COMPUTATIONAL PREDICTION OF GENETIC DRIVERS IN CANCER

Alice B. Djotsa Nono[1], Ken Chen[2,3] and Xiaoming Liu[1]

[1]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, [2]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, [3]Corresponding Author: Ken Chen (kchen3@mdanderson.org)

Abstract

Cancer is a complex genetic disease driven by somatic mutations in the genomes of cancer cells. Distinguishing pathogenic "driver" mutations from non-pathogenic "passenger" mutations is a central task for functionalizing cancer genomics in patient care. With the outpouring of genomic information from next generation sequencing, predictive algorithms become relevant to filter the outnumbered pathogenic driver mutations from non-pathogenic passenger mutations. In this paper, we review recent published computational approaches for predicting cancer drivers at mutation, gene and pathway levels. We highlight the statistical approaches used for these methods and discuss their advantages, drawbacks including recent improvements. The current trend is to use multiple and complementary methods for a more accurate prioritization of cancer driver candidates available for targeted therapy at the clinical level.

Key words:

Cancer, driver, genomics, function, predictor, computational, mutation, tools, bioinformatics

Key Concepts:

- Cancer is a disease driven by mutations in the genome
- Only a small fraction of mutations are drivers that are responsible for cancer initiation and progression
- Distinguishing drivers from passengers is essential for genomic medicine
- Computational prediction of drivers is challenging due to complexity of biology and genomics
- Statistical and machine-learning approaches have been applied to discover the signature of drivers
- A mutation can affect the function of multiple genes and pathways
- The function of a mutation is context dependent and can vary in different diseases

Introduction

Cancer is a complex genetic disease (Stratton et *al.*, 2009; Vogelstein *al.*, 2013). Its initiation and progression are driven by somatic mutations in the genomes of cancer cells. Most of the somatic mutations are "passengers" that occur stochastically as a result of mutagenesis. These passenger mutations have no functional impact and do not contribute to tumorigenesis. Only a small fraction of somatic mutations are genetic "drivers" that lead to dysfunction of genes and pathways and provide growth advantage to cancer cells.

The identification of driver mutations from passenger mutations and germline polymorphisms usually starts with sequencing of matched tumor and normal DNA samples from cohorts of cancer patients. Identified after comparing sequencing reads against the human reference genome are somatic mutations, those present in only the tumor samples, and germline mutations, those present in both the tumor and the matched normal samples. A crucial next step is to prioritize the list of somatic mutations and identify driver mutations that are truly responsible for cancer initiation and progression. Over the last three decades, many analytical tools have been developed to help predicting the relationships between somatic mutations and cancer phenotypes. While some of these strategies focus on individual altered loci, other techniques assess mutated genes in a cohort of samples or evaluate group of genes involved in metabolic pathways. Patterns of recurrence or abundance, pattern of loss/gain function (impaired functional impact), signs of clustering in functional regions, genes or pathways, are the signatures that have fueled the development of predictive methods aiming at discriminating cancer drivers from passengers. In this article, we will describe computational methods for predicting somatic driver mutations at nucleotide, gene, or pathway levels. These three broader categories are shown in Figure 1.

I. Types of genomic alterations

The biology of cancer is driven by different types of somatic mutations, which include single nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations (CNAs), fusion genes, chromosomal/structural rearrangements, and epigenetic reprogramming. SNVs are sequence alterations that involve a single nucleotide and they are the most abundant variants observed in sequencing data. Synonymous SNVs (sSNVs) do not change protein sequences, whereas non-synonymous single nucleotide variants (nsSNVs) change protein sequences. Indels usually refer to insertions or deletions of short (1bp to 50 bp) nucleotide sequences in a genome. CNAs are gains or

losses of DNA segments, which can result in non-diploid copies of DNA segments in a genome. There are many ways to classify copy number alterations, depending on their sizes and types of alterations. For instance, aneuploidies usually refer to losses or gains affecting whole chromosomes; whereas, CNAs usually refer to alteration of segments between 1kbp and 1Mbp in length. Chromosomal rearrangements refer to gross changes in the structure of a chromosome due to duplication (increase of the number of copies of a chromosomal region), inversion (partial rotation of a chromosomal segment), deletion, translocation (one part of a chromosome attaches to another chromosome) and transpositions (short DNA segments moves from one position to the next position). A fusion gene can be introduced either by inter-chromosomal rearrangements, which combine two or multiple genes from different chromosomes into one fused gene, or by intra-chromosomal rearrangements such as deletion, inversion or duplication of large DNA segments on the same chromosomes. To date, many algorithms have been developed to enable the discovery of not only nsSNVs, but also impaired genes and pathways associated with cancer initiation, progression and development. **See also**: DOI: 10.1002/9780470015902.a0023379.

II. Algorithms for identifying driver mutations from passenger mutations

Mutations in protein coding regions can be classified into different categories based on their functional impacts on proteins. Four categories of mutational consequences are causatively involved in cellular outcome. While gain of function mutations activate new and abnormal function in oncogenes, loss of function mutations incapacitate tumor suppressors, drug resistance mutations impede the effect of a drug on the targeted protein, and finally switch of function mutations are intermediate between gain and loss of function and can switch from one kind to another (Reva *et al.*, 2011).

sSNVs are assumed to have no or low functional impact, whereas nsSNVs can have medium or high functional impact, often resulting in gain or switch of functions, and finally stop-gain and frameshift indels are assumed to have high functional impact, often resulting in loss of functions. Most algorithms predicting the functional impact (deleteriousness) of nsSNVs exploit four main types of features including evolutionary conservation of the site, physicochemical properties of the protein, protein domains and sequence context (Mao *et al.*, 2013, Dong *et al.*, 2015a). In this section, we discuss some algorithms screening for individual variants, especially nsSNVs, which are deleterious to proteins.

Because some cancer driver mutations exhibit the same features found in other disease causing mutations (Shihab *et al.*, 2013); many pathogenicity prediction programs originally developed for

separating disease causing mutations from neutral polymorphism have been used successfully to prioritize cancer driver mutations. These include but not limited to Sorted Intolerant from Tolerant (SIFT, Kumar *et al.,* 2009), PolyPhen-2 (Adzhubei *et al.,* 2010), LRT (Chun and Fay, 2009), MutationTaster2 (Schwarz *et al.,* 2014), CONsensus DELeteriousness score of missense mutations (Condel, Gonzalez Perez and Lopez-Bigas 2011), Protein Variation Effect Analyzer (PROVEAN, Choi *et al.,* 2012), CADD (Kircher *et al.,* 2014), MetaSVM (Dong *et al.,* 2015a) and MetaLR (Dong *et al.,* 2015a). Still the aforementioned tools have shown to lack the specificity of discriminating driver from passenger mutations involved in carcinogenesis. In the following paragraphs, we will delve into some ensemble scoring metapredictors for disease causing mutations and we will also discuss tools designed specifically for the detection of cancer driver mutations (Table 1), including CanPredict (Kaminker *et al.,* 2007), Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM, Carter *et al.,* 2009), Mutation Assessor (Reva *et al.,* 2011), Functional Analysis through Hidden Markov Models (FATHMM**,** Shihab *et al.,* 2013), and CanDrA (Cancer driver annotation, Mao *et al.,* 2013).

II.1 Ensemble Predictive Methods

Metapredictors or ensemble algorithms combine the results of multiple predictors and are therefore expected to outperform any individual methods they included. The rationale behind ensemble methods is to exploit the complementary performance of different prediction algorithms (Gonzalez-Perez & Lopez, 2011). In recent years, some metapredictors have been developed to answer some of the limitations of individual methods mentioned earlier.

Condel (Gonzalez-Perez & Lopez, 2011) algorithm was originally a weighted average of the normalized scores (WAS) approach combining five predictive tools (SIFT, LogRE, MAPP, Mutation Assessor and Polyphen2) which classify nsSNVs as either deleterious or neutral. The method sequentially (1) fits each of the five predictors to the same lists of neutral and deleterious mutations screened from two human mutations databases (HumVar and HumDiv) to generate predictive scores; (2) for each predictor, builds the probability of corresponding complementary cumulative distribution of the outputted scores to compute the weights and also plots the receiver operator characteristic (ROC) curve; (3) integrates the outputted scores of all the five methods using two techniques of simple vote score (SVS) and weighted vote score (WVS) to obtain internal scores; (4) combines the internal scores of each method (SVS and WVS) through a simple average score (SAS) and weighted average score (WAS) to conduct the classification. Higher scores and higher weights (probability that a mutation is not a false positive) are predictive of deleterious mutation whereas lower scores and lower weights (the probability

that a mutation is not a false negative) are indicative of neutral mutations. The authors conducted comparative studies on the four techniques used to incorporate scores from five methods and concluded WAS is the best. WAS did outperform its competitors (SVS, WVS, and SAS) as well as the five individual predictors assayed in the analysis. The current web version integrates only the scores of two predictive methods (MutationAssessor and FatHMM). In Condel, variants are predicted to be either deleterious (scores ≥0.468) or neutral (scores <0.468).

