Cancer Letters 369 (2015) 307–315

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Mini-review

Mining the coding and non-coding genome for cancer drivers

Jia Li [a], Damien Drubay [b,c], Stefan Michiels [b,c], Daniel Gautheret [a,*]

[a] Institute for Integrative Biology of the Cell (I2BC), CNRS, CEA, Université Paris-Sud, Université Paris-Saclay, 91198 Gif sur Yvette, France
[b] Service de Biostatistique et d'Epidemiologie, Gustave Roussy, Villejuif, France
[c] INSERM U1018, CESP, Université Paris-Sud, Université Paris-Saclay, Villejuif, France

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form 24 September 2015
Accepted 24 September 2015

Keywords:
Cancer drivers
Non-coding drivers
Somatic mutation scoring
Bioinformatics

ABSTRACT

Progress in next-generation sequencing provides unprecedented opportunities to fully characterize the spectrum of somatic mutations of cancer genomes. Given the large number of somatic mutations identified by such technologies, the prioritization of cancer-driving events is a consistent bottleneck. Most bioinformatics tools concentrate on driver mutations in the coding fraction of the genome, those causing changes in protein products. As more non-coding pathogenic variants are identified and characterized, the development of computational approaches to effectively prioritize cancer-driving variants within the non-coding fraction of human genome is becoming critical. After a short summary of methods for coding variant prioritization, we here review the highly diverse non-coding elements that may act as cancer drivers and describe recent methods that attempt to evaluate the deleteriousness of sequence variation in these elements. With such tools, the prioritization and identification of cancer-implicated regulatory elements and non-coding RNAs is becoming a reality.

Introduction

Cancer is caused by the accumulation of genetic alterations and consequent disruption of cell functions. Over the past decade, the introduction of fast and relatively inexpensive sequencing methods has provided unprecedented opportunity to characterize cancer genomic landscapes. A variety of bioinformatics tools are now available to discover genetic variations from high throughput sequencing of tumor DNA, such as GATK [1], CRISP [2], LoFreq [3], VarScan 2 [4], and SNVer [5], which have been recently evaluated [6,7]. Depending on cancer type, tumors harbor hundreds to tens of thousands of somatic mutations, most of which are located in the non-coding portion of the genome [8]. A critical challenge in this context is to distinguish "driver" mutations and cancer genes that actively contribute to tumor growth or metastasis from "passenger" mutations that are mere results of the cancerous process. A number of reviews provide guidelines for the discovery of cancer-causing variants [9,10]. The most common strategy is first to prioritize non-synonymous variants in protein-coding regions and then seek recurrently mutated genes in a cohort of cancer patients [11–15]. Diverse computational methods have been explored to prioritize non-synonymous variants with respect to their disease-causing potential. Most are based on the assumption that coding mutations impacting functionally important residues, as inferred from evolutionary conservation and protein domain analysis, are more likely damaging [16]. Other software, used in conjunction with these scoring systems, perform recurrence search in patient cohorts. Currently, 547 cancer genes are described the COSMIC catalogue of somatic mutations in cancer (version 71) [17].

The immense majority of the human genome (98%) is non-coding, and consequently most somatic mutations/alterations observed in tumors occur in this non-coding fraction. Because non-coding mutations are more difficult to interpret, these regions have been mostly discounted from the wider search for driver mutations. However, mutations in non-coding regions can have a profound impact on cell fate. Indeed, functional regions in the non-coding genome include mRNA splice sites, UTR regulation elements, promoters, transcription factor binding sites, enhancers and a wide variety of non-coding RNA (ncRNA) genes. Among ncRNA genes, one particular class is now receiving focused attention due to its vast extent: long non-coding RNA (lncRNA). According to the latest estimate [18], over 58,000 lncRNA genes are expressed in the human genome, which makes this class the biggest contributor to the "black matter" transcriptome.

There is ample evidence for disease-related mutations in the non-coding genome. A large fraction of disease or trait-relevant single nucleotide polymorphisms (SNPs) detected by Genome-wide Association Studies (GWAS) [19] is located in the non-coding genome, preferentially within enhancers, exons and mRNA promoters [20]. Inherited disease-causing variants are strongly enriched in non-coding regions under strong purifying selection, which comprise binding sites of transcription factors (TFs) and critical motifs from TF Families [21]. Further studies have shown that altered ncRNA

* Corresponding author. Tel.: +33169154632; fax: +33169157296.
E-mail address: daniel.gautheret@u-psud.fr (D. Gautheret).

http://dx.doi.org/10.1016/j.canlet.2015.09.015
0304-3835/

functions initiated by genetic or regulatory changes play an important role in tumorigenesis [22–27].

