Chapter 3
Identifying Driver Mutations in Cancer

Jack P. Hou and Jian Ma

Abstract A key question in cancer genomics is how to distinguish "driver" mutations, which contribute to tumorigenesis, from functionally neutral "passenger" mutations. Driver mutation is critically important for understanding the molecular mechanisms of cancer development and progression, which will ultimately help tailor more targeted and effective treatments for patients. In this chapter, we introduce recent developments in computational methods for identifying driver mutations. We summarize existing methods into several major categories and discuss challenges in discovering the whole spectrum of driver mutations in cancer for future computational and systems biology studies.

Keywords Cancer · Genomics · Driver mutation · Systems biology

3.1 Introduction

3.1.1 What is Driver Mutation?

Rapid advances in next-generation sequencing technologies have paved the way for comprehensive analysis for large numbers of cancer genomes (Stratton 2013). Through these advances, scientists have uncovered a large number of genetic mutations and other alterations (e.g., copy number changes, epigenetic changes,

J. P. Hou · J. Ma (✉)
Department of Bioengineering, University of Illinois, Urbana–Champaign, IL, USA
e-mail: jianma@illinois.edu

J. P. Hou
Medical Scholars Program, University of Illinois, Urbana–Champaign, IL, USA

J. Ma
Institute for Genomic Biology, University of Illinois, Urbana–Champaign, IL, USA

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_3,

and structural variations) pertaining to cancer (Green et al. 2011). To understand the significant alterations that cause cancer is to discover the source of carcinogenesis—information that we can utilize to improve treatments for patients. However, the complexity of cancer and the tremendous amount of genomic data remain a daunting obstacle for us to fully understand cancer mutations. Cancer cells may often exhibit hundreds upon thousands of different mutations and other alterations in its genome that affect a wide array of genes representing many diverse functions. However, the vast majority of these genes do not have a significant impact on tumorigenesis (Hanahan and Weinberg 2011). A key question in cancer genomics is how to distinguish "driver" mutations, which contribute to tumorigenesis (Greenman et al. 2007), from functionally neutral "passenger" mutations. Such driver mutations (e.g., point mutations or copy number changes) are critically important to elucidate key biological pathways that are perturbed in cells and eventually lead to proliferation, angiogenesis, or metastasis (Hanahan and Weinberg 2011).

Detecting driver mutations is necessary for understanding the molecular mechanisms of carcinogenesis. Determining the driver will also aid in verifying and discovering new prognostic and diagnostic markers in cancer as well as therapeutic targets for potential cancer drugs. Therefore, recently in the field of computational cancer genomics, many researchers have developed computational methods to identify driver mutations (Zhang et al. 2013). Overall, these methods have different underlying principles to achieve similar goals. We can group these different methods that identify driver mutations in cancer into four broad categories:

- Sequence-Based Approaches: methods that assess the functional impact a mutation has on the candidate driver gene and its protein product (Kumar et al. 2009; Adzhubei et al. 2010; Yue et al. 2006; Reva et al. 2011; Gonzalez-Perez et al. 2012; Gonzalez-Perez and Lopez-Bigas 2012) (i.e. MutationAssessor, SIFT, Polyphen2, TransFic, SNPs3D, Oncodrive-FM).
- Machine Learning-Based Approaches: methods that use machine-learning algorithms to model existing knowledge of drivers and passengers to classify driver mutations (Hanahan and Weinberg 2011; Adzhubei et al. 2010; Carter et al. 2009; Bromberg and Rost 2007; Douville et al. 2013) (i.e. CHASM, Polyphen2, SNAP, CRAVAT).
- Frequency-Based Approaches: methods that differentiate drivers and passengers by the number of mutations seen in the candidate driver gene in contrast to the expected number of mutations from functionally neutral passengers (Boca et al. 2010; Dees et al. 2012; Reimand and Bader 2013; Lawrence et al. 2013) (i.e. MutSig, ActiveDriver, MuSiC).
- Pathway-Based Approaches: methods that identify drivers based on the impact a mutated gene would have on gene interactions and biological pathways (Wendl et al. 2011; Ciriello et al. 2012; Vandin et al. 2012; Ng et al. 2012; Bashashati et al. 2012) (i.e. MEMo, Dendrix, DriverNet, PARADIGM-Shift).

The methods described above all excel in explaining some of the biological properties associated with driver mutations (Zhang et al. 2013). Unfortunately, no model exists that can identify all the driver mutations in any given cancer with great accuracy and precision, and many existing models tend to disagree with each other (Zhang et al. 2013). Because of this, there is no computational gold standard for driver mutations in cancer (Tran et al. 2012). In this chapter, we will discuss in detail the methods associated with each of the four broad categories. We will also introduce the strengths and potential limitations of each method.

3.1.2 Properties of Driver Mutations

As stated earlier, driver mutations differ from passenger mutations in that drivers will actively alter a cell's function to display tumorigenic properties, hence "driving" the cancer, whereas passenger mutations simply occur by happenstance. Not providing functions that "drive" the cancer, passenger mutations are simply along for the ride. Drivers can have a wide variety of functions and operate on a variety of mechanisms; however, all drivers provide selective advantage to a mutant cell, allowing it to thrive, grow, and most importantly, divide rapidly to out-compete the non-mutant cells (Bunz 2008). The selective advantage, illustrated in review by Hanahan and Weinberg fall under one of six functions, called "hallmarks" of cancer cells: (1) Sustaining Proliferative Signaling, (2) Evading Growth Suppressors, (3) Resisting Cell Death, (4) Enabling Replicative Immortality, (5) Inducing Angiogenesis, and (6) Invasion and Metastasis (Hanahan and Weinberg 2011).

3.1.3 Evolutionary Model of Cancer

The concept of driver mutations can be best explained by the clonal evolution model of cancer. The clonal evolution model of cancer, as first presented by Peter Nowell in 1976, states that cancer neoplasms originate from a single cell, or clone (Nowell 1976). Over time, the original clone accumulates somatic mutations (Nowell 1976). Although the vast majority of somatic mutations induced this way are functionally neutral or damaging to the clone, in rare instances, a mutation in a hallmark gene will be advantageous to a clone. For this reason, mutated genes with hallmark properties are considered cancer genes (Nowell 1976). The cancer gene, with a hallmark property, will provide the clone with a unique advantage and higher overall fitness that allow it to survive, prosper, and out-compete other cells. This results in an outgrowth of the clone with the new mutation called a neoplasia (Bunz 2010).

A single mutation in a cancer gene is often not enough to trigger cancer (Knudson 1971). The vast majority of neoplasia are not equipped to sustain its expansion and will fail to progress and eventually die, marking the end of the particular clone (Nowell 1976; Bunz 2010). This is due to selective pressures such as the body's immune system response, changes in the cellular microenvironment, or even self-induced pressures such as a shortage of oxygen as a result from its proliferative success (Kim et al. 2009). Just as most somatic mutations will not lead to cancer genes, most neoplasia will not lead to cancer. However, in rare cases, the clone will accumulate new mutations over time, some of which will lead to the formation of new cancer genes that will provide additional growth and fitness advantages for the clone, allowing for the clone to adapt and thrive in the microenvironment and even spread to others (Nowell 1976; Bunz 2010).

The clonal evolution model illustrates many concepts that are required to better understand cancer driver mutations. First, for a mutation to be considered a driver, it must have a significant functional impact on a hallmark gene and/or biological pathway (Hanahan and Weinberg 2011). Second, since a single cancer gene gone awry is not enough to trigger cancer, cancers generally have multiple drivers (Torkamani and Schork 2008). Third, although cancer is driven by multiple drivers with hallmark properties, there are many combinations of different drivers that may lead to the same end result of cancer (Leiserson et al. 2013). Therefore, the drivers within each individual tumor may vary, highlighting the concept of tumor heterogeneity.

3.1.4 Types of Cancer Genes

There are two main types of cancer genes: oncogenes and tumor suppressors. Oncogenes are genes in which a gain of function alteration contributes to the development of cancer (Bunz 2010; Croce 2008). Genes that can become oncogenes are considered proto-oncogenes. Mutations in oncogenes are considered activating mutations as the oncogenic version of these genes present increased activity, thereby being classified as Gain-of-Function mutations. Oncogenes are generally dominant and only one mutated allele of a proto-oncogene is required for the gene to show cancer-like properties. Examples of oncogene functions are involved in functions such as Growth Factors, Receptor and Cytoplasmic Tyrosine Kinases, Serine and Threonine kinases, Regulatory GTPases, and transcription factors. Examples of oncogenes include EGFR, RAS, WNT, MYC, ERK, and TRK (Bunz 2010; Croce 2008).

In contrast, a tumor suppressor is a gene that protects a cell from becoming cancerous. A loss of function of a tumor suppressor through genetic alteration contributes to the development of cancer (Bunz 2010; Sherr 2004). Mutations in tumor suppressors are considered inactivating mutations, resulting in Loss-of-Function mutations. Unlike oncogenes, tumor suppressors are generally recessive, and for that reason, both alleles of a tumor suppressor are required to be inactivated for a functional effect, i.e. the so called "two-hit" model (Knudson 1971). Examples of tumor-suppressor gene functions include repression of genes responsible to continue the cell cycle, triggering apoptosis, blocking contact-inhibition, and repairing DNA. Examples of tumor suppressors include TP53, RB1, PTEN, BRCA1, BRCA2, PIK3CA, AKT, and APC (Bunz 2010; Sherr 2004).

3.1.5 Types of Genetic Alterations in Cancer

There are many different ways a gene can be altered. The question of where and how a gene is altered is very crucial to assessing the impact of a particular mutation. Not all mutations and genetic alterations will have the same impact on the gene (Bunz 2010; Yokota 2000). For example, a mutation in a coding region is more likely to have an impact on a gene's activity than one in a non-coding region (Kryukov et al. 2005). Even though recent studies have shown that alterations in non-coding sequences can be impactful to cancer progression (Vinagre et al. 2013; Landa et al. 2013), most current methods in detecting drivers tend to narrow the scope in coding regions only (Bunz 2010). Nevertheless, even in exonic regions, some types of mutations tend to have more impact on the overall well-being of the cell than others.

The simplest and most intuitive type of genetic mutation is the point mutation. Single base-pair substitutions refer to the replacement of a single nucleotide with another and they can be divided into three groups: silent, missense, and nonsense mutations. Silent mutations occur in the third "wobble" position of a codon (Crick 1966). Due to the redundancy of amino acid codes, silent mutations are substitutions that do not occur in a change in a protein. Silent mutations generally have the least impact, as they do not alter the primary structure of the resulting protein, although they have been shown to have minor effects on the secondary and tertiary structure of the resulting protein. A missense mutation occurs when the single base-pair change results in a single amino acid change. A missense mutation can affect all structures of the resulting protein: primary, secondary, tertiary, and quaternary. The effects of a missense mutation depend both on the similarity of the replacement protein to the original and the position of the mutation. A nonsense mutation is a mutation in which the single base pair substitution transforms an amino acid codon to a stop codon. Nonsense mutations lead to premature truncation of the protein, rendering it non-functional.

In addition to point mutations, small insertions and deletions (indels) can cause frame-shift mutations, resulting in a completely new set of codons as an indel will shift the reading frame. Like nonsense mutations, these proteins are nonfunctional. These faulty proteins are usually degraded and are responsible for the formation of null alleles (Bunz 2010).

Point mutations and indels are not the only form of genetic alterations that can lead to cancer genes. An example of large-scale mutations is copy number variation (CNV). CNVs cause changes of the number of copies of a chromosomal region. CNVs may be either amplifications, presentation of multiple copies of a gene, or deletions, the loss of gene copies. Other examples of large-scale mutations include chromosomal translocations, the interchange of genetic parts from non-homologous chromosomes; chromosomal inversions, reversing sections of a chromosome; and loss of heterozygosity, the deletion of an allele (Bunz 2008). There are other forms of genetic alterations that are epigenetic in nature. Even though these alterations have no effect on the genomic sequence itself (mainly through DNA methylation and histone modification), they can sometimes have profound effects in tumor progression. For example, DNA methyltransferases target CpG islands in the promoter region leading to spontaneous deamination and lowering of gene expression by restricting transcription, effectively silencing the gene. Promoters often are unmethylated in normal cells but hypermethylated in cancer cells.

3.2 Overview of Computational Methods to Identify Driver Mutations

The initial type of methods that identified driver mutations in cancer relied on simple recurrence as a measurement. In simple recurrence, drivers and passengers were classified by the number of times they were observed in patient populations (Jones et al. 2008). Although this method was crucial in identifying *some* common drivers such as TP53 and EGFR (Jones et al. 2008), it soon became clear that based on the biological properties of driver mutations, several difficult challenges need to be overcome in order to determine all of the driver mutations in cancer.

3.2.1 Challenges for Driver Mutation Identification

Many difficulties in identifying driver mutations arise from the concept of tumor heterogeneity, the concept that no two cancer genomes will exhibit the same mutation profiles (Stratton 2013; Pe'er and Hacohen 2011). Therefore, two patients with the same cancer may have vastly different drivers. Additionally, drivers and passengers may switch roles such that a driver in one patient may be a passenger in another patient (Cooke et al. 2010). The advent of cancer subtypes has explained some of the heterogeneity; however, it is at best a compromise. Tumor heterogeneity contributes to the long-tail distribution of the frequency cancer mutations. The long-tail hypothesis states that cancer is driven not only by a few common genes that are mutated in many patients, but also many genes that are not mutated in many patients (i.e. less frequently mutated genes) (Ding et al. 2010). This implies that there will be many rare, yet undiscovered driver mutations that are obscured by tumor heterogeneity.

Another challenge in driver mutation identification is determining what constitutes a mutation. Not all mutations are created equal, some mutations display greater functional impact on a gene in terms of its protein structure and will be more damaging (Kumar et al. 2009). Even genes that have functionally damaging mutations across many patients are not necessarily drivers. Some genes have little functionality in cancer development and progression but are mutated frequently by chance. The most famous example of a highly recurrent passenger gene is the TTN gene. TTN is the largest gene in the human genome, and it functions as a molecular spring for the passive elasticity in muscle cells (Nair and Banerji 2013).

TTN does not have a large impact in many of the flagship cancer pathways (Lawrence et al. 2013). However, due to its large size, it is often mutated in cancer cells due to random chance alone, confounding the results of many methods.

A third challenge is to map the biological function of potential driver mutations. As demonstrated in the TTN example, some genes may present damaging mutations but due to the gene's function being unrelated to cancer pathways, they are most likely to be passengers. Individual driver genes do not operate by themselves; rather they interact with many other genes in complex biological networks (Bashashati et al. 2012). Therefore, driver mutations must be verified by their biological functions. A driver mutation is expected to interact with other genes in various cancer pathways to further promote different hallmarks of cancer (Hanahan and Weinberg 2011; Schwartzentruber et al. 2012).

3.2.2 Resources Available for Driver Mutation Identification

For researchers interested in identifying driver mutations, there exists a wealth of publicly-available data regarding molecular signature data, compendiums on driver mutations, pathway databases, and comparison tools that all can be utilized to achieve a greater understanding of driver mutations in cancer. Perhaps the most comprehensive of these resources is The Cancer Genome Atlas (TCGA), a resource of molecular alterations over large cohorts of patients representing a wide array of cancers (Cancer Genome Atlas Research Network 2008). With regards to curated catalogs of known somatic mutations in cancer, the Sanger Institute's COSMIC and the Cancer Gene Census, maintain a well-defined comprehensive list of common mutations already identified as drivers (Bamford et al. 2004; Futreal et al. 2004). Other tools such as Biocarta (Kim et al. 2012), NCI Pathway interaction Database (PID) (Schaefer et al. 2009), Reactome (Croft et al. 2011), or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) all provide valuable information on curated cancer pathways for evaluating potential driver genes.

3.2.3 Summary of Different Algorithms for Driver Mutation Identification

Name	Type	Website
SIFT	Sequence-based	http://sift.jcvi.org/
PolyPhen2	Sequence-based Machine learning-based;	http://genetics.bwh.harvard.edu/pph2/

(continued)

(continued)

Name	Type	Website
MutationAssessor	Sequence-based	http://www.bitnos.com/info/mutation-assessor
TransFic	Sequence-based; aggregate method	http://bg.upf.edu/transfic/help
Oncodrive-FM	Sequence-based; aggregate method	http://bg.upf.edu/group/projects/oncodrive-fm.php
SNPs3D	Sequence-based; aggregate method	http://www.snps3d.org/
CHASM	Machine learning-based	http://wiki.chasmsoftware.org/index.php/Main_Page
CRAVAT	Machine learning-based; aggregate method	http://www.cravat.us/
SNAP	Machine learning-based	https://rostlab.org/services/snap/
MutSig	Frequency-based	http://www.broadinstitute.org/cancer/cga/mutsig
MutSigCV	Frequency-based	http://www.broadinstitute.org/cancer/cga/mutsig
ActiveDriver	Frequency-based	http://www.baderlab.org/Software/ActiveDriver
MuSiC	Frequency-based; Pathway-based	http://gmt.genome.wustl.edu/genome-music/0.2/doc/
MEMo	Pathway-based	http://cbio.mskcc.org/tools/memo/
HotNet	Pathway-based	http://compbio.cs.brown.edu/projects/hotnet/
Dendrix	Pathway-based	http://compbio.cs.brown.edu/projects/dendrix/
DriverNet	Pathway-based	http://www.bioconductor.org/packages/2.12/bioc/html/DriverNet.html
Paradigm-Shift	Pathway-based	http://sysbio.soe.ucsc.edu/paradigm/tutorial/

3.3 Sequence-Based Approaches

The underlying belief in these approaches is that mutations that have functional impact on a gene are more likely to be driver mutations in cancer. These methods assess the functional impact of mutations by predicting the consequences, either through evolutionary impact on conserved regions or changes in the resulting amino acid and potential effects on the protein's secondary and tertiary structure. Examples of these approaches include Separating Tolerant from Intolerant (SIFT) which performs multiple sequence alignments (MSA) to determine the evolutionary impact of altered amino acids in protein homologs to predict functional impacts (Kumar et al. 2009); Polyphen2 combines a multiple sequence alignment to detect mutations with a Naïve Bayes Classifier (NBC) to train the potential functional impact (Adzhubei et al. 2010).

Results from many sequence-based approaches applied to cancer studies have shown that mutations in driver genes tend to have a much higher functional impact to the sequence and resulting protein structure than those of non-driver genes (Reva et al. 2007, 2011; Gonzalez-Perez et al. 2012; Gonzalez-Perez and Lopez-Bigas 2012). These methods also have the advantage of being able to evaluate individual patients' mutations to identify the drivers (Reva et al. 2007, 2011; Bashashati et al. 2012). However, these approaches also present several drawbacks as well. These methods are unable to separate mutations that provide a selective advantage to the overall cell fitness (Zhang et al. 2013). By definition, only mutations that provide a selective advantage to the tumor's growth and survival can be considered driver mutations (Hanahan and Weinberg 2000). Therefore, sequence-based approaches often struggle in separating driver mutations from passenger mutations. This drawback has prompted many groups to look into other methods to detect driver mutations, and for this reason, sequence-based approaches are not commonly used as the sole determinant of novel driver mutations (Zhang et al. 2013; Adzhubei et al. 2010; Yue et al. 2006). Nevertheless, these tools are widely applied as filters, comparison tools, and confirmation for more cancer-specific driver mutation methods.