MetaSVM and ***MetaLR*** (Dong *et al.*, 2015a) are two ensemble scoring machine learning algorithms for accurate prediction of deleterious nsSNVs. Both algorithms combine individual scores from ten popular predictors [six function prediction scores (SIFT, PolyPhen-2, MutationTaster, Mutation Assessor, FATHMM and LRT), three conservation scores (GERP++ (Davydov *et al.*, 2010), MutationTaster2 (Schwarz *et al.*, 2014), SiPhy (Garber *et al.*, 2009) and PhyloP (Cooper *et al.*, 2005) and the maximum minor allele frequency (MMAF) from the 1000 Genomes Project populations (Abecasis *et al.*, 2010) to predict the deleteriousness of each mutation. The MetaSVM algorithm uses a two step-approach. Beginning with a list of mutations curated from Uniprot database (Uniprot Consortium, 2011), it obtained output scores of each of the ten algorithms for a training set of mixed known deleterious and non-deleterious mutations from dbNSFP (Liu *et al.*, 2011; Liu *et al.*, 2013). Then, it applied a support vector machine (SVM) technique on the outputted scores to build a SVM with linear kernel, radial kernel and polynomial kernel. MetaLR follows the same design; however, it used a different algorithm based on logistic regression. The comparison of MetaSVM and MetaLR ensemble methods showed that LR scores slightly outperform SVM scores. Most importantly, MetaSVM and MetaLR exhibited higher separating performance compared to any single algorithms that they combined.

II.2 Cancer Specific Prediction Methods

CanPredict (Kaminker *et al.*, 2007) webserver and algorithm is one of the first machine learning techniques designed specifically to discriminate driver from passenger mutations in cancer. The method uses a random forest classifier to integrate the results of three algorithms including two methods predicting the effect of a particular mutation on a protein (SIFT and Pfam-based LogR.E-value metric) and the Gene Ontology Similarity Score (GOSS) that measures the similarity between known cancer genes and any given mutated gene. The algorithm begins with a list of mutations curated from two databases: known somatic cancer mutations were obtained by searching COSMIC and nsSNVs from dbSNP. Briefly, a random forest classifier was trained upon the individual scores generated by three

algorithms (SIFT, LogR.E-value, and the GO log-odds score) with a training set of mixed 200 deleterious and 800 non-deleterious mutations. The output predict whether an amino acid change is likely cancer-related or not.

CHASM (Carter *et al.,* 2009) algorithm considers 49 predictive features and uses a random forest which is an ensemble of decision trees used to discriminate driver from passenger somatic missense mutations. The training dataset contains a mixture of driver mutations or positive set (curated from the COSMIC database and other cancer-related databases) and passenger mutations or negative set synthetically generated. The final score for each mutation is the fraction of votes of individual tree classifying a mutation as driver or passenger. The researchers evaluated the performance of CHASM using two threshold-independent measures including the ROC and the Precision-recall (PR) curves. They found values of area under the curve (AUCs) of 0.91 and 0.79 for the two measures respectively. Additionally, they concluded that CHASM performed better and exhibited higher specificity, sensitivity, and precision than previous classifiers (PolyPhen's PSIC score, SIFT, CanPredict, KinaseSVM and SIFT-PolyPhen consensus).

MutationAssessor **(**Reva *et al.,* 2011**)** uses evolutionary conservation information to differentiate functional from nonfunctional mutations. The prediction of the functional impact scores (FISs) of "switch of function" mutations involved in cancer is the main strength of MutationAssessor over its competitors. Switch of function (SOF) mutations can switch between loss of function (LOF) and gain of function (GOF) mutations. According to Reva *et al.* (2011), SOF accounted for more than 5% of 10K predicted functional mutations in COSMIC database**.** From a list of genomic coordinates or proteins coordinates, the algorithm searches for sequence homologs in the UniProt database. Based on the homologs it builds a partitioned multiple sequence alignments (MSA) using a combinatorial entropy measurements that differentiate the conserved residues found in aligned families (conservation scores) from the specificity residues clustered in subfamilies (specificity score). The FIS are the averages of the conservation score and the specificity score. MutationAssessor provides four classes of predictions including neutral, low, medium and high. Neutral and low variants are predicted to have no impact on protein function; whereas, medium and high variants are predicted to alter protein function. The authors obtained a specificity of 79 % when the FIS was applied to a large dataset of mixed polymorphic and disease-associated variants.

CanDrA (Mao *et al.,* 2013) is a machine learning method which integrates 95 structural and evolutionary features computed by ten functional prediction algorithms [CHASM's SNVBOX (Carter *et*

al., 2009), SIFT, MutationAssessor, ENSEMBL Variant Effect Predictor (McLaren et al., 2010), Mutation Assessor, ANNOVAR (Wang *et al.*, 2010), SIFT, PolyPhen-2, CONDEL, PhyloP, GERP ++, and LRT] to predict cancer type specific missense driver mutations. For any cancer type, a driver mutation is defined as a mutation that occurs in a gene mutated in this cancer type and fulfil one of the following requirements: (i) it is observed in at least 3 primary tumor samples (regardless of cancer type), or (ii) its site intersects at least 4 mutations (including indels, dinucleotide or trinucleotide mutations), or (iii) it is centered in a 25 bp region that intersects at least 5 mutations in the COSMIC database. Passenger mutations are those absent within a 31bp window in COSMIC cancer census gene, but seen only once in a primary tumor of a cancer type. The algorithm was applied to 15 types of cancer and all cancer combined. The algorithm uses a weighted support vector machine to compute two scores: the CanDrA_GEN score predicts whether a mutation is a generic driver for all cancers and the CanDrA_CTS score predicts whether a mutation is a driver in a given cancer type. CanDrA classifies mutations into three categories: driver, no-call, and passenger. Using a 10-fold cross-validation and the AUC, CanDrA outperformed CHASM, MutationTastor, and MutationAccessor when discriminating driver from passenger mutations.

FATHMM (Shihab *et al.*, 2013) combines sequence-based conservation features (the alignment of homologous sequences and conserved protein domains) and pathogenicity weights (indicative of the tolerance of the corresponding model to mutations) in a Hidden Markov Model (HMM) based method to predict the functional consequences of missense mutations. FATHMM offers two different algorithms. The first algorithm covers regions falling within conserved protein domains and takes protein sequences and nsSNVs as inputs. Next, using the SUPERFAMILY and Pfam databases, protein domain annotations are generated. Variants are mapped into corresponding HMMs match states and significant HMMs are extracted. The relative frequencies of cancer-associated (CanProVar) and putative neutral polymorphisms (UniProt) found in these regions are computed. The final predictive score is weighted using cancer-specific pathogenicity weights. The second algorithm is for variants falling outside conserved protein domains. It is similar to the above-mentioned method except aligning homologous sequences from SwissProt/TrEMBL database and using the JackHMMER algorithm to constructs HMMs. FATHMM was reported to perform better than SIFT, PolyPhen-2, SPF-Cancer (Capriotti and Altman, 2011) and CHASM, but shows comparable performance with MutationAssessor and TransFIC when tasked to discriminate driver mutations from passenger mutations. FATHMM is the algorithm of choice used in the COSMIC database to generate mutational impact scores.

II.3 Comparison of the performance of multiple predictive methods

Many methods have been discussed above and each of them comes with its own specific biases. These tools may yield conflicting results which hampered their comparability and reproducibility (Gnad and colleagues, 2013; Castellana and Mazza, 2013; Martelotto et al., 2014) and many of them have not been comprehensively compared to one another. To better understand these issues, Gnad and colleagues (2013) have assessed the utility and performance of eight nsSNV prediction tools including two cancer-specific methods and six general methods (CHASM, mCluster, Condel, SIFT, PolyPhen2, logRE, SNAP, and MutationAssessor), regarding their coverage, accuracy, availability and dependence on other tools. A positive dataset of 2,682 variants was curated from COSMIC and a negative set of 7,170 from dbSNP. They ran each individual classifier on the datasets and also combined individual methods into metapredictors using the Condel technique of weighted average scores. They concluded MutationAssessor outperformed all its competitors including any other metapredictors they constructed: "Polyphen-2 and MutationAssessor", "SIFT and MutationAssessor", and "Polyphen-2 and SIFT".

A comprehensive benchmarking study (Martelotto *et al.*, 2014) compared the performance of 15 mutation effect prediction algorithms and their agreement using 3591 experimentally validated cancer SNVs. Eleven individual algorithms [SIFT, PolyPhen-2, Mutation Assessor, PROVEAN, VEST, MutationTaster, FATHMM (cancer, missense) and CHASM (breast, lung, melanoma)] and five metapredictors [Condel and CanDrA (breast, lung, melanoma)] were tested. The agreement among cancer-specific predictors was more consistent compared to non-cancer-specific predictor which exhibit less agreement according to their unweighted Cohen's Kappa coefficient. FATHMM (cancer) was the most accurate single predictor whereas CanDrA (lung) outperformed other metapredictors. However, using a composite score (sum of sensitivity, specificity, positive predictive value and negative predictive value) to evaluate each predictor's overall performance, the authors found that CHASM (lung) was superior among the individual predictors and CanDrA (lung) was the best among the metapredictors. The study revealed that compared to both the best individual algorithm and the best metapredictor, some combination of mutation effect predictors, such as CHASM (breast) and MutationTaster, significantly improved accuracy, composite score and more importantly negative predictive value; thus reducing the amounts of false-negative predictions considerably.