In the absence of a clear and uniform functional code for these highly diverse non-coding elements, their variations are much more difficult to interpret than those of amino acid-coding regions. In this review we describe the methods and data available to interpret and prioritize non-coding genome mutations. As many basic principles in this field were laid for protein-coding sequence analysis, we start by reviewing the methods developed for scoring protein-coding variants. We then describe the specific non-coding elements that may be the subject cancer-driving mutations and we address the specific methods that were set up to characterize these variations.

Prioritizing coding variants

Prioritization of non-synonymous mutations for cancer is a mature field built upon decades of experience in protein sequence and cancer pathway analysis. Table 1 provides a listing of the most commonly used tools. We distinguish below three classes of scoring systems, using either probabilistic, machine learning or hybrid approaches.

Probabilistic models

The pioneering SIFT (Sorting Intolerant From Tolerant) uses sequence homology to predict whether an amino acid substitution will affect protein function and hence, potentially alter phenotype [28]. SIFT identifies conserved protein residues based on multiple sequence alignments of homologous proteins and calculates the likelihood that an amino acid at a position is tolerated, conditional on the most frequent amino acid being tolerated. Mutations in higher conserved coding regions intend to be predicted as more likely deleterious than those in lower conserved protein regions.

The mCluster method [32] aggregates mutation data by mapping known disease-related mutations to positions along conserved domains, and then mapping novel variants to those same conserved domains. The program identifies conserved mutation-enriched clusters, which are hotspots for cancer driving functional alterations, across multiple proteins. The mCluster score is the likelihood of a cluster of certain size occurring, given the number of positions in the domain and the mutation frequency.

MutationAssessor [30] implements a more elaborate conservation-based approach. It computes residue distribution entropy in multiple sequence alignments and estimates mutation impact by measuring the entropy difference caused by the mutation (conservation score). Moreover, the algorithm classifies protein alignment into distinct subfamilies with a clustering algorithm and quantifies the entropy difference initiated by a mutation in protein subfamilies (specificity score). The final "functional impact score" combines these two independent scores.

Machine learning models

PolyPhen2 [29] integrates eight sequence and three structure-based attributes for the description of an amino acid substitution, and predicts the damaging effect of a coding mutation. Most PolyPhen2 features compare a property of the wild-type allele (ancestral, normal) and the corresponding property of the mutant allele (derived, disease-causing) and characterizes how likely the two human alleles are to occupy the site given the pattern of amino-acid replacements in a multiple-sequence alignment. The probability of a deleterious allele replacement is predicted using a Naïve Bayes classifier trained on HumDiv and HumVar [41], two databases of damaging alleles.

CHASM uses a random forest classifier to discriminate driver missense mutations from synthetically generated passenger mutations [31]. It includes 49 predictive features ranging from exon conservation to UniProt annotation and frequency of the missense change type in the COSMIC database of cancer mutations [17]. The program computes a classification score for each missense mutation. A mutation is determined to be driver or passenger by comparing its score to a null distribution made of scores from a filtered set of synthetic passengers that were held out from the Random Forest training.

SNAP (Screening for Non-acceptable Polymorphisms) is a neural network-based tool that predicts the effect of a missense variant [33]. It uses PMD (the Protein Mutant Database) [45] and incorporates evolutionary constraints, transition frequencies for mutations, biophysical characteristics of the substitution, secondary structural information, relative solvent accessibility, and SwissProt annotations information to build a neural network model, which is trained on known mutations from PMD.

MutPred [35] is another Random Forest classifier trained on five databases of human amino acid substitutions, CANCER [46], KINASE [47], The Human Gene Mutation Database (HGMD) [48], SwissProt [49] and a broad array of attributes describing structure features (such as secondary structure, solvent accessibility), a variety of functional sites (such as DNA-binding or phosphorylation sites), evolutionary conservation and transition frequencies. The MutPred model then associates a given non-synonymous mutation to a probability of gain or loss of structural and functional features.

Hybrid models

The current trend for increasing the accuracy of impact measure is to integrate different methods. For example, CanPredict [34] uses a random forest classifier to predict whether a change is likely to be cancer-associated, based on analyses of three scores: the SIFT score determining functional impact of change, the Pfam-based LogR.E-value metric [50] and the Gene Ontology Similarity Score (GOSS), which measures how similar a given mutated gene is to known cancer-causing genes [51].