3.3.1 MutationAssessor

The aforementioned sequence-based methods are generic methods to identify functionally relevant mutations and are not specific to cancer driver mutations. However, some methods have shown to perform well in detecting impactful mutations. One method is MutationAssessor, which predicts the consequence of a mutation using a Functional Impact Score (FIS). The FIS is a metric used to quantify a mutation's impact on a gene by observing the evolutionary conserved patterns from a MSA using combinatorial entropy formalism (Reva et al. 2011).

The FIS of any non-synonymous mutation can be calculated as the average of two conservation scores: the general conservation score S_i^C and the subtype conservation score S_i^S. A mutation in a conserved region is more likely to have a functional impact than a mutation in a non-conserved region (Henikoff and Henikoff 1992). MutationAssessor measures the impact of a mutation from the wild-type amino acid residue α to the mutant β using an entropy score. The general conservation score at position i with respect to the MSA to go from $S_i^C(\alpha \rightarrow \beta)$ therefore is:

$$S_i^C(\alpha \rightarrow \beta) = -\ln\left(\frac{n_i(\beta) + 1}{n_i(\alpha)}\right) \tag{3.1}$$

where $n_i(\alpha)$ is the number of sequences which display the residual α (the wild type) at position i and $n_i(\beta)$ is the number of sequences which display the residual β (the mutant) at position i. This change predicts the functional impact of a protein by determining if a change in the amino acid sequence is highly conserved or not. MutationAssessor takes one step further by assessing the entropy difference of the particular subfamily of the observed difference $S_i^S(\alpha \rightarrow \beta)$. The rationale of determining subfamily impact is to model different interaction partners or substrates on the background of similar, conserved biochemical or cellular function (Sarid et al. 1987). To determine subfamilies, a clustering algorithm is used to divide the MSA into subfamilies and the subfamily conservation score $S_i^S(\alpha \rightarrow \beta)$ is a measure of the entropy difference between the $\alpha \rightarrow \beta$ change with regards to the subfamily that the β residual belongs.

$$S_i^S(\alpha \rightarrow \beta) = -\ln\left(\frac{n_i^p(\beta) + 1}{n_i^p(\alpha)}\right) \tag{3.2}$$

where $n_i^p(\beta)$ and $n_i^p(\alpha)$ in equation are the residual counts of α and β with respect to a particular subfamily p. The FIS score for MutationAssessor is simply the average of the two aforementioned conservation scores.

MutationAssessor applied the FIS score for 10,000 mutations cataloged in COSMIC and it was shown that genes with a high FIS score were much more likely to become drivers (Reva et al. 2007).

3.3.2 TransFic

There have been methods that combine the predictive value of several methods to determine the impact of genes in cancer. One example is TransFic, a method that combines the scores from MutationAssessor, SIFT, and Polyphen2, and compares their scores to the distribution of scores of alterations observed in genes with similar functional annotations to select for drivers (Gonzalez-Perez et al. 2012). The use of functional annotations in TransFic was applied to obtain a better grasp on the function of a particular driver in question.

The process of selection is illustrated below:

1. Obtain the Functional Annotations of the gene of interest using four sources: Gene Ontology Biological Process (GOBP) and Molecular Function (GOMF) categories, canonical pathways (CP), and Pfam domain (Dom) (Henikoff and Henikoff 1992; Dejongh et al. 2004; Chagoyen and Pazos 2010; Yu et al. 2012; Punta et al. 2012).
2. Determine the alterations associated with all genes related to the most specific functional term of the original gene of interest. This allows TransFic to not only calculate the impact of an altered gene, but also predict its biological function.

3. If less than 20 alterations are found, the user may choose to add other alterations in genes that have similar functions as the original gene of interest. This allows for an accurate reading of the functional impact score even with less available input.
4. Calculate and normalize the SIFT, Polyphen2 and MutationAssessor scores. The SIFT and Polyphen2 scores first undergo a logit transformation.
5. Calculate the mean, standard deviation and other summary statistics to determine the aggregate FIS score of both the gene and the potential function.

The authors compared their method to each of the individual methods that they aggregated and found that the aggregated results that were more concordant with COSMIC's category of driver mutations. They tested their score with the breast cancer driver PIKC3A and found that the impact of the mutation was more mild than previously thought. Another software developed by the same lab was Oncodrive-FM (Gonzalez-Perez and Lopez-Bigas 2012). Oncodrive-FM uses SIFT and Polyphen2, along with other driver mutation software such as MutSig in order to determine to select driver genes that present accumulated functional impact mutations across a gene (Gonzalez-Perez and Lopez-Bigas 2012).

3.3.3 SNPs3D

SNPs3D is another sequence-based approach that attempts to combine information from many different sources to draw conclusions (Yue et al. 2006). SNPs3D is made up of three gene modules: one concerning the impact a non-synonymous SNP (in our case, a point mutation in a tumor) has on the network, one that connects genes to other related genes based on a PubMed literature search, and a third which provides users with a literature score to measure how likely a gene is related to certain diseases. SNPs3D is unique in that it associates literature scores as a direct measurement to disease association (Yue et al. 2006).

SNPs3D covers the sequence-based data of a driver mutation using two methods: the first determining the amino acid substitution's stability on a proteins folded state (Yue et al. 2005) and the second being a conservation score similar to the one presented in MutationAssessor (Yue and Moult 2006). SNPs3D also links genes together to form gene to gene interactions based on the number of PubMed search results returning the pair of genes. It also counts abstracts from PubMed to link a mutated gene with a disease (Stapley and Benoit 2000). Using this integrated approach, SNPs 3D discovered candidate genes for a long list of diseases, including around 200 potential candidates for Lung Cancer (Yue et al. 2006).

3.4 Machine Learning-Based Methods

Machine-learning approaches operate by training a classifier on a gold standard of driver and passenger mutations to develop a model, which is utilized to determine the drivers and passengers of a new dataset. Generally, these methods train their data from a catalog of missense mutations, and the classifiers themselves range from Naïve Bayes Classifiers to Random Forests to Neural Networks. Machine-learning based approaches have better ability to distinguish drivers from passengers than methods that only consider mutation's functional impact. Once a model is classified, the model can be fitted to any number of patients or groups. However, the machine-learning approaches heavily rely on a gold standard of driver and passenger mutations as a training set, which could be problematic as there currently is no established computational gold standard. Even though COSMIC and the CGC have good compendium for common drivers, they do not take into account rare drivers (Futreal et al. 2004).

3.4.1 CHASM

One example of a machine learning-based method is the Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM). CHASM seeks to identify and rank missense mutations most likely to augment tumor cell proliferation (Carter et al. 2010). CHASM applies a Random Forest Classifier on 49 predictive features including amino acid substitution properties, alignment-based estimates of evolutionary conservation at the mutated position, predicted structural changes at the mutated position and annotations from the UniProtKB feature table. The Random Forest Algorithm is a decision tree classifier that uses a set of random classification trees to vote on a classification of a particular mutation as "driver" and "passenger". Each tree then "votes" for the eventual classification of the alteration (Carter et al. 2009, 2010; Gnad et al. 2013).

The authors selected 2,488 missense mutations breast, colorectal, and pancreatic cancers. The driver mutations selected were from COSMIC and various biological studies in which specific genes were demonstrated to have proliferative roles, and the passenger mutations were computer generated via simulation with an algorithm that recapitulates base substitutions found in brain tumors (Carter et al. 2010). The authors reported higher sensitivities and specificities than traditional sequence-based methods such as SIFT and Polyphen2. Additionally, when training the classifier, the authors reported that many of the variables by themselves only explained a small percentage of the model, which the authors used to justify their rationale behind Random Forests. Random Forests work with each variable jointly rather than as individuals. When applied to a GBM dataset, the authors predicted that 49 of the 607 missense mutations in the GBM dataset, or 8 %, were drivers (Carter et al. 2009).

3.4.2 CRAVAT

A recent machine learning-based method, Cancer-Related Analysis of Variants Toolkit (CRAVAT), seeks to provide predictive scores on the importance of somatic alterations of in cancer genes using a variety of classifier tools (Douville et al. 2013). CRAVAT is unique because it (1) combines the results of multiple classifiers to hone in on both the impact of the driver and the biological function of a somatic alteration; (2) provides a user-friendly workflow where users can submit their jobs to the server and receive both the gene's importance rating and a variety of PubMed literature sources that relate to the important drivers that CRAVAT predicted; and (3) is not limited by the size of the dataset (Douville et al. 2013).

CRAVAT uses three machine learning tools for its workflow: SnvGet, CHASM, and VEST (Carter et al. 2009, 2013; Wong et al. 2011). CRAVAT uses SnvGet to get classifier information for the subsequent CHASM and VEST runs. SnvGet returns 86 pre-computed features for each alteration such as physiochemical properties of amino acid residues; scores derived from multiple sequence alignments of protein or DNA; region-based amino acid sequence composition; predicted properties of local protein structure; and annotations from the UniProtKB feature tables (Wong et al. 2011). The features are then used by CHASM to predict whether or not the alteration in question is a driver, and then VEST (designed by the same authors as CHASM), which also utilizes a Random Forest classifier to determine the function impact of the predicted protein. The *p*-values from both tests are aggregated to return a list of functional driver genes for the user (Douville et al. 2013).

3.4.3 Polyphen2 and SNAP

In addition to CHASM, several other machine-learning approaches have been used to identify driver mutations. Polyphen2, as mentioned earlier as a sequence based method, uses the Naïve Bayes Classifier (NBC) to predict functional impact, improving on the traditional multiple sequence-based approach with knowledge from machine learning (Adzhubei et al. 2010). The alignment output from Polyphen2 is used to select the features for the Naïve Bayes Classifier, which is then used to classify them on function. The NBC works by solving the probability of a sampling belonging to a group c from all groups C using Baye's rule with respect to features $F_1, F_2 \ldots F_n$. The group with the highest probability that a sample could belong is the predicted classifier.

Another method, SNAP, utilizes a neural network to predict the functional effects of non-synonymous SNP, which can be applied to missense mutations to predict drivers (Bromberg and Rost 2007). Both Polyphen2 and SNAP are general functional impact algorithms that can be applied to cancer but are not necessarily created to specifically model cancer mutations.

3.5 Frequency-Based Methods

The third class of driver mutation identification software is methods based on mutation frequency. In the early days of driver mutation identification, simple recurrence was the first method to determine driver mutations. Drivers were defined by the number of times a gene was mutated (Schwartzentruber et al. 2012). Although many common driver mutations were detected using this method, simple recurrence has since fallen out of favor as it does not account for (1) rare mutations in the long tail of driver gene distribution and (2) propensity to select genes that have a high probability due to chance to be mutated by being large or having a high background mutation rate.

Frequency-based methods are among the most powerful methods in classifying common driver genes and passenger genes, and these methods have been some of the most widely-adopted and widely-utilized methods in driver mutation detection (D'Antonio and Ciccarelli 2013). However, one drawback of frequency-based methods is that these methods, like machine-learning based methods, require a large amount of input data from many patients to operate.

3.5.1 *MutSig*

One of the most-utilized frequency-based methods is MutSig (Banerji et al. 2012). The original MutSig assumes a single average background mutation rate, μ, which can be tailored to be category-specific: μ_c.

Examples of category specific criteria taken from a Lung Carcinoma study were (1) transitions in C's or G's in CpG dinucleotides; (2) transversions in C's or G's in CpG dinucleotides; (3) transitions in other C's or G's; (4) transversions in other C's or G's; (5) transitions at A's or T's; (6) transversions in A's or T's; and (7) small insertions/deletions, nonsense and splice site mutations (Lawrence et al. 2013). Then to calculate a p-value for each gene based on category-specific background rates, a score s is calculated for each gene. The score of each gene's mutation significance s_g is based on the binomial probability distribution given the parameters of the number of mutations in the category n_c, the number of bases covered by those mutations N_c, and that category's background mutation rate: μ_c.

$$s_g = \sum_c -10 \times \text{binomial}(n_c, N_c, \mu_c) \tag{3.3}$$

After calculating the score, the background distributions of all the mutation rates are convoluted and a p-value is calculated by calculating the probability that the convoluted mutation rates can exceed the score s_g. A Benjamin-Hochberg correction is used to correct for multiple testing (Lawrence et al. 2013). The authors of the original MutSig applied the data to a Lung cancer dataset and found 450 candidate drivers that were mutated at a frequency much higher than the expected frequency as assumed from the background mutation rate (Greulich et al. 2012).

3.5.2 *MutSigCV*

Recently, a newly-published version of MutSig, MutSigCV (Lawrence et al. 2013), has been released. MutSigCV offers additional features to the original MutSig. MutSig corrects for the extensive false positive findings of previous driver mutation identification software by correcting for the heterogeneity of the mutation rates among genes, the mutations rates among patients, and among the mutation types themselves by allowing separate models for multiple types of heterogeneity. MutSigCV also incorporates molecular properties of the gene that may co-vary with the mutation rate of the gene into their model. Examples include gene expression, DNA replication time, open versus closed chromatin status, local GC content, and local gene density (Lawrence et al. 2013).

In MutSigCV, each gene is placed in a high-dimensional covariate space and the gene's nearest neighbors are identified to supplement information to the background mutation rate of the gene in question. The information from the nearest neighbors of the gene, dubbed "Bagel", is combined with the gene's own mutation rates to estimate the background mutation rate. This process, combined with category and patient-specific background mutation rates (calculated via the original MutSig model) provide the mutation rates used to calculate the significance of each gene.

The authors of MutSigCV analyzed 3,083 tumor normal pairs to both look for sources of heterogeneity and for novel driver mutations. The authors found that tissue type mutation rate are highly variable and that lung and skin cancers tend to have high mutation rates although much of the variation can also be attributed to the patients themselves (Lawrence et al. 2013). The authors also studied the type of mutation present for tissue types, and found that lung cancer tended to have more C→T mutations while melanoma patients tended to have more C→A mutations. The regional heterogeneity was one of the most variable, meaning that certain genes are much more likely to mutate by chance than others, and that that mutation rates tended to coincide with gene expression and the time of DNA replication. Taking into account this heterogeneity, the method assigned each gene and tumor type a score, which was used to correct the background rate of mutations in specific genes for specific tumors, and patients. This approach was used to confirm common drivers, eliminate false positive drivers, and suggest possible new drivers (Lawrence et al. 2013).

3.5.3 *ActiveDriver and MuSiC*

Other recent methods include ActiveDriver and MuSiC (Dees et al. 2012; Reimand and Bader 2013). ActiveDriver is a method developed to discover driver genes in among genes with phosphorylation single nucleotide variants (pSNV). ActiveDriver performs a hypothesis test to determine whether or not the phosphosite-specific mutation rate is the same as the gene-wide mutation rate for particular genes using generalized linear regression tests. The authors of ActiveDriver found that their approach identified many common phospho-specific drivers such as TP53 and EGFR as well as new candidate genes in FLNB, GRM1, POU2F1 (Reimand and Bader 2013).

MuSiC also employs the concept of selecting for genes that tend to be mutated more than a background mutation rate in their novel test, Significantly Mutated Gene (SMG) test. The background rate was a combination of mutated genes in the entire sample set of all patients, mutated genes in the patient, and mutated genes within the subgroup of the gene in question (Dees et al. 2012). MuSiC also supplement their results using pathway analysis through the PathScan algorithm (Wendl et al. 2011), which combines individual selection of driver genes to a multiple-sample value using the Fisher-Lancaster approach (Wendl et al. 2011) to determine the mutated pathway of the driver genes in their analysis.

3.6 Pathway-Based Methods

The most recent type of model to determine driver mutation relies on biological pathways. Pathway-based models have been shown to be effective not only in reliably determining common driver mutations, but also have been able to pinpoint the biological pathways that could be the source of the cancer (Ciriello et al. 2013). As a result, pathway-based methods have a unique advantage over other types of methods in that they take into account gene interactions and potential biological effects rather than simply viewing driver genes individually (Wu et al. 2010). For example, a particular candidate driver gene that shows significantly more mutations in cancer than in normal cells may still not be a true driver gene (Michor and Polyak 2010). If the candidate gene does not affect a cancer pathway or does not interact with many genes that are crucial in cancer-pathways, the candidate gene may have no true biological connection to cancer. Pathway-based approaches allow us to verify functional impactful candidate drivers. These methods are sometimes used to supplement other methods as was demonstrated in the case of ActiveDriver and MuSiC, as measure of the biological significance of their methods (Dees et al. 2012; Reimand and Bader 2013; Wendl et al. 2011).

3.6.1 MEMo

Some pathway-based approaches are not built with specific cancer genes in mind, but rather, these approaches are aimed at discovering driver pathways, groups of genes that may interact together to promote tumorigenesis. Mutual Exclusivity Modules in cancer (MEMo) serves to determine groups of genes that contribute to tumorigenesis (Ciriello et al. 2012). These gene groups, or modules, together are highly recurrent, have similar pathway impact in terms of biological processes, and also are mutually exclusive meaning that only one gene in each gene group is mutated at a time in any given patient. This idea follows the mutual exclusivity rule in cancer pathways, i.e., generally one mutated gene in a pathway is enough to alter the pathway's function. The algorithm for MEMo is described below:

1. Build binary event matrix of significantly altered genes. The binary event matrix (B) is an $n \times m$ matrix where n is the number of genes in the dataset and m is the number of samples (patients) being observed. As a binary event matrix, a cell in the matrix $B_{i,j}$ will be 0 if a gene i is altered in the sample j.
2. Build a gene network to identify gene pair interactions. This step involves the building of a gene network that will gauge the interactions and pathways present in cancer genes. The authors at MEMo built two gene networks: the first being a combination Human Interaction Network based on both curated and non-curated networks, and the second one simply based on manual curation.
3. Extract Cliques: MEMo then finds all cliques in the network. A clique is a fully connected subgraphs such that each subgraph cannot be contained by another fully connected subgraph.
4. Assess each clique for mutual exclusivity. The idea of this step is to determine whether or not the clique has both highly recurrent gene alterations, and also whether or not only one gene in the subgraph is mutated at once. MEMo tests on whether the set of genetic alterations occurs by chance. MEMo builds a null model by randomly permuting the event matrix, and then applies a Markov Chain Monte Carlo method called "permutation switching" to randomly generate networks to find simulated cliques. The cliques are tested for mutual exclusivity under the null model, thus allowing MEMo to determine an empirically derived p-value to gauge the mutual exclusivity of the cliques.

The authors of MEMo discovered several mutually-exclusive modules in GBM such as EGFR, PDGFRA, and NF1 and TP53, CDKN2A, and GLI1. One of the genes in these modules is likely to be altered in any given patient. MEMo is a unique approach at observing cancer as it acknowledges that although patients may have different mutations to drive the cancer, many of those mutations have similar biological effects eventually (Ciriello et al. 2012).

3.6.2 HotNet and Dendrix

In the spirit of finding subnetworks in cancer, Vandin et al. developed two algorithms to determine the impact of mutated genes have on biological pathways: HotNet and Dendrix (Vandin et al. 2011, 2012). HotNet algorithm combines mutation data and protein–protein interaction network information to find subnetworks of genes that are mutated in a significant number of cancer patient (Vandin et al. 2011). Using mutation and gene interaction data on an undirected graph, HotNet uses a heat diffusion algorithm where a mutated sends a "heat" signature based on the number of mutations present in that gene evenly to its neighbors such that genes with lower degrees of connectivity receive a larger proportion of "heat" than those with high connectivity. The idea behind HotNet is that genes with lower connectivity will define the boundaries of the neighborhood (the subnetwork) as they will retain heat better, allowing HotNet to pinpoint subnetworks.

Dendrix, on the other hand, determines driver pathways using two concepts: Mutually Exclusivity (as demonstrated in MEMo) and coverage (recurrence). Modeling a gene interaction network as an adjacency matrix, Dendrix finds the submatrix within the matrix that will maximum coverage, that is, cover the most patients while being mutually exclusive, that is not having any two genes in the submatrix mutated simultaneously within a patient (Vandin et al. 2012). Dendrix uses a greedy MCMC method to do so. After selecting a starter gene, Dendrix selects the neighbor that has the most mutations without any of those mutations being in a patient that already had a mutation in a previously selected gene. One frequently sampled gene set from Dendrix's application to GBM was CDKN2B, RB1, CYP27B1 (Vandin et al. 2012).

3.6.3 DriverNet

One of the most recent pathway-based methods is DriverNet (Bashashati et al. 2012). DriverNet models both gene mutation events and differential expression events of a group of patients into a bipartite graph. The algorithm then applies pathway information to select for mutated genes that are the most well-connected to genes that are differentially expressed. The DriverNet algorithm is a greedy optimization algorithm aimed at determining driver genes as genes with the most pathway impact, which they measure as genes that create the most outlying differentially expressed genes. The greedy optimization algorithm is described below:

1. Create a bipartite graph $B(V^m, V^0, E)$, a graph whose vertices can be divided into two disjoint sets V^m and V^0 such that every edge connects a vertex in V^m to one in V^0. In DriverNet's case, V^m is a mutation matrix built in a similar fashion as MEMo's binary event matrix. V^0 is a binary $n \times m$ matrix where n is the number of genes in the dataset and m is the number of samples (patients) being observed. V^0 is equal to 1 for gene i with respect to patient j if the normalized difference between the tumor and normal expression exceeds a certain threshold. E is an adjacency matrix representing the gene network that connects V^m and V^0 in the bipartite graph. E can be built by similar procedures as MEMo's adjacency matrix.
2. Let Z be the set of all connected outlying events, and z be the set of covered outlying events (initially a null set).

3. Choose the mutated gene that contains the largest number of uncovered outlying expression events. Add that to the driver mutation list. Add the outlying events to z.
4. Remove the mutated gene and its connecting edges from the bipartite graph B.
5. Stop when all connected outlying events all covered (when $Z = z$).

DriverNet combines gene expression, mutation information among groups of patients, and biological pathways (Bashashati et al. 2012). The authors of DriverNet tested their results in Breast Cancer and Glioblastoma datasets and found an abundance of infrequently mutated genes: 22 in the Breast Cancer dataset and 13 in Glioblastoma. The advantage of DriverNet is that it is less dependent on recurrence and therefore can detect rare mutation.

3.6.4 PARADIGM-Shift

PARADIGM-Shift predicts functions of driver genes as gain-of-function or loss-of-function genes in specific cancer pathways (Ng et al. 2012). PARADIGM-Shift has the ability to determine not only if a candidate driver is functionally impactful, but also the type of impact that the driver gene may show. The authors utilized PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) (Vaske et al. 2010), using gene expression and cy number change signals as inputs to determine the impact of upstream and downstream genes of a candidate driver. The difference activity in upstream and downstream genes of the driver determines a gain-of-function (high downstream, low upstream activity) or loss-of-function (high upstream, low downstream activity).

The activity score was determined by PARADIGM, which uses belief-propagation on a factor graph to compute the log-posterior odds score called inferred pathway levels (IPLs) for each gene, complex, protein family and cellular process using gene expression, copy number and/or genetic interaction. Genes that are more active in a tumor with more activity have positive IPL scores while genes with less activity in the tumor than normal cells have negative IPL scores (Vaske et al. 2009). PARADIGM-Shift runs two iterations of PARADIGM, one with the gene of interest and its upstream genes in the pathways to measure the loss of function score, and one with only the gene of interest and its downstream genes to measure the gain of function score. The PARADIGM-Shift score is the difference of the two paradigm runs. The authors of PARADIGM-Shift applied their approach to both common, TP53, and uncommon, NFE2L2, genes to analyze the impact (Ng et al. 2012).

3.7 Discussion

Each of the four approaches, and the various methodologies associated with each of the approaches, has different advantages and addresses many of the challenges associated with driver mutations. Unfortunately, no method can solve all the challenges, and no perfect model exists that can fully reverse engineer the clonal evolution model of cancer and select only drivers that serve a function relating to the hallmarks of cancer. The task of accounting for tumor heterogeneity, genetic function, and mutation severity is indeed daunting. Many researchers, therefore, have applied multiple methods to determine driver mutations (Adzhubei et al. 2010; Bashashati et al. 2012). The multi-step approach allows for researchers to address multiple challenges in driver mutation identification at the same time.

In addition to the current challenges involved in driver mutation identification, there are also many future avenues of studying driver mutations that have yet to be identified and modeled. Some examples include analyzing the cumulative effects of passenger mutations, accounting for intra-tumor heterogeneity, and predicting the effects of mutations in non-coding regions.

A study from McFarland et al. found that even though a single passenger mutation has a negligible impact on tumorigenesis, the cumulative effect of all passengers may affect a cell's tumor progression model in ways not explainable by widely accepted driver mutation models (McFarland et al. 2013). Much intra-tumor heterogeneity is also ignored by driver mutation methods as most cancer genome sequencing project sequences a bulk tumor tissue from a population of cancer cells. In other words, the sequencing is a simple average of the cells, and no model exists to explain intra-tumor heterogeneity (Michor and Polyak 2010).

The methods described in this chapter are mostly only applicable to point mutations in coding regions of the genome. As described earlier, only a small subset of cancer mutations are point mutations. Detailed impacts of larger scale mutations and structural rearrangements have yet to be described. Additionally, only 2 % of the genome codes for proteins, leaving 98 % of the genome in non-coding regions unexplained. Mutations in non-coding regions can have profound impact on gene regulation related to cancer development and progression. Currently, no driver mutation software can systematically predict the effects of alterations in non-coding sequences. All these challenges need to be addressed by future computational methods.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91:355–8.

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhim R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486:405–9.

Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13:R124.

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. Genome Biol. 2010;11:R112.

Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35:3823–35.

Bunz F. Principles of cancer genetics. Dordrecht: Springer, 2008. p. 325.

Bunz F. Principles of Cancer Genomics. Berlin: Springer; 2010.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. 2008;455:1061–8.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69:6660–7.

Carter H, Samayoa J, Hruban RH, Karchin R. Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). Cancer Biol Ther. 2010;10:582–7.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3.

Chagoyen M, Pazos F. Quantifying the biological significance of gene ontology biological processes—implications for the analysis of systems-wide data. Bioinformatics. 2010;26:378–84.

Ciriello G, Cerami E, Aksoy BA, Sander C, Schultz N. Using MEMo to discover mutual exclusivity modules in cancer. *Curr Protoc Bioinformatics*, Chap. 8: Unit 8 17 (2013).

Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22:398–406.

Cooke SL, Ng CK, Melnyk N, Garcia MJ, Hardcastle T, Temple J, Langdon S, Huntsman D, Brenton JD. Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. Oncogene. 2010;29:4905–13.

Crick FH. Codon–anticodon pairing: the wobble hypothesis. J Mol Biol. 1966;19:548–55.

Croce CM. Oncogenes and cancer. N Engl J Med. 2008;358:502–11.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691–7.

D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14:R52.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.

Dejongh M, Van Dort P, Ramsay B. Linking molecular function and biological process terms in the ontology for gene expression data analysis. Conf Proc IEEE Eng Med Biol Soc. 2004;4:2984–6.

Ding L, Wendl MC, Koboldt DC, Mardis ER. Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet. 2010;19:R188–96.

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29:647–8.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.

Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14(Suppl 3):S7.

Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40:e169.

Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012;4:89.

Green ED, Guyer MS, National Human Genome Research I. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470:204–13.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153–8.

Greulich H, Kaplan B, Mertins P, Chen TH, Tanaka KE, Yun CH, Zhang X, Lee SH, Cho J, Ambrogio L, Liao R, Imielinski M, Banerji S, Berger AH, Lawrence MS, Zhang J, Pho NH, Walker SR, Winckler W, Getz G, Frank D, Hahn WC, Eck MJ, Mani DR, Jaffe JD, Carr SA, Wong KK, Meyerson M. Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of ERBB2. Proc Natl Acad Sci USA. 2012;109:14476–81.

Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89:10915–9.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321:1801–6.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

Kim Y, Lin Q, Glazer PM, Yun Z. Hypoxic tumor microenvironment and cancer cell differentiation. Curr Mol Med. 2009;9:425–34.

Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. Biol Direct. 2012;7:21.

Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA. 1971;68:820–3.

Kryukov GV, Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet. 2005;14:2221–9.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.

Landa I, Ganly I, Chan TA, Mitsutake N, Matsuse M, Ibrahimpasic T, Ghossein RA, Fagin JA. Frequent Somatic TERT promoter mutations in thyroid cancer: higher prevalence in advanced forms of the disease. J Clin Endocrinol Metab. 2013;98:E1562–6.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CW, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.

Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol. 2013;9:e1003054.

McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. Proc Natl Acad Sci USA. 2013;110:2910–5.

Michor F, Polyak K. The origins and implications of intratumor heterogeneity. Cancer Prev Res (Phila). 2010;3:1361–4.

Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 2012;28:i640–6.

Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194:23–8.

Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell. 2011;144:864–73.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290–301.

Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.

Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8:R232.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.

Sarid J, Halazonetis TD, Murphy W, Leder P. Evolutionarily conserved regions of the human c-myc protein can be uncoupled from transforming activity. Proc Natl Acad Sci U S A. 1987;84:170–3.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674–9.

Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, Sturm D, Fontebasso AM, Quang DA, Tonjes M, Hovestadt V, Albrecht S, Kool M, Nantel A, Konermann C, Lindroth A, Jager N, Rausch T, Ryzhova M, Korbel JO, Hielscher T, Hauser P, Garami M, Klekner A, Bognar L, Ebinger M, Schuhmann MU, Scheurlen W, Pekrun A, Fruhwald MC, Roggendorf W, Kramm C, Durken M, Atkinson J, Lepage P, Montpetit A, Zakrzewska M, Zakrzewski K, Liberski PP, Dong Z, Siegel P, Kulozik AE, Zapatka M, Guha A, Malkin D, Felsberg J, Reifenberger G, von Deimling A, Ichimura K, Collins VP, Witt H, Milde T, Witt O, Zhang C, Castelo-Branco P, Lichter P, Faury D, Tabori U, Plass C, Majewski J, Pfister SM, Jabado N. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature 2012;482:226–31.

Sherr CJ. Principles of tumor suppression. Cell 2004;116:235–46.

Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput. 2000:529–540.

Stratton MR. Journeys into the genome of cancer cells. EMBO Mol Med. 2013;5:169–72.

Thasni KAT, Ratheeshkumar T, Rojini G, Sivakumar KC, Nair RS, Srinivas G, Banerji A, Somasundaram V, Srinivas P. Structure activity relationship of plumbagin in BRCA1 related cancer cells. Mol Carcinog. 2013;52:392–403.

Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. Cancer Res. 2008;68:1675–82.

Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AM, Zhang T, Shaw P, Onetto N, Stein L, Hudson TJ, Neel BG, Siu LL. Cancer genomics: technology, discovery, and translation. J Clin Oncol. 2012;30:647–60.

Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18:507–22.

Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22:375–85.

Vaske CJ, House C, Luu T, Frank B, Yeang CH, Lee NH, Stuart JM. A factor graph nested effects model to identify networks from genetic perturbations. PLoS Comput Biol. 2009;5:e1000274.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, Melo M, Rocha AG, Preto A, Castro P, Castro L, Pardal F, Lopes JM, Santos LL, Reis RM, Cameselle-Teijeiro J, Sobrinho-Simoes M, Lima J, Maximo V, Soares P. Frequency of TERT promoter mutations in human cancers. Nat Commun. 2013;4:2185.

Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics. 2011;27:1595–602.

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27:2147–8.

Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11:R53.

Yokota J. Tumor progression and metastasis. Carcinogenesis. 2000;21:497–503.

Yu N, Seo J, Rho K, Jang Y, Park J, Kim WK, Lee S. hiPathDB: a human-integrated pathway database with facile visualization. Nucleic Acids Res. 2012;40:D797–802.

Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006;356:1263–74.

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005;353:459–73.

Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform. 2006;7:166.

Zhang J, Liu J, Sun J, Chen C, Foltz G, Lin B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. Brief Bioinform. 2013.