III. Algorithms for identifying driver genes (gene level approaches)

To be a driver, a mutation must be functional and alter the activity of proteins at some stages of tumor development. A driver gene needs to contain at least one driver mutation. Due to genetic heterogeneity among different cancer patients and limitation in sample size, the recurrence and clustering of driver mutations are often better observed at gene or pathway level. Here we discuss three approaches of gene level analysis including mutation rate based approaches, function prediction based approaches, and ensemble approaches. Mutation rate based approach first estimates the background mutation rate of the assumed non-functional mutations (e.g. synonymous, non-coding, etc.) of the gene and then test whether the observed rate of functional mutations (e.g., non-synonymous) is higher than the BMR. Function prediction based approach exploits a similar idea but avoids the difficulties of estimating BMR and instead interrogates the distribution of functional predictions among mutations. Assuming most of the observed mutations are passenger mutations and given the quantitative predictions for the functional impact of the mutations, genes harboring more functional mutations are more likely to be drivers than those harboring less functional mutations. An alternative approach based on clustering hypothesizes that passenger mutations are randomly distributed along the genome, while driver mutations are concentrated in functionally important regions, or around specific functional sites. The goal is to detect genes that contain non-randomly distributed mutations in functionally important regions or sites.

III.1 Mutation Rate Based Approaches

Mutation rate based approaches [Mutational Significance in Cancer (***MuSiC***, Dees *et al.*, 2012); Mutation Significance (***MutSig, MutSigCV, MutSigFN***, Lawrence, *et al.*, 2013); ***ActiveDriver*** (Reimand *et al.*, 2013), ***ContastRank*** (Tian *et al.*, 2014), and Driver Genes and Pathways (***DrGaP***, Hua *et al.*, 2013)] hypothesize that mutation rates of functional SNVs are higher than non-functional SNVs in driver genes while that is not true in non-driver genes. Basically, they rely on mutation frequency of sSNVs and SNVs in non-coding regions to build a null model of the background mutation rate (BMR, defined as the probability that a base is mutated by chance) which is usually unknown (Table 1).

MuSiC (Dees *et al.*, 2012) is one of the well-known mutation rate based approaches with the benefit to discriminate not only driver genes, but also altered pathways and gene sets. The algorithm accepts four input files [list of somatic mutations in mutation annotation format (MAF format), a list of mapped reads of matched tumor and normal samples in BAM files, a set of target regions for genes in MAF file and some clinical data (categorical and or numeric)]; then it performs seven statistical analyses consecutively in its final execution module "MuSiC Play". The significantly mutated genes (SMG) test

incorporates gene size and BMR to discover genes with significantly higher mutation rate than expected by chance. SMG test outputs composite p-values and false discovery rates (FDRs) for each gene using three tests: Fisher's Combined P-value test (FCPT), Likelihood Ratio test (LRT), and the Convolution test (CT). The Significantly mutated pathway/genes set analysis (PathScan) searches for significant mutations in the whole pathway defined by Kyoto Encyclopedia of Genes and Genomes (KEGG) or other databases. The Mutation Relation test (MRT) assesses each set of two genes in the MAF file to label them as harboring positively correlated or negatively correlated mutations. The Pfam annotation classifies genes by proteins domains and adds a new column of Pfam annotation domains to the input MAF file. The Proximity analysis searches for dense cluster of mutations around each identified mutation within a window of seven amino acids both upstream and downstream. The COSMIC/OMIM analysis cross-references the list of predicted mutations in the COSMIC and OMIM databases to find variants that have been previously reported. The Clinical Correlation test (CCT) uses a generalized linear model to find mutations or genes associated with a specific clinical feature such as tumor subtype.

From MutSig to MutSigCV, MutSigCL and then MutSigFN

MutSig **(**Lawrence *et al.*, 2013**)** algorithm identifies genes that are more frequently mutated compared to the BMR. The algorithm assumes the BMR is constant across the entire genome. It inputs three files (list of somatic mutations and indels in MAF format, a coverage table file and a covariates table file lists genomic features for each gene). For each gene found in a cohort of patients, it estimates the distribution of its mutations by aggregating them in each tumors sample, and then tallies them and computes gene score and p-values. The next step compares this distribution to the BMR of observed sSNVs present in the same gene. The original version focuses only on one signal that is the abundance of mutations in the impaired gene. The current version ***MutSigCV*** accounts for mutational heterogeneity as it incorporates features such as DNA replication time, chromatin state (open/closed) and gene-expression level (highly transcribed vs. not transcribed at all) to compute gene-specific BMR, patient-specific mutation rate and spectrum. The inclusion of these additional features has shown to significantly improve the BMR estimation because they are highly correlated to the BMR. MutSigCV was tested on the same datasets as MuSic and MutSig; and uncovered only a small set of 11 driver genes out of a longer list of 450 genes classified as driven tumorigenesis by its two competitors (Lawrence et al., 2013). This drastic decrease of the number of driver genes is seen as a crucial improvement in finding truly biological significant cancer drivers by the developers of MutSigCV.

Other improvements include the development of ***MutSigCL*** that exploits the clustering of mutations. This algorithm computes the probability that clustered mutations are due by chance. Another algorithm ***MutSigFN*** takes the functional impact of each mutation into account. Ultimately, the novelty of the current MutSig algorithm is that it can combines three complementary signals of positive selection (CV, CL and FN) to separate drivers from non-driver genes; it can also computes a joint p-value for CL and FN to account for any correlation between clustering and conservation (Marx V, 2014).

ActiveDriver **(**Reimand *et al.*, 2013**)** algorithm uses a gene-centric logistic regression model to predict genes disrupting protein phosphorylation. It focuses mainly on phosphorylation sites and searches for kinase domains harboring more mutations than expected by chance. The method can be conducted at the gene level or at the module level (pathway and network analysis). At the gene level, ActiveDriver assesses the significance of mutation enrichment (or depletion) in each active site by integrating multiples features [mutation frequency of predefined regions (a phosphorylation site and its neighboring amino acids), distribution of active sites, their positions with respect to mutations (direct and flanking) and structured and disordered regions of proteins] in the analysis. It use a likelihood ratio test to compare the null model of observing same mutation rate at all the phosphosite regions to the alternative of seeing higher or low phosphosite-specific mutation rate considering the gene-wide mutation rate.

ContrastRank **(**Tian *et al.*, 2014**)** algorithm assumes that (1) rare variants especially nsSNVs are more likely to have an impaired functional impact and (2) genes that are rarely mutated in normal samples are more likely to be driver genes if frequently mutated in multiple tumor samples. It considers the background mutation rate for each gene as the maximum value of its mutations rates in the normal TCGA samples and in the 1000 Genomes Project samples. Here is how this algorithm works. Obtain minor allele frequency (MAF) of nsSNVs from the 1000 Genomes Project and also search for putative deleterious variant (PDV) in TCGA which are nsSNVs with MAF $\leq$0.5% in 1000 Genomes. Identify putative impaired genes (PIGs) harboring at least one PDV. Compute the putative defective rate (PDR) of each gene which is the fraction of samples in which the genes harbors at least one PDV. Calculate a gene prioritization score for each gene and use it to differentiate TCGA normal and tumor samples. Using the binomial distribution, a gene prioritization score is the probability that gene g is labelled as PIG in k or more samples over N tumor samples. ContastRank was applied to three lists of known cancer-related genes curated from previous studies: the Bushman list containing 2,125 genes (Bushman,

2013), the COSMIC Census list containing 522 genes (Forbes *et al.*, 2011) and the Vogelstein list containing 125 driver genes (Vogelstein *et al.*, 2013). The authors reported for all three lists, ContrastRank consistently outperforms MutSigCV using the AUC.

Overall, mutation rate-based methods are hindered by some limitations. The main challenge is to build a strong model of background mutation rate that significantly reduces false discovery rates (Tamborero et al., 2013a). They are also prone to overlook functional driver genes with mutations occurring at very low rate particularly with small sample sizes (Pon and Marra, 2014; Creixell *et al.*, 2015).

III.2 Function Prediction Based Approaches

Function prediction based approaches [***OncodriveFM*** (Gonzalez-Perez & Lopez-Bigas 2012); ***OncodriveCLUST*** **(**Tamborero *et al.*, 2013b**),** ***Oncodrive-CIS*** (Tamborero *et al.*, 2013c), ***Oncodrive-ROLE*** (Schroeder *et al.*, 2014), ***InVeX*** **(**Hodis *et al.*, 2012**)**] assume that across a cohort of samples, candidate driver genes are more likely to contain mutations with higher functional impact than those in non-driver genes (Table 1). The main advantage of these approaches is that they avoid the difficulties of estimating BMR and further refine the function classes for SNVs. Additionally; they can be applied to the genome of a single cancer patient as well as across the genomes of many patients.

From Oncodrive-FM to Oncodrive-CLUST and Oncodrive-CIS

OncodriveFM (Gonzalez-Perez & Lopez-Bigas 2012) method was introduced to discover driver genes or genes modules exhibiting bias toward accumulation of mutations of high functional impact (FM bias) across cohort of tumor samples. The algorithm takes a list of variants (nsSNVs, stop-gain SNVs, frameshift indels, and sSNVs) found in each gene in all samples with their corresponding SIFT, PolyPhen2, and Mutation Assessor functional impact (FI) scores. The next step averages the three FI scores of mutations occurring in the same gene or pathway across all tumor samples; then, assesses whether the distribution of the averaged FI scores diverge from the null distribution of averaged FI scores. Null distributions can be generated by randomly pick with replacement of (i) somatic variants in the tumor (internal null distribution) or (ii) nsSNVs found in genes exhibiting the same broad biological process across human populations (external null distribution). Genes with high FI averages are prioritized as drivers whereas genes with low averages FI are passengers. The algorithm computes for each gene or pathway an integrated p-value predicting how biased it is compared to the null distribution. Candidate driver genes are classified as RFM (Recurrent and FM biased) describing those with high FI, or as LRFM (Lowly Recurrent and FM biased) or as RnFM (Recurrent but not-FM biased).

OncodriveCLUST (Tamborero *et al.*, 2013b) algorithm exploits the tendency of certain genes to harbor mutations clustered in certain regions of the protein sequence. The algorithm inputs three files; one of protein affecting mutations (nsSNVs, stop-gain SNVs, and splice site mutations) found in each gene across tumor samples, the second with coding sSNVs that will be used for the background model and the third contains CDS lengths of gene transcripts. For each gene the algorithm builds a binomial cumulated distribution function considering gene length and percentage of gene mutations at each position. Then, it searches for positions with probability of occurrence above the threshold and further groups them into mutations clusters joining positions within a window of 5 or less amino acid residues. Next, its scores each cluster and produce a gene score by tallying the scores of all clusters. The significance value of each gene is assessed by comparing its gene clustering score distribution to the null distribution obtained from gene clustering scores of sSNVs. The method computes transformed p-value measuring the clustering bias. In COSMIC database and other projects of the Cancer Genome Atlas, OncodriveCLUST has identified known cancer driver genes and candidate driver genes not detected by OncodriveFM and MutSig.

Oncodrive-CIS (Tamborero *et al.*, 2013c) algorithm identifies driver genes harboring copy number alterations (CNAs) biased toward misregulation (overexpressed or underexpressed). The program assesses the accurate expression impact of CNAs through tumors and normal samples and computes the functional impact of CNAs considering gene dosage and expression. The algorithm sequentially computes the expression impact scores (EIS) for each gene, calculates two standard scores for each gene (Z_{NORMAL} and Z_{TUMOR}), combines these two scores using the Stouffer's method to obtain a measurement of the gene expression bias (ZCOMB), and uses the combined score to prioritize genes. Genes with a higher score are predicted to have larger bias towards misregulation caused by the CNAs. Compared to other algorithm, this method is independent of the frequency of the CNAs and therefore can identify driver genes with low CNA frequency. Additionally, the analysis of gene misregulation due to changes of their CNAs is done across tumor and normal samples.

InVeX **(**Introns Vs Exons, Hodis *et al.*, 2012**)** leverages intron and untranslated (UTR) sequences in a gene locus to control for gene-specific basal mutation rates, and have found to be particularly effective at identifying driver genes in highly mutated tumors such as cutaneous melanoma. For each gene, the method tallies the mutations across all samples to compute an observed mutation burden of non-silent mutations. Next, it uses a permutation-based approach to generate a null distribution of all mutations (exon, intron and UTR) where the locations of mutations are randomly permutated 10^8 times

on a per-patient, per-trinucleotide-context basis across the covered exon, intron and UTR. The non-silent mutations burden from the null is compared to the observed burden to identify genes harboring positively-selected non-silent mutations. Additionally, mutations are weighted with their PolyPhen-2 p-value; mutation with the largest weight is identified per sample and the sum of these largest weights represents the functional mutation burden. Synonymous mutation burden is the number of samples with ≥1 synonymous mutation. Loss-of function (LoF) mutation burden is the number of samples with ≥1 nonsense mutation, frameshift indel, or splice site mutation.

III.3 Ensemble Approaches

Because each gene level prediction method has a specific combination of features, it has been suggested that combine results derived from diverse methods will significantly increase accuracy. One interesting study by Tamborero *et al.*, 2013a integrated mutation rate based methods and functional prediction based methods. The five methods (MuSiC-SMG, MutSigCV, OncodriveFM, OncodriveCLUST, and ActiveDriver) were used to predict different signatures of positive selection across the whole pan-cancer dataset of 3,205 samples from 12 cancers types and effectively uncovers 291 "high-confidence drivers" (HCDs) genes identified by at least three classifiers. Each of the aforesaid methods identified respectively 155, 177, 55, 63 and 131 HCDs, but only 12 genes were in common. In another study (Cheng *et al.*, 2014) the predictive scores from eight algorithms (MutsigCV, Simon, OncodriverFM, ActiveDriver, MEMo, Dendrix, MDPFinder (Mutated Driver Pathway Finder, Zhao *et al.*, 2012), and NetBox) were summarized into a new database "DriverDB" hosting 6,079 mutation profiles from 14 cancer types. From this database, analysts can assess data at three level of biological analysis including gene oncology, pathway and protein/genetic interaction. The web interface provides a cancer section, a gene section as well as a meta-analysis section. The eight methods were applied to the Glioblastoma multiform' (GBM) dataset and predicted 14 driver genes identified by at least 4 classifiers. Ten of these genes were already known driver genes in GBM and the remaining 4 have been reported in other types of cancer.

It is important to note that many of the methods have been integrated under a single platform or pipeline. As discussed above, MutSigCV has seen the addition of and MutSigFN. Along the same line, the Oncodrive methods family regroups OncodriveFM, OncodriveCLUST, OncodriveCIS and OncodriveROLE. Moreover, the Integrative OncoGenomics (IntOGen) platform (Gonzalez-Perez *et al.*, 2013) was created to run some of these methods in a single pipeline to thoroughly analyze tumor samples from various cohort of whole genome sequencing or whole exome sequencing projects.

IV. Algorithms for identifying driver gene modules (gene module level approaches)

One weakness of previous methods is that they are unable to account for any interaction between genes or protein complexes. To overcome that, holistic approaches including pathways or network analyses focus on groups of genes called "modules". Compared to mutation or gene level analysis, the commons of these holistic approaches is to regroup genomic alterations by exploiting existing knowledge on cellular mechanism, biological pathway or molecular network. Because cancer is a complex disease and driver mutations have been identified in many genes across the genome, these methods hypothesize that a single gene is hardly accountable for a given phenotype; but rather there are multiples genes networking with one another in a pathway to deeply affect the genotype (Leiserson *et al.*, 2013; Liu *et al.*, 2015; Raphael *et al.*, 2014, Pon and Marra, 2014). A pathway refers to the smallest functional unit of a network of genes acting together to perform a single task; whereas a network is a combination of multiple pathways. These methods assume most driver mutations are rare, hence a driver pathway may harbor no more than one driver mutation per gene and per patient. In this perspective, most patients end up with a single mutation in any given significant pathway. For that reason, we might be confronted with driver genes harboring recurrent mutation accruing at low frequency and unable to pinpoint the real impaired ones. If these genes are involved in the same pathway; at least, it will be reasonable to suggest the set of genes is the driver of the cancer. A closer look at any particular type of cancer data supports the fact that individual tumors are unique in the sense that they bring different genomic alterations to oncogenesis; however, these unique variations usually altered significantly the same biological pathways across different tumors. Three classes of methods have been used to identify driver pathways: known gene set approaches, de novo gene interaction approaches, and interaction networks approaches (Table 1).

IV.1 Known Gene Set Approaches

The first group of methods includes MuSiC; Gene Set Enrichment analysis (GSEA, Subramanian *et al.*, 2005) and PathScan (Wendl *et al.*, 2011) among others. These methods rely heavily on prior knowledge of genes as they attempt to identify genes that are recurrently mutated in known genes sets. The majority of these techniques use enrichment analysis to search for functionally mutated gene sets or those that are differentially expressed in known pathway databases, protein complexes databases and network databases. These methods ignore other important information (genes and protein-protein interaction, topology of the pathway, type of genes, gene position on the pathways) and by such, they

use prior knowledge to simply list impaired genes in a pathway as if they have equal importance (Mitrea *et al.*, 2013, Raphael *et al.*, 2014).

(***GSEA***, Subramanian *et al.*, 2005) method was designed to assess whether predefined gene set shows concordant expression differences between two phenotypes. The algorithm starts with a list of genes (*L*) ordered based on the degree of association between their expression levels and two phenotype groups (tumor and normal samples, diseased and normal, etc.). Then, for any given predefined gene set (*S*), it uses the Kolmogorov-Smirnoff method to test whether *S* is enriched in *L*. The enrichment score (*ES*) is the degree to which members of *S* are overrepresented either at the top or at the bottom of *L*. This is done by running down *L* and starting from 0 applying a weighted increment if a gene from *S* is seen or a weighted decrement if the gene is absent. The *ES* score refers to the maximum deviation from zero found when running down *L* and it is calculated for each of the two phenotypes. Next a permutation test identifies significantly enriched genes sets in each phenotype. Therefore, gene sets at the top have positive ES; whereas, those located at bottom have negatives ES. In other words, highly expressed genes sets from one phenotype will be found at the top whereas the opposite is true for the second phenotype.

PathScan **(**Wendl *et al.*, 2011**)** algorithm tests for enrichment of mutations in predefined gene sets for each patient. Under the null hypothesis, somatic mutations are assumed to be randomly distributed; most genes harbor zero mutations which characterized the background mutation rate. The test considers each gene as either mutated (one or more mutations) or not mutated. The method identifies metabolic pathways that are significantly mutated by considering the effect of gene length and the distribution of mutations by combining p-values of enrichment across tumor samples using the Fisher–Lancaster theory. The novelty of PathScan is that it accounts for the distribution of mutations among samples and also the effect of gene lengths because larger genes are more likely to be mutated compared to small genes within a pathway. The method starts with a predetermined set of genes and can be used in a single patient or a cohort of patients. PathScan can be used as a single tool, or through MuSiC as mentioned earlier.

IV.2 De Novo Gene Interaction Approaches

The second group of approaches was designed to construct driver pathways harboring genes exhibiting the mutual exclusivity criterion. This property hypothesizes that multiple functionally related driver genes involved in the same pathway will harbor genomic alterations that do not overlap; in other words, those genes are not impaired all together in the same cancer patient. Accordingly, a single driver gene in an individual pathway can be sufficient to turn it into a putative driver pathway and even disturb

many other pathways. So their commonality is to discover mutually exclusive genes sets at the patient level. This group includes HotNet (Vandin *et al.,* 2011), Recurrent and Mutually Exclusive (RME Miller *et al.,* 2011); Dendrix (Vandin *et al.,* 2012); Multiple Pathways De novo Driver Exclusivity (Multi-Dendrix, Leiserson *et al.,* 2013); MDPFinder (Mutated Driver Pathway Finder, Zhao *et al.,* 2012); CoMEt (Combinations of Mutually Exclusive Alterations, Leiserson *et al.,* 2015), etc. The main advantage of these methods is their ability to detect genes with rare mutations (Liu *et al.,* 2015) and also novel genes as they are independent of any prior knowledge.

Dendrix (Vandin *et al.,* 2012) algorithm assumes the driver pathways contain sets of genes, domains or nucleotides whose mutations concurrently have a high coverage (most patients have at least one mutation in the set), and exhibit high pattern of mutual exclusivity (most patients have exactly one zero mutation in the set). From a list of somatic mutations, the algorithm constructs a binary mutation matrix. Then it chooses the best pair of genes and using the Markov Chain Monte Carlo (MCMC) it interactively adds the best genes and builds larger sets of genes with both high coverage and approximate exclusivity. The Dendrix method suffers from two disadvantages: mutations present in different pathways are not always mutually exclusive and some mutations demonstrate in contrary a pattern of co-occurrence in many patients. ***Multi-Dendrix*** (Leiserson *et al.,* 2013) algorithm is an expansion of Dendrix as it combines SNVs and CNVs data into a mutation matrix. It uses the Integer Linear Programming (ILP) to identify sets of genes with high coverage and simultaneously find multiple driver pathways.

IV.3 Interaction Networks Approaches

The third group of methods uses prior gene information to search for protein subnetworks that recurrently mutated than expected by chance, and they also strive to account for interaction information among genes. This group of methods includes PARADIGM (Vaske *et al.,* 2010**),** NetBox (Cerami *et al.,* 2010) and Mutually Exclusive Modules in Cancer (MEMo, Ciriello *et al.,* 2012) and MUDPAC (Liu & Hu, 2014) among others**.**

PARADIGM (Vaske *et al.,* 2010**)** method exploits existing pathway information (Reactome, KEGG, National Cancer Institute's Pathway Interaction Database) in order to identify recurrent pathway perturbations. The algorithm posits that different genetic alterations can be observed among cancer patients, and they usually occur in common pathways. The algorithm predicts the degree to which the activities of a given pathway are altered in a particular patient by combining different types of genomic data (gene expression, CNV, etc.) and information found in pathway databases into a matrix of

integrated pathway activities (IPAs). The matrix is later used for unsupervised clustering paired with survival analysis. The algorithm uses a probabilistic approach to construct a factor graph model that illustrates interactions and dependencies between genes where the nodes refer to different biological activities within the cell. For instance, each gene has a set of nodes describing its presence or absence, whether or not there was a DNA transcription into RNA, and later RNA translation into proteins and so on. Therefore, positive or negative scores refer to active or inactive nodes. Aggregated pathways scores are ranked by averaging nodes scores and associated p-values are computed. The novelty of this method is the ability to identify cancer subtypes and its application on patient specific dataset. PARADIGM outperformed two competitors including GSEA and signaling pathway impact analysis (SPIA, Tarca *et al.*, 2009) in a comparative study.

MEMo (Ciriello *et al.*, 2012**)** algorithm identifies recurrently mutated set of genes appearing in the same pathways or involving in the same biological processes and also exhibiting patterns of mutually exclusivity across many patients. The method takes advantage of prior biological knowledge seen in large databases of known pathways, interaction networks, proteins structure or protein sequence conservation. The algorithm uses correlation analysis and statistical tests to search for mutually exclusive driver networks considering three properties: (i) member genes are recurrently altered across tumor samples, (ii) member genes are likely belong to the same pathway and (iii) individual alterations in the same module are mutually exclusive. The algorithm starts with filtering significantly mutated genes (SMG), genes with recurrently altered copy number amplification or deletion, both exhibiting concordant mRNA expression. The obtained list of altered entities is used to build a binary event matrix of significantly altered genes. Next, it identifies gene pairs likely involved in the same pathway and builds a network of gene pairs and extract cliques. The mutual exclusivity criterion is assessed using random permutations to the test significance of each clique for overlap.

Mutex (Babur *et al.*, 2015) expanded beyond ***MEMo*** and the aforementioned mutual exclusivity testing methods by searching for groups of mutually exclusively altered genes that have a common downstream event. Introducing additional biological constraints effectively reduced the search space of pathways and modules, and improved the power of the analysis. Consequently, Mutex outperformed early methods such as RME, Dendrix, MEMo, MDPFinder, Multi-dendrix and ME (Mutual exclusivity, Szczurek *et al.*, 2014) in simulation and generated interesting novel results using TCGA data.

MUDPAC (Liu & Hu, 2014**)** was developed to identify pathways mutated together across multiple samples with accounting for pathway interaction. The method is a two-step approach. The first

step uses a mutational pathway enrichment analysis where it constructs a mutation matrix from two input files (mutations data and pathway data). For each gene it computes two scores. The gene mutation factor (MF) score is a weighted difference between its non-synonymous and synonymous mutations rates; whereas, the gene interaction factor (IF) account for mutational patterns such as mutual exclusivity or topology distance. The two scores are combined into a gene score and these ranked scores are used to find candidate pathways harboring more genes with higher scores than expected by chance. Genes outside of the pathway are assigned scores and the statistical significance of each pathway is computed correcting for multiple testing. The second step looks for collaborative driver pathways. Starting with the pathway with the highest coverage rate (number of sample where the pathways harbors at least one nsSNV), a greedy algorithm iteratively adds new mutated pathways into a previous collaborative set with the Maximal Coverage Rate (MCR) until it discovers combination of pathways with genes linked uniquely to a particular cancer subtype. The MCR is defined as the percentage of samples covered by current set of driver pathways.

The major advantage of gene module approaches is that they consider gene interactions while gene level methods do not. Additionally, they can integrate a broader range of data types including SNVs, CNVs, expression data, etc. In addition, they can identify driver genes with rare or low frequency mutations which are undetectable by gene based approaches. On the other hand, module approaches are built on certain summaries of genes, therefore they shall be considered as complements other than replacements of the gene level approaches. One such example is the ensemble classifier machine learning method (EC**,** Liu *et al.,* 2015). It combines predictions of ten classifiers including five gene-level methods (OncodriveFM, OncodriveCLUST, MutSig, ActiveDriver and Simon) based on mutation rate and five module level methods [Functional linkage network (FLN, Linghu *et al.,* 2009), Functional Linkage Network Partitioning (FLNP (Huang *et al.,* submitted), NetBox, MEMo, and Dendrix]. The EC algorithm is described as follow. The individual scores obtained from the ten predictors are inputted into an $n \times m$ matrix G of driver candidates where n labels the genes, m are the number of features (10 algorithms) and an entity g_{nm} is equals 1 if gene n is classified as a diver by predictor m or equals to 0 otherwise. Next the meta-classifier DECORATE (Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) is used to build ensemble classifier by interactively training a base classifier on the union of a training dataset and an artificial training dataset, which increase diversity. The final classification uses the posterior probabilities of each gene that are computed by averaging the results of four base classifiers including Naïve Bayes, Sequential Minimal Optimization

(SMO) algorithm, C4.5 decision tree, and Random forest. DECORATE has shown to have a predictive power superior to each of its individual predictors in separating driver genes from passengers.

V. Other aspects in driver prediction

V.1 Gene Fusion

In addition to methods developed for the identification of driver nsSNVs and CNVs, other sophisticated algorithms have been developed for the prioritization of fusion genes. A comprehensive characterization of current fusion detection tools aimed at identifying fusion-transcript candidates mostly from RNA-seq data is published elsewhere. **See also**: DOI: 10.1002/9780470015902.a0025848.

Concept signature (***ConSig***, Wang *et al.*, 2009) algorithm specifically searches and ranks biologically meaningful genes based on prior knowledge of association to certain signatures of "molecular concepts" (molecular interactions, pathways and functional annotations) of known cancer genes. The first step compiles molecular concept from diverse databases (Gene Ontology, Reactome, KEGG, Biocarta, HPRD, and Entrez Gene conserved domain), screens cancer-causal gene from Mitelman database, removes duplicates and maps the fusion or point mutation gene lists against the compendia of molecular concepts. Then the Fisher's exact test is applied to identify two sets of "concept signatures" that differentiate fusion genes from point mutation genes. The fusion concept signatures include concepts enriched for fusion genes and the mutation concept signatures are those enriched for point mutation genes. The next step assesses the "relevance" of each concept that is computed as the log10 of the number of fusion (or point mutation) genes present divided by the square root of the total number of genes in the concept. Finally, "fusion ConSig score" or "mutation ConSig score" of genes are generated by tallying all the relevancies of the fusion signature concepts (or mutation signature concepts) in the gene and normalized for the total number of assigned concepts. High fusion or mutation ConSig score is indicative of a high probability that this gene is driving tumorigenesis based of the strength of its association with signature of molecular concepts.

Oncofuse (Shugay *et al.*, 2013) algorithm uses a Bayesian classifier to assign a functional predictive score to already identified fusion genes responsible for tumorigenesis. It prioritizes gene fusions based on their probability of being a driver rather than a passenger event. The classifier integrates 24 features including 12 features "retained" and 12 features "lost" in the gene fusion process belonging to four

categories (six functional profile features, two promoter features, one 3'UTR feature, and three protein interaction features).

V.2 Genomic Variations in Noncoding Regions

The human genome comprises less than 2% coding regions; the remaining 98% are non-coding regions harboring regulatory elements with different biochemical tasks such as regulating gene expression. Most of the predictive algorithms described above fall under the discovery of genomic alterations in coding regions. Recently, analyses outside of coding regions have attracted great attention as the price of whole genome sequencing become more affordable for larger cohorts of patients. Weinhold and colleagues (2014) have investigated the role of noncoding regions in cancer as they characterized the landscape of functional noncoding alterations in matched tumor DNA and normal tissue of 20 types of cancers among 863 patients curated from TCGA. For this comprehensive analysis they applied three complementary methods. The hotspot analysis detected clusters within 50bp window. The regional recurrence analysis targeted annotated regulatory elements considering length and replication time. The transcription analysis searched for mutations affecting ETS binding sites. The study discovered recurrent mutations in regulatory elements upstream of PLEKHS1, WDR74, SDHD, and TERT promoters. Specifically, recurrent mutations in promoter TERT make this gene hyperactive causing an uncontrolled cell division. In another study, Fredriksson and colleagues (2014) assessed all genes for association between RNA-level changes and presence of somatic mutations in regulatory regions up to 100 kb across 505 tumor samples from 14 cancer types. They discovered relevant recurrent promoter mutations in several genes including a unique mutation-expression association for TERT as well as the CLPTM1L controlled by TERT promoter mutations. Li and colleagues (2015) take advantage of the ENCODE and modENCODE projects to investigate the functional role of noncoding variants found in high-occupancy target (HOT) regions. Using matched cancer cells with healthy counterpart and the associated genes, a search for HOT regions revealed a large cancer-specific HOT region in the promoter of gene BRCA2, and also around c-MYC gene, CANT1 gene, and FAS gene in hepatocellular carcinoma cells, prostate carcinoma cells, and malignant melanoma cells. The study concluded that cancer cells acquire cancer-specific HOT regions at key oncogenes through various mechanisms; thus HOT regions should be used as biomarkers to identify oncogenes.

SASE-hunter (***Signatures of Accelerated Somatic Evolution – hunter,*** Smith *et al.*, 2015) is the latest computational approach designed to identify regulatory segments in noncoding regions harboring excess of somatic mutations than expected by chance. The algorithm considers regional variation in

mutation rate, evolutionary conservation and genomic context while making the inference. The algorithm was applied to the promoter regions of protein-coding genes in a Pan-cancer dataset of 906 samples (from 14 cohorts from 12 different cancer types). They discovered SASE in the promoters of known cancer genes MYC, BCL2, RBM5 WWOX and other genes and these findings were associated with the hyperactivity of these genes, age of onset of cancer, aggressiveness of the disease, and survival.

Funseq2 **(**Fu *et al.,* 2014**)** was developed to annotate and prioritize regulatory somatic mutations in noncoding regions. The method was built around two main components. A small scale informative data context summarizes data from large-scale genomics and cancer resources such as gene lists, conservation, functional annotations and network centrality. The variant prioritization pipeline annotates non-coding mutations and prioritizes them against the data context using a weighted scoring scheme accounting for the relative importance of various features in a two-step approach. First it computes integrated variant core scores from diverse analyses using different features (in parenthesis) including functional annotations (regulatory and HOT regions), conservation analysis (GERP score>2, ultra-conserved elements and sensitive elements), nucleotide-level analysis (motif-breaking score and motif-gaining score) and network analysis (linking regulatory elements with genes centrality, differential gene expression, network centrality score). The core scores are tally of the weighted values of all features present in each variant. Each feature is weighted with the mutation patterns observed in the 1000 Genomes Project polymorphisms and features frequently observed in the polymorphism data are down weighted. The next step searches a recurrence database to generate variants' recurrent scores. Recurrent regulatory elements are those mutated in more than two samples and they exhibit higher scores compared to non-recurrent elements. Final score for each variant sums up core scores and recurrence score. High scores are indicative of high deleteriousness. The variant prioritization step can also highlight noncoding variants likely involved in tumorigenesis using prior knowledge of cancer genes, DNA-repair genes, a module to detect differentially expressed genes and other user annotations. Funseq2 was claimed to outperform the genome-wide annotation of variants (GWAVA, Ritchie *et al.,* 2014) and the Combined Annotation–Dependent Depletion (CADD, Kircher *et al.,* 2014) as it was the only method that identifies the TERT promoter mutation found in one medulloblastoma sample as a gain of function. Also, the method prioritized the mutation in 2nd place compared to 25th and 224th ranks for CADD and GWAVA respectively.

V.3 Personalized Cancer Driver Prioritization Tools

The ultimate goal of prioritizing cancer drivers is to enable precision medicine specifically at patient level. However, the majority of the tools described above are limited to population based cohort of normal and tumor samples size. In recent years, other methods have been developed to analyze personal genome and suggest personalized medicine. In a recent study, Rubio-Perez and colleagues (2015) devised a novel strategy combining a cancer driver database and a cancer driver actionability database. Their strategy can be implemented in three steps. The first step is to identify "driver events" (SNVs, CNAs, and gene fusions) using three tools (MutSigCV, OncodriveFM and OncodriveCLUST). From 6792 samples from 28 cancer types from 48 cohorts of patients the authors uncovered 475 driver genes (459 via mutations and 38 via CNAs and gene fusions), and 63% of the genes were mutated in less than one sample. The mode of action of the 459 mutational genes was then ascertained by OncodriveROLE, which categorized 169 genes as harboring activating mutations (either from gain of function or switch of function), 207 with loss of function and 83 genes as unclassified. Finally in silico drug targeting strategy discovered that a total of 96 driver genes could benefit (from FDA approved drugs, drug in clinical trial and preclinical ligands) and be targeted directly (74) or indirectly (13) and the remaining 7 from both therapies.

Another recent study by Dong *et al.,* 2015b proposed the ***iCAGES (integrated CAncer GEnome Score),*** an algorithm with three layers specifically tailored to accomplish three purposes: (i) analyze patient-specific cancer genomic data, (ii) prioritize personalized cancer driver events and (iii) predict personalized therapies. The algorithm starts with a list of somatic mutations data either in ANNOVAR or VCF or BED format. The first layer prioritizes mutations and computes three iCAGES mutation scores: radial SVM scores for point mutations considering the scoring patterns of eleven predictors (SIFT, PolyPhen-2, GERPtt, PhyloP, CADD, VEST LRT, MutationTaster, Mutation Assessor, FATHMM, and SiPhy), Copy Number Variation (CNV) normalized peak scores for structural variation and FunSeq2 (Fu *et al.,* 2014) scores for point non-coding SNVs. The second layer prioritizes specific patient cancer driver genes where four features (the aforementioned three scores and a Phenotype Based Gene Analyzer (Phenolyzer, Yang *et al.,* 2015) score accounting for prior knowledge of candidate genes) are combined to generate a logistic regression score (iCAGES gene score or predicted probability) for each mutated gene. The final layer predicts personalized therapies and outputs iCAGES drug scores through a three-stage process where it screens BioSystems database for genes outputted earlier and computes the relatedness probabilities for its top four neighboring genes; then it queries

DGIdb database for potential drugs acting on cancer suppressor genes, oncogenes or other genes, and finally it computes the joint probability for each drug being the most effective (iCAGES drug score).

VI. Conclusion and Future Directions

Cancer is a complex disease and different computational methods may discover different drivers at different levels (e.g. mutation-level, gene-level and module-level). Due to lack of golden standard approach, their performance varies by dataset, which make them incomparable because of non-reproductive results. Therefore, the current tendency is to use complementary methods and then integrate the results. On the other hand, a vast array of genomic alterations has been linked to cancer. Many cancer driver prediction methods focus on using nsSNVs, but some methods can take account of in-frame indels (PROVEAN), CNAs (OncodriveCIS), CNVs (Multidendrix, Memo, Dendrix) and fusion genes (Oncofuse, ConSig). The identification of drivers in the coding regions may be starting to culminate; whereas, the study of noncoding regions has just started. With the rapid expansion of whole genome sequencing, more data will be available for future identification of genomic alterations in noncoding regions. To ease this process, the development of better computational algorithms is in urgent need in order to overcome some of weakness of the current methods.

Acknowledgements:

This work was supported in part by the NIH [CA172652] and the NCI Cancer Center Support Grant [P30 CA016672].

References

Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* **458**: 719-724.

Vogelstein B, Papadopoulos N, Velculescu VE *et al*. (2013) Cancer genome landscapes. *Science* **339**:1546-1558.

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

Mao Y, Chen H, Liang H *et al*. (2013) CanDrA: Cancer-Specific Driver Missense Mutation Annotation with Optimized Features. *PLoS ONE* **8**:e77945.

Dong C, Wei P, Jian X, Gibbs R *et al*. (2015a) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* **24**:2125-2137.

Shihab HA, Gough J, Cooper DN *et al*. (2013a) Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics* **29**:1504-1510.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**:1073-1081.

Adzhubei IA, Schmidt S, Peshkin L *et al*. (2010): A method and server for predicting damaging missense mutations. *Nature Methods* **7**: 248-249.

Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. *Genome Research* **19**:1553-1561.

Schwarz JM, Cooper DN, Schuelke M *et al.* (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods* **11**:361-362

Gonzalez-Perez A and Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American Journal of Human Genetics* **88**: 440-449.

Choi Y, Sims GE, Sean Murphy S *et al.* (2012) Predicting the functional effect of amino acid substitutions and indels. *PloS One* **7**: e46688

Kircher M, Witten DM, Jain P *et al.* (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**:310-315.

Kaminker JS, Zhang Y, Watanabe C *et al.* (2007a) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Research* **35** (Web Server issue):W595-598.

Davydov EV, Goode DL, Sirota M *et al.* (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**: e1001025.

Garber M, Guttman M, Clamp M et al. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**: i54- i62.

Cooper GM, Stone EA, Asimenos G *et al.* (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**: 901-913

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-1073.

UniProt, Consortium (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research* **39**: D214-D219.

Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a light-weight database of human nonsynonymous SNPs and their functional predictions. *Human Mutation* **32**: 894-899.

Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human Mutation* **34**: E2393- E2402.

Carter H, Chen S, Isik L *et al*. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Research* **69**: 6660-6667.

McLaren W, Pritchard B, Rios D *et al.* (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069-2070.
Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**: e164.
Capriotti E and Altman RB (2011). A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**: 310–317.
Gnad F, Baucom A, Mukhyala K *et al.* (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, 14(Suppl. 3), S7.
Martelotto LG, Ng CK, De Filippo MR *et al.* (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. *Genome Biology* **15**: 484.
Castellana S and Mazza T (2013) Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Brief Bioinformatics***14**: 448-459.
Dees ND, Zhang Q, Kandoth C *et al.* (2012) MuSiC: Identifying mutational signifiance in cancer genomes. *Genome. Res.* **22**: 1589-1598.
Lawrence MS, Stojanov P, Polak P *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214-218.
Marx V (2014) Cancer genomes: discerning drivers from passengers. *Nature Methods* **11**: 375–379
Reimand J, Wagih O, Bader GD (2013) The mutational landscape of phosphorylation signaling in cancer. *Scientific Reports* **3**: 2651; DOI:10.1038/srep02651.
Tian R, Basu MK, Capriotti E (2014) ContrastRank: a new method for ranking putative cancer driver genes and classification of tumor samples. *Bioinformatics* **30**: i572-i578.
Hua X, Xu H, Yang Y *et al.* (2013) DrGaP: a powerful tool for identifying driver genes and pathways in cancer sequencing studies. *American Journal of Human Genetic* **93**: 439-451.
Bushman, F. (2013) Cancer Gene List. http://www.bushmanlab.org/links/genelists
Forbes SA, Beare D, Gunasekaran P *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* **43**: D805–D811.
Tamborero D, Gonzalez-Perez A, Perez-Llamas C *et al.* (2013a) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3:** 2650; DOI:10.1038/srep02650.
Pon JR and Marra MA (2015) Driver and passenger mutations in cancer. Annual Review of Pathology **10**: 25-50.
Creixell P, Jüri R, Haider S *et al.* (2015) Pathway and network analysis of cancer genomes. *Nature Methods* **12**: 615-621.
Gonzalez-Perez A and Lopez-Bigas N (2012) Functional impact bias reveals cancer drivers *Nucleic Acids Research*. **40**: e169.
Tamborero D, Gonzalez-Perez A, Lopez-Bigas N (2013b) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**: 2238-2244.
Tamborero D, Lopez-Bigas N, Gonzalez-Perez A (2013c) Oncodrive-CIS: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PloS One* **8**: e55489.
Schroeder MP, Rubio-Perez C, Tamborero D *et al.* (2014) OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action *Bioinformatics* **30**: i549-i555.
Hodis E, Watson IR, Kryukov GV *et al.* (2012) A landscape of driver mutations in melanoma. Cell **150**:251-63.
Cheng WC, Chung IF, Chen CY *et al.* (2014) DriverDB: an exome sequencing database for cancer driver gene identifiation. *Nucleic Acids Research* **42:** D1048–D1054.

Zhao J, Zhang S, Wu LY *et al.* (2012) Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics* **28**: 2940-2947.
Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J *et al.* (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nature Methods* **10**: 1081-1082.
Leiserson MD, Blokh D, Sharan R *et al.* (2013) Simultaneous identification of multiple driver pathways in cancer. *PLoS Computational Biology* **9**: e1003054.
Liu Y, Tian F, Hu Z *et al.* (2015) Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers. *Science Reports* **5**:10204.
Raphael BJ, Dobson JR, Oesper L *et al.* (2014) Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Medicine* **6**:5.
Subramanian A, Tamayo P, Mootha VK *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States* **102**: 15545-15550.
Wendl MC, Wallis JW, Lin L *et al.* (2011) PathScan: a tool for discerning mutational significance in groups of putative cancer genes. *Bioinformatics* **27**:1595-1602.
Mitrea C, Taghavi Z, Bokanizad B *et al.* (2013) Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology* **4**: 278.
Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. Journal of *Computational Biology* **18**: 507-22.
Miller CA, Settle SH, Sulman EP *et al.* (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Medical Genomics* **4**:34.
Vandin F, Upfal E, Raphael BJ (2012) De novo discovery of mutated driver pathways in cancer. *Genome Research* **22**: 375-85.
Leiserson MD, Wu H-T, Vandin F *et al.* (2015) CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology* **16**:160.
Vaske CJ, Benz SC, Sanborn JZ *et al.* (2010) Inference of patient-specic pathway activities from multidimensional cancer genomics data using paradigm. *Bioinformatics* **26**: 12.
Tarca AL, Draghici S, Khatri P *et al.* (2009) A novel signaling pathway impact analysis. Bioinformatics **25**: 75-82.
Cerami E, Demir E, Schultz N *et al.* (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE* **5**: e8918.
Ciriello G, Cerami E, Sander C *et al.* (2012) Mutual exclusivity analysis identifies Oncogenic network modules. *Genome Research* **22**: 398-406.
Liu Y and Hu Z (2014) Identification of collaborative driver pathways in breast cancer. *BMC Genomics* **15**:605.
Babur Ö, Gönen M, Aksoy BA *et al.* (2015) Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. *Genome Biology* **16**: 45. doi:10.1186/s13059-015-0612-6.
Szczurek E, Beerenwinkel N (2014) Modeling Mutual Exclusivity of Cancer Mutations. *PLoS Computational Biology* **10**: e1003503. doi:10.1371/journal.pcbi.1003503.
Linghu B, Snitkin ES, Hu Z *et al.* (2009) Genome-wide prioritization of disease genes and identification of disease associations from an integrated human functional linkage network. *Genome Biology* **10**: R91.
Wang XS, Prensner JR, Chen G *et al.* (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nature Biotechnology* **27:** 1005 - 1011
Shugay M, Ortiz de Mendíbil I, Vizmanos JL *et al.* (2013) Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. *Bioinformatics* **29**:2539-2546.

Weinhold N, Jacobsen A, Schultz N *et al.* (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics* **46**:1160-1165.
Fredriksson NJ, Ny L, Nilsson JA *et al.* (2014) Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics* **46**:1258-63.
Li H, Chen H, Liu F *et al.* (2015) Functional annotation of HOT regions in the human genome: implications for human disease and cancer. *Scientific Reports* **5**:11633.
Smith KS, Yadav VK, Pedersen BS *et al.* (2015) Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic acids research* **43**: 5307-5317.
Fu Y, Liu Z, Lou S *et al.* (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology* **15**: 480.
Ritchie GR, Dunham I, Zeggini E *et al.* (2014) Functional annotation of noncoding sequence variants. *Nature Methods* **11**: 294-296.
Kircher M, Witten DM, Jain P *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46**:310–315.
Rubio-Perez C, Tamborero D, Schroeder MP *et al.* (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**: 382-396.
Dong C, Yang H, He Z *et al.* (2015b) iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing cancer driver genes in personal genomes. bioRxiv, 015008.
Yang H, Robinson PN, Wang K (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature Methods* **12**: 841–843.

Further Reading list

Zhang J, Liu J, Sun J *et al.* (2014) Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Briefings in Bioinformatics* **15**: 244-255.
Hou JP and Ma J (2013) Identifying Driver Mutations in Cancer in B. Shen (ed.), Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases, Translational Bioinformatics
Shihab HA, Rogers MF, Gough J *et al.* (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*; **31**: 1536-1543.
Kaminker JS, ZhangY, Waugh A *et al. et al.* (2007b) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Research* **67**: 465-473
Cooper GM and Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* **12**:628-640.
Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine* 4:89. doi: 10.1186/gm390.
Grimm DG, Azencott CA, Aicheler F *et al.* (2015) The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutations* **36**: 513-523.

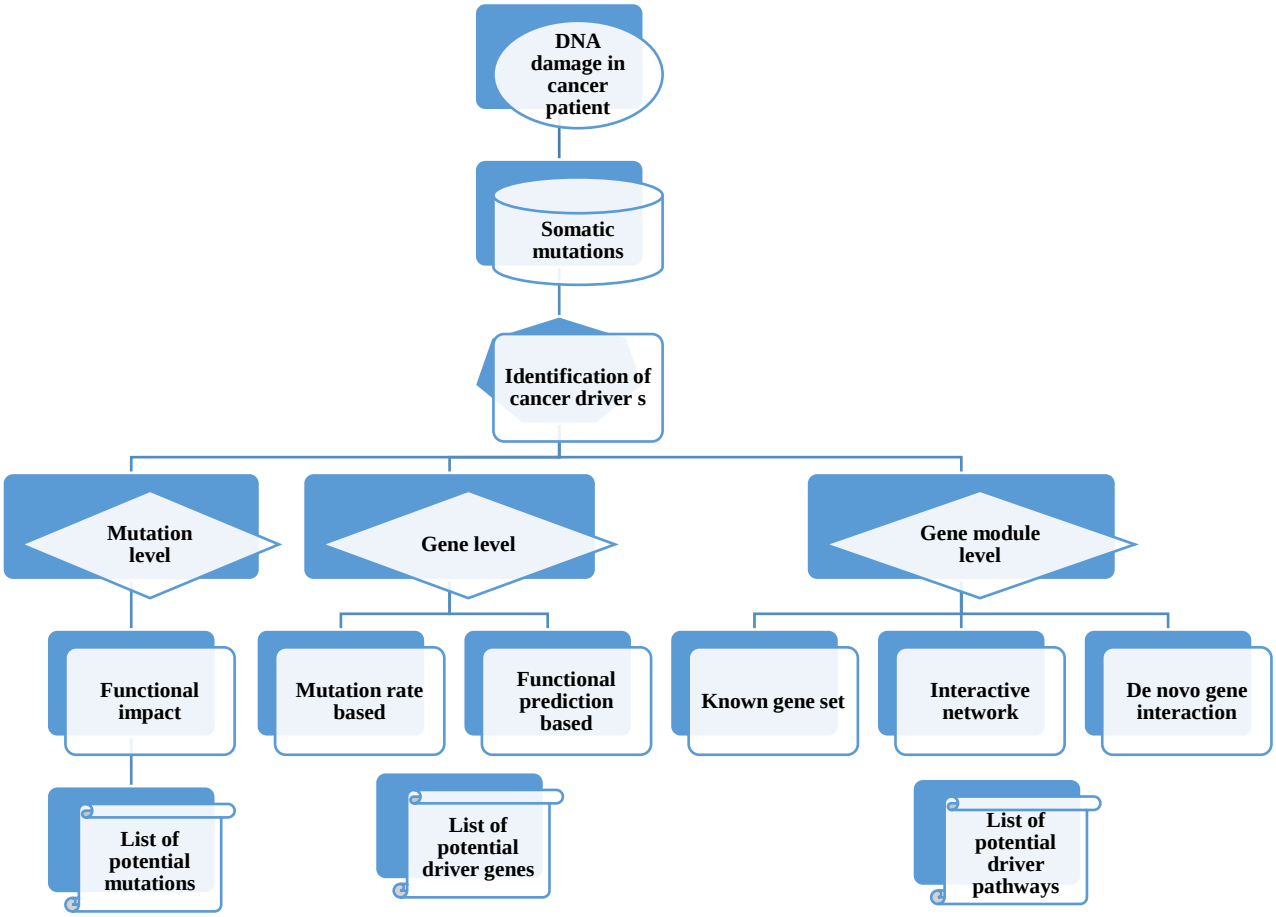

Figure 1: An overview of three broader approaches for predicting potential candidate driver (mutations, genes and pathways) from somatic mutations.

Table 1: Computational Predictive Algorithms of Genetic Drivers in Cancer

DRIVER MUTATIONS ALGORITHMS		DESCRIPTION	URL
Non Cancer specific	*Condel*	Uses a weighted approach to integrate two functional impact scores from MutationAssessor and FatHMM for nsSNVs	http://bg.upf.edu/condel
	MetaSVM	Uses a support vector machine approach to integrate ten functional impact scores for nsSNVs	
	MetaLR	Uses a logistic regression approach to integrate ten functional impact scores for nsSNVs logistic regression	
Cancer-specific	*CanPredict*	Uses a random forest classifier to integrate the metrics from SIFT, Pfam-based LogR.E-value and Gene Ontology Similarity Score (GOSS)	http://research-public.gene.com/Research/genentech/canpredict/index.html
	CHASM	Compute scores using a random forest classifier to integrate 49 predictive features and predict the functional significance of nsSNVs	http://wiki.chasmsoftware.org/
	MutationAssessor	Conservation-based method using position specific scoring matrices where a combinatorial entropy approach clusters conserved residues from specific residues in multiple sequence alignments from proteins families and subfamilies	http://mutationassessor.org
	CanDrA	Uses a weighted support vector machine to integrate the metrics from 95 structural and evolutionary features computed by ten functional prediction algorithms	http://bioinformatics.mdanderson.org/main/CanDrA
	FATHMM	Conservation-based method using position specific scoring matrices where a Hidden Markov Model approach combines sequence-based conservation features and pathogenicity weights to predict the functional impact of nsSNVs and non-coding variants	http://fathmm.biocompute.org.uk/
DRIVER GENES ALGORITHMS			
Mutation rate based approaches	*MuSiC*	Computes the SMG test by incorporating gene size and BMR to discover SMG	http://gmt.genome.wustl.edu/packages/genome-music/
	MutSigCV	Computes GSBMR integrating DNA replication time, chromatin state and gene-expression level to discover SMG	http://www.broadinstitute.org/cancer/cga/mutsig
	ActiveDriver	Use a gene-centric logistic regression approach to discover SMGs disrupting protein phosphorylation sites	http://individual.utoronto.ca/reimand/ActiveDriver/
	ContrastRank	Computes GSBMR using the maximum PDR for the gene in TCGA normal and 1000 Genomes samples. The PDR is the fraction of samples in which a given gene carries at least one PDV	http://snps.biofold.org/contrastrank/
Function prediction based approaches	*OncodriveFM*	Computes integrated functional impact scores from three other functional impact scores (SIFT, PolyPhen2 and MutationAssessor).	http://bg.upf.edu/oncodrive-fm
	OncodriveCLUST	Uses the binomial cumulated distribution from nsSNVs, stop-gain SNVs, and splice site mutations and coding silent mutations to compute gene clustering score	http://bg.upf.edu/oncodrive-clust
	Oncodrive-CIS	Computes an gene expression score integrating two standard scores using the Stouffer's method	http://bg.upf.edu/oncodrive-cis
	InVeX	Uses a permutation based approach to compute GSMR using synonymous mutations and/or mutations in introns and UTR sequences	http://www.broadinstitute.org/cancer/cga/invex
GENES MODULES ALGORITHMS			
Known Gene Set Approaches	*GSEA*	Uses the Kolmogorov-Smirnoff method and a permutation test to discover significantly enriched genes in known gene sets	http://www.broadinstitute.org/gsea/index.jsp
	PathScan	Computes enrichment significance test accounting for both differences in gene length and in mutation probabilities	http://tvap.genome.wustl.edu/tools/pathscan/
De Novo Gene Interaction Approaches	*Dendrix*	Uses the Markov Chain Monte Carlo approach to score mutually exclusivity and find sets of genes with coverage	http://compbio.cs.brown.edu/projects/dendrix/
	Multi-Dendrix	Uses the Integer Linear Programming to score mutually exclusivity and find sets of genes with high coverage	https://github.com/raphael-group/multi-dendrix
Interaction Networks Approaches	*PARADIGM*	Uses a factor graphs model which assigns weights to each interaction and then aggregates multiple scores to search for consistent pathways	http://sbenz.github.com/Paradigm
	MEMo	Uses a permutation test to score coverage and finds a combination of genes with mutual exclusivity in interaction network	http://cbio.mskcc.org/memo
	Mutex	Uses a permutation approach to model the null distribution and selects genes that significantly exhibit mutual exclusivity pattern	https://code.google.com/p/mutex/
	MUDPAC	Uses a mutational pathway enrichment analysis combining gene Mutation Factor (MF) and gene Interaction Factor (IF); followed by a greedy algorithm using the Maximal Coverage Rate (MCR) to search for collaborative driver pathways	http://www.visantnet.org/misi/MUDPAC.zip

SMGs: Significantly mutated genes; BMR: background mutation rate; GSBMR: gene specific background mutation rate ; GSMR: gene specific mutation rate; SMG: Significantly mutated Genes; PDR: putative defective rate; PDV: putative deleterious variant