Condel [36] combines the output from PolyPhen2, SIFT, Mutation Assessor, Pfam-based LogR.E-values and MAPP [43], which predicts deleterious mutations based on their disruption of physicochemical protein characteristics. Another hybrid tool, CoVEC (Consensus Variant Effect Classification) [37] integrates prediction results from SIFT, PolyPhen2, Mutation Assessor and SNPs&GO [42], a scoring system based on functional protein features such as sequence conservation and GO-terms. Finally, Combined Annotation scoRing toOL (CAROL) combines the scores of PolyPhen-2 and SIFT to predict the effect of non-synonymous coding variants [38]. Expectedly, the authors of Condel, CoVEC and CAROL demonstrate that these tools outperform most individual methods in classifying variants as damaging or neutral, highlighting the benefits of combined approaches [36–38].

Comparing coding mutation scoring tools

The authors of CoVEC [37] assessed the classification performance of their tool and nine other prediction softwares: SIFT, PolyPhen2, SNPs&GO, PhD-SNP, PANTHER, Mutation Assessor, MutPred, Condel and CAROL. Based on the programs' ability to properly classify HGMD inherited disease-related variants [48] and neutral SNPs, MutPred had the best performance in terms of true positive rate, followed by PolyPhen2. SNPs&GO showed most applicability in cases requiring minimal false positive rates. Most of the individual tools had similar overall (ROC curve-based) performances, however, combined tools such as CoVEC were shown to outperform the-individual tools. In an independent benchmark, Thusberg et al [52] tested nine scoring tools for their ability to distinguish 40,000 pathogenic variants of the PhenCode database [53] from neutral variants. Tested tools included MutPred, Panther,

Table 1
Summary of computational methods for scoring missense mutations.

	Based on	Machine learning	Cancer-specific	Other tools used	Web server, references
SIFT	Conservation	Alignment scores	No		http://sift.jcvi.org/ [28]
Polyphen 2	Conservation Structure Training set	Bayesian classification	No		http://genetics.bwh.harvard.edu/pph2/ [29]
Mutation assessor	Conservation		No		http://mutationassessor.org/ [30]
CHASM	Conservation Structure Annotation Training set	Random forest	Yes		http://wiki.chasmsoftware.org/index.php/MainPage [31]
mCluster	Training set		Yes		http://www.mcluster.org [32]
SNAP	Conservation Structure Annotation Training set	Neural network	No	Gene ontology	http://rostlab.org/services/snap/ [33]
Canpredict	Conservation Annotation	Random forest	Yes	SIFT LogR.E GOSS	http://research-public.gene.com/Research/genentech/canpredict/ [34]
MutPred	Conservation Structure Annotation Training set	Random forest	No	SIFT	http://mutpred.mutdb.org/ [35]
Condel	Hybrid scoring system (weighted score)	NA	No	PolyPhen2, SIFT, Mutation Assessor, Pfam-based LogR.E-values and MAPP	http://bg.upf.edu/fannsdb/ [36]
CoVEC	Hybrid scoring system	SVM	No	SIFT, PolyPhen2, SNPs&GO, Mutation Assessor	http://www.dcs.kcl.ac.uk/pg/frousiok/variants/index.html [37]
CAROL	Hybrid scoring system	No	No	SIFT, PolyPhen2	http://www.sanger.ac.uk/resources/software/carol/ [38]
nsSNPAnalyzer	Structural and evolutionary information	Random forest	No		http://snpanalyzer.utmem.edu/ [39]
PANTHER	Conservation	Alignment scores	No		http://www.pantherdb.org/tools/csnpScoreForm.jsp [40]
PhD-SNP	Conservation Training set	Support vector machine	No		http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi [41]
SNPs&GO	Conservation Swissprot features	Support vector machine	No		http://snps-and-go.biocomp.unibo.it/snps-and-go/ [42]
MAPP	Physicochemical constraints	NA	No		http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005/ [43]
IntOGen-mutations	Hybrid scoring system	NA	Yes	PolyPhen2, SIFT, Mutation Assessor	http://www.intogen.org/web/mutations/v04/search [44]

PhD-SNP, PolyPhen, PolyPhen2, SIFT, SNAP, SNPs&GO and nsSNPAnalyzer [39]. Programs SNPs&GO and MutPred had best overall prediction accuracy.

Integrating recurrence for driver prediction

Further to prioritizing individual mutations as shown above, a variety of approaches predict driver genes by combining mutation scores and recurrence patterns. The assumption underlying these methods is that genes critical to the development of a specific cancer type should be recurrently mutated in a cohort of cancer samples. Several programs are available to identify such genes [11–15].

IntOGen-mutations is a web server aiming to identify cancer drivers across tumor types [44]. The system first determines the consequences of mutations using the Ensembl variant effect predictor tool which offers a comprehensive database of variations, their effects and context [54] and uses three of the above tools (SIFT, PolyPhen2 and MutationAssessor) to compute the functional impact score of a somatic mutation. These functional scores are then transformed into a uniform score which measures the damaging impact of somatic mutations with transFIC [36]. This pipeline also computes each mutation's frequency of occurrence within and across cancer projects and groups mutations occurring in the same gene (or pathway). Subsequently, OncodriveFM [55] which detects genes accumulating mutations with high functional impact (FM bias) and OncodriveCLUST tools [56] which determine genes whose mutations cluster in particular regions of the protein sequence in comparison with synonymous mutations (CLUST bias) are used to identify positively selected genes, *i.e.* genes whose mutations are selected during tumor development and are therefore likely drivers. Finally, the pipeline computes the frequency of mutation of each gene (and pathway) within a cancer class.

The MutSigCV method [8] assesses the background mutation rate for each gene–patient–category combination based on the observed silent mutations in the gene and non-coding mutations in the surrounding regions. It pools data from other genes with similar properties (for example replication time, expression level) to increase accuracy. Significance levels (P values) are determined by examining whether observed mutations in a gene significantly exceed the expected counts based on the background model.

MuSiC relies on the calculation of a background mutation rate (BMR) [57]. The algorithm counts the number of bases with sufficient aligned read-depth based upon user-defined coverage. Counts are determined for A, T, C and G as CpG dimers, and non-CpG C and G. Discovered mutations are categorized according to mutational mechanism, with separate categories for AT transitions, AT transversions, CpG transitions, CpG transversions, CG (non-CpG) transitions and transversions, and a seventh 'indel' category. The BMR of each mutational mechanism category is calculated by dividing the number of mutations found in that category by the total number of bases available in which such a call could have been made. Significantly mutated genes are generated by comparisons of mutation rates to BMR, using statistical tests.

InVEx is a random permutation-centered algorithm [58] that relies on the assumption that a gene under positive selection for nonsilent mutations during cancer formation displays a higher rate of high-scoring non-synonymous mutations than silent and intronic mutations. A random permutation test is performed across each gene and a "mutation burden" score is calculated for each randomized instance, providing a null model of score distribution. The actual mutation burden observed for a gene across all samples is then compared to this distribution and a P-value is computed, assessing whether the observed coding mutations and genes undergo positive selection.

Although genes that are mutated with high recurrence are easily recognized, some cancer drivers are mutated in a small fraction (*e.g.* <1%) of tumors [59]. Thus, methods that can classify mutations as either drivers or passengers on the basis of data that is independent of mutation frequency clearly become important. There are many ways of combining mutation deleteriousness, recurrence and knowledge of mutational background. Computational options in this area are far from fully explored and we may thus expect improved driver predictors in the future. Furthermore, the application of these methods to the non-coding genome is a fascinating perspective, as so little is known about driver elements in these regions. This challenge may soon become accessible thanks to development of scoring systems for non-coding mutations, as explained in the next sections.

Non-coding elements and cancer

The list of non-coding elements involved in gene expression regulation has been steadily increasing over the years. Promoters, enhancers, splicing regulators and the expanding family of regulatory ncRNA (mainly miRNAs and lncRNAs) are central elements of the cell regulatory network. Their function in the control of gene expression is similar to that of many protein-coding cancer drivers, half of which are involved in transcriptional and posttranscriptional regulation. Therefore, it comes as no surprise that mutations within these non-coding elements are responsible for the initiation and progression of cancer, among other diseases [20,21,60–62].

The first non-coding cancer hotspots to be suspected were promoters and TF binding sites. Indeed, among 4,492 phenotype-associated SNPs from the GWAS Central Database [19], 12% are located in binding regions of transcription factors, which is significant as these loosely defined regions represent 8.1% of the genome [63]. Genetic variations at TF binding sites, including single-nucleotide polymorphisms and larger structural variants, are frequently associated with binding affinity [64–66], gene expression [67,68] and cancer susceptibility, progression and outcome [69–71]. A well-known such locus is the TERT promoter, whose mutations were established as drivers in melanomas and gliomas [60–62].

Another important class of regulatory element is that of splicing regulators. Misregulation of RNA splicing initiated by genetic variants is a cause of human disease, including cancer. Alteration of 5'and 3' splicing sites and adjacent bases accounts for 10% of human inherited disease mutations [72,73] and the number of tumor-relevant splicing variants detected by GWAS in cancers reaches 15,000 [74–76]. For example, a germline mutation in the splicing site of hSNF5 is causative of exon 7 skipping and subsequent frameshift, which, as a result, renders infants susceptible to develop malignant brain tumors [77]. Likewise, a mutation at the acceptor site of the APC gene intron 3–exon 4 junction causes the loss of exon 4, which accordