cancer in a holistic manner. However, there are other aspects that influence cancer development which, for technical reasons, could not be taken into account. These include the cancer heterogeneity, microenvironment and epigenetic factors.

The stumbling block in trying to incorporate these other factors is the lack of data. In a tumor it is difficult to identify how many sub-populations of cancer cells there really are and which mutation belongs to which sup-population due to the heterogeneous nature of tumors. The more accurately we can identify the key mutations that are important for each of the sub-populations the better we can choose which

therapeutics to use. Different analytical tools are being developed to improve the cancer heterogeneity issue.

Another important area we need data on is the tumor microenvironment. This is difficult because it is not clear cut where the microenvironment starts and ends in a patient, how heterogeneous the microenvironment may be, and whether other parts of the host's system further away may also play a role. Currently, for the vast amount of tumor samples and their matched normal that we have, we do not have data on their matched microenvironment.

Environmental and epigenetic factors no doubt are responsible for the development of cancer. However, to measure and quantify these factors are perhaps the most difficult. To keep a track of a patient's daily activities, behavioral patterns, and emotional status is not only not possible, but it is also time sensitive. However, to identify which environmental factors to look out for is even a bigger challenge. The variabilities from one patient to another vary greatly due to lifestyle choices. There are some large datasets that have the contributions of epigenetics (e.g. methylation) to carcinogenesis, but more factors need to be included.

In our work, we have selected a few cancer types to show how a systems biology approach can better help understand cancer. However, our analysis was developed so it may be applicable to multiple cancer types. New projects at the NRC are currently

applying the methodologies developed in our work for other cancer types such as lung and colon cancer and have shown similar results to the ones presented in this thesis.

Our work on luminal and basal breast cancer subtypes, in Chapter 2, was the most accurate in predicting subtype-specific drug targets and the first paper to include RNAi screening data. Our work in Chapter 3 that tries to predict somatic mutations that may give rise to cancer in a patient based on their germline mutation is one of a kind. In the field of cancer research, the vast majority of research projects have as a goal the identification of mutations that are specific to a cancer type. The reasoning behind why we have not been able to identify these cancer type-specific mutations is because we have not sequenced enough samples. However, we believe that it is the combination of pre-existing germline variants with new somatic mutations that give rise to cancer. Concentrating only a newly-acquired mutations cannot provide the whole picture. To our knowledge, our approach of predicting somatic mutations based on germline variants is unique.

The works in this thesis have shown the advantages of using a systems biology approach. It is multidisciplinary, integrative and demands scientists from different backgrounds to share ideas to solve problems. It is a new field and one of its shortcomings is that the datasets are not always complete (i.e. protein interaction network). Hence, it is possible to draw incomplete conclusions. However, with an ever

increasing number of platforms and data generation, its power and accuracy will only increase over time.

Deep learning is quickly becoming "the next big thing" in artificial intelligence. It has been shown to "learn" or recognize patterns on its own using neural networks (which is completely different from protein interaction networks). Recently, a deep learning-based algorithm has beaten the world campion of the board game Go[121]. They have also shown great success in the image and voice recognition areas as evidenced in todays cameras that recognize faces and cell phones that respond to voice commands. Unlike traditional computational methods which calculate every possible outcome of every permutation, a neural network learns, in a similar fashion, the way a human brain does, by optimizing internal parameters to recognize patterns via computer "intuition" or experience. The datasets required for deep learning are vastly greater than what is currently used in biology, but with more genetic datasets being generated every day, deep learning could become a necessity. While it is making a huge splash in the computer science field, it will not be long before it has a similar impact on biology, as the two are more integrated now than ever. It could allow us to identify the impacts of genetic alterations that are both cancer-specific and personalized.

Smart phones and the app world may have changed our lives for the better or worse, but what they have also done is to generate piles of lifestyle data. There are now apps for nearly everything and many of them are health-related, such as measuring

your steps, calories, heartbeat, sleeping pattern and various other things. New methodologies will be required to be able to integrate so many different parameters from these datasets in a meaningful way, and to understand how they relate to disease progression. Here again, the precision of the methods could benefit from genomic information.

The development of new algorithms can produce new insights on cancer mechanism and improve patient care. Most significantly, they could help in early detection of cancer and to identify exactly which patients are most likely to respond to expensive new drugs. It is becoming increasingly possible to measure smaller and smaller quantities of DNA from the blood, some of which is released as a result of cancer cells deaths. Patients can be monitored this way to identify whether new somatic mutations are arising in their body which would be indicative of early stage cancer. In addition, given the heterogeneity nature of tumors resulting from multiple sub-populations of cancer cells, algorithms can help choose a combinatorial drug approach that can target the different sub-populations simultaneously and not give competitive advantage for the other sub-populations to grow afterwards.

We hope this work has shed some light on the advantages of a systems biology approach and how it may help to bring about a new era in personalized medicine.

References

1 Müller-Wille, S. Carolus Linnaeus *britannica*.
2 Bailey, R. Cell Theory. *About.com* (2015).
3 Nobel Lectures. *Elsevier Publishing Company* (1964).
4 Watson, J. D. C., F. H. C. . A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737-738 (1953).
5 Wang, E. *Cancer Systems Biology*. (CRC Press, 2010).
6 Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-1664, doi:10.1126/science.1069492 (2002).
7 Cancer Facts & Figures 2016. *American Cancer Society* (2016).
8 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).
9 *Cancer statistics at a glance*, <http://www.cancer.ca/en/cancer-information/cancer-101/cancer-statistics-at-a-glance/?region=on> (
10 *Are the number of cancer cases increasing or decreasing in the world?*, <http://www.who.int/features/qa/15/en/> (2008).
11 *The History of Cancer*, <http://www.cancer.org/cancer/cancerbasics/thehistoryofcancer/the-history-of-cancer-what-is-cancer> (2014).
12 Russell, W. An Address on a Characteristic Organism of Cancer. *Br Med J* **2**, 1356-1360 (1890).
13 Matthews, J. B. & Mason, G. I. Immunoglobulin producing cells in human periapical granulomas. *Br J Oral Surg* **21**, 192-197 (1983).
14 Livingston, V. W. & Alexander-Jackson, E. A specific type of organism cultivated from malignancy: bacteriology and proposed classification. *Ann N Y Acad Sci* **174**, 636-654 (1970).
15 Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311-1315 (1984).
16 Epstein, M. A., Achong, B. G. & Barr, Y. M. Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma. *Lancet* **1**, 702-703 (1964).
17 Durst, M., Gissmann, L., Ikenberg, H. & zur Hausen, H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proc Natl Acad Sci U S A* **80**, 3812-3815 (1983).
18 Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-173 (1976).
19 Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med* **2**, 54, doi:10.1186/gm175 (2010).
20 Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625-1629 (1994).
21 Miyaki, M. *et al.* Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat Genet* **17**, 271-272, doi:10.1038/ng1197-271 (1997).
22 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038 (1993).
23 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).
24 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).
25 Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40, doi:10.1016/j.cell.2010.11.055 (2011).
26 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat Genet* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

27 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).
28 Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med* **353**, 1659-1672, doi:10.1056/NEJMoa052306 (2005).
29 Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med* **353**, 1673-1684, doi:10.1056/NEJMoa052122 (2005).
30 Wang, E. *et al.* Cancer systems biology in the genome sequencing era: part 2, evolutionary dynamics of tumor clonal networks and drug resistance. *Semin Cancer Biol* **23**, 286-292, doi:10.1016/j.semcancer.2013.06.001 (2013).
31 Wang, E. *et al.* Cancer systems biology in the genome sequencing era: part 1, dissecting and modeling of tumor clones and their networks. *Semin Cancer Biol* **23**, 279-285, doi:10.1016/j.semcancer.2013.06.002 (2013).
32 Wang, E. *et al.* Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin Cancer Biol* **30**, 4-12, doi:10.1016/j.semcancer.2014.04.002 (2015).
33 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).
34 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).
35 Wang, E. Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Lett* **340**, 261-269, doi:10.1016/j.canlet.2012.11.050 (2013).
36 Yuan, T. L. & Cantley, L. C. PI3K pathway alterations in cancer: variations on a theme. *Oncogene* **27**, 5497-5510, doi:10.1038/onc.2008.245 (2008).
37 Pollack, J. R. A perspective on DNA microarrays in pathology research and practice. *Am J Pathol* **171**, 375-385, doi:10.2353/ajpath.2007.070342 (2007).
38 Bertucci, F., Finetti, P. & Birnbaum, D. Basal breast cancer: a complex and deadly molecular subtype. *Curr Mol Med* **12**, 96-110 (2012).
39 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).
40 Diaz-Cano, S. J. Tumor heterogeneity: mechanisms and bases for a reliable application of molecular marker design. *Int J Mol Sci* **13**, 1951-2011, doi:10.3390/ijms13021951 (2012).
41 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536, doi:10.1038/415530a (2002).
42 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).
43 Carlson, J. J. & Roth, J. A. The impact of the Oncotype Dx breast cancer assay in clinical practice: a systematic review and meta-analysis. *Breast Cancer Res Treat* **141**, 13-22, doi:10.1007/s10549-013-2666-z (2013).
44 Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).
45 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).
46 Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-273, doi:10.1038/ng1180 (2003).
47 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* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

48 Wetterstrand, K. *DNA Sequencing Costs*, <http://www.genome.gov/SequencingCosts/> (2016).
49 Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141, doi:10.1016/j.tig.2007.12.007 (2008).
50 Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46, doi:10.1038/nrg2626 (2010).
51 Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811, doi:10.1038/35888 (1998).
52 Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-437, doi:10.1038/nature02371 (2004).
53 Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov* **2**, 172-189, doi:10.1158/2159-8290.CD-11-0224 (2012).
54 Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J Bacteriol* **169**, 5429-5433 (1987).
55 Maeder, M. L. *et al.* Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Mol Cell* **31**, 294-301, doi:10.1016/j.molcel.2008.06.016 (2008).
56 Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* **39**, e82, doi:10.1093/nar/gkr218 (2011).
57 Wood, A. J. *et al.* Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307, doi:10.1126/science.1207773 (2011).
58 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).
59 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).
60 Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29-59, doi:10.1146/annurev.genom.7.080505.115623 (2006).
61 Pawson, T. & Nash, P. Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445-452, doi:10.1126/science.1083653 (2003).
62 Bader, S., Kuhner, S. & Gavin, A. C. Interaction networks for systems biology. *FEBS Lett* **582**, 1220-1224, doi:10.1016/j.febslet.2008.02.015 (2008).
63 Cusick, M. E., Klitgord, N., Vidal, M. & Hill, D. E. Interactome: gateway into systems biology. *Hum Mol Genet* **14 Spec No. 2**, R171-181, doi:10.1093/hmg/ddi335 (2005).
64 Berggard, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**, 2833-2842, doi:10.1002/pmic.200700131 (2007).
65 Kerrien, S. *et al.* IntAct--open source resource for molecular interaction data. *Nucleic Acids Res* **35**, D561-565, doi:10.1093/nar/gkl958 (2007).
66 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451, doi:10.1093/nar/gkh086 (2004).
67 Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res* **35**, D572-574, doi:10.1093/nar/gkl950 (2007).
68 Stark, C. *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704, doi:10.1093/nar/gkq1116 (2011).
69 Salwinski, L. *et al.* Recurated protein interaction datasets. *Nat Methods* **6**, 860-861, doi:10.1038/nmeth1209-860 (2009).
70 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

71 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).
72 Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* **37**, D619-622, doi:10.1093/nar/gkn863 (2009).
73 Prieto, C. & De Las Rivas, J. APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res* **34**, W298-302, doi:10.1093/nar/gkl128 (2006).
74 Tarcea, V. G. *et al.* Michigan molecular interactions r2: from interacting proteins to pathways. *Nucleic Acids Res* **37**, D642-646, doi:10.1093/nar/gkn722 (2009).
75 Chaurasia, G. *et al.* UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res* **35**, D590-594, doi:10.1093/nar/gkl817 (2007).
76 Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* **10**, 1108-1115, doi:10.1038/nmeth.2651 (2013).
77 Morgan, G., Ward, R. & Barton, M. The contribution of cytotoxic chemotherapy to 5-year survival in adult malignancies. *Clin Oncol (R Coll Radiol)* **16**, 549-560 (2004).
78 Maude, S. L. *et al.* Chimeric antigen receptor T cells for sustained remissions in leukemia. *N Engl J Med* **371**, 1507-1517, doi:10.1056/NEJMoa1407222 (2014).
79 Grupp, S. A. *et al.* Chimeric antigen receptor-modified T cells for acute lymphoid leukemia. *N Engl J Med* **368**, 1509-1518, doi:10.1056/NEJMoa1215134 (2013).
80 Porter, D. L., Levine, B. L., Kalos, M., Bagg, A. & June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *N Engl J Med* **365**, 725-733, doi:10.1056/NEJMoa1103849 (2011).
81 Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405-409, doi:10.1038/nature11154 (2012).
82 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404, doi:10.1038/nature11017 (2012).
83 Chin, L., Hahn, W. C., Getz, G. & Meyerson, M. Making sense of cancer genomic data. *Genes Dev* **25**, 534-555, doi:10.1101/gad.2017311 (2011).
84 Schlabach, M. R. *et al.* Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620-624, doi:10.1126/science.1149200 (2008).
85 Silva, J. M. *et al.* Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617-620, doi:10.1126/science.1149185 (2008).
86 Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235**, 177-182 (1987).
87 Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274, doi:10.1126/science.1133427 (2006).
88 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).
89 Rozenblatt-Rosen, O. *et al.* Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* **487**, 491-495, doi:10.1038/nature11288 (2012).
90 Cui, Q. *et al.* A map of human cancer signaling. *Mol Syst Biol* **3**, 152, doi:10.1038/msb4100200 (2007).
91 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**, 56-68, doi:10.1038/nrg2918 (2011).
92 Li, J. *et al.* Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat Commun* **1**, 34, doi:10.1038/ncomms1033 (2010).
93 Awan, A. *et al.* Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signalling network. *IET Syst Biol* **1**, 292-297 (2007).

94 Li, L. *et al.* The human phosphotyrosine signaling network: evolution and hotspots of hijacking in cancer. *Genome Res* **22**, 1222-1230, doi:10.1101/gr.128819.111 (2012).
95 Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A* **104**, 20007-20012, doi:10.1073/pnas.0710052104 (2007).
96 Wang, E., Lenferink, A. & O'Connor-McCourt, M. Cancer systems biology: exploring cancer-associated genes on cellular networks. *Cell Mol Life Sci* **64**, 1752-1762, doi:10.1007/s00018-007-7054-6 (2007).
97 Chan, K. C. *et al.* Cancer genome scanning in plasma: detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clin Chem* **59**, 211-224, doi:10.1373/clinchem.2012.196014 (2013).
98 Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* **4**, 162ra154, doi:10.1126/scitranslmed.3004742 (2012).
99 Murtaza, M. *et al.* Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108-112, doi:10.1038/nature12065 (2013).
100 Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575, doi:10.1038/nature11005 (2012).
101 Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci U S A* **109**, 2724-2729, doi:10.1073/pnas.1018854108 (2012).
102 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).
103 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).
104 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).
105 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).
106 Reimand, J., Tooming, L., Peterson, H., Adler, P. & Vilo, J. GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res* **36**, W452-459, doi:10.1093/nar/gkn230 (2008).
107 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).
108 Zaman, N. *et al.* Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep* **5**, 216-223, doi:10.1016/j.celrep.2013.08.028 (2013).
109 Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat Methods* **9**, 1069-1076, doi:10.1038/nmeth.2212 (2012).
110 Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* **6**, e1000641, doi:10.1371/journal.pcbi.1000641 (2010).
111 Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* **32**, 1202-1212, doi:10.1038/nbt.2877 (2014).

112 Global Burden of Disease Pediatrics, C. *et al.* Global and National Burden of Diseases and Injuries Among Children and Adolescents Between 1990 and 2013: Findings From the Global Burden of Disease 2013 Study. *JAMA Pediatr* **170**, 267-287, doi:10.1001/jamapediatrics.2015.4276 (2016).
113 Committee, A. C. P. R. W. *et al.* AACR Cancer Progress Report 2013. *Clin Cancer Res* **19**, S4-98, doi:10.1158/1078-0432.CCR-13-2107 (2013).
114 Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224, doi:10.1126/scitranslmed.3007094 (2014).
115 Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-308, doi:10.1038/nature12981 (2014).
116 Bao, L., Pu, M. & Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics*, doi:10.1093/bioinformatics/btt759 (2014).
117 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).
118 Douville, C. *et al.* CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* **29**, 647-648, doi:10.1093/bioinformatics/btt017 (2013).
119 Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-362, doi:10.1038/nmeth.2890 (2014).
120 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).
121 Gibney, E. Google AI algorithm masters ancient game of Go. *Nature* **529**, 445-446, doi:10.1038/529445a (2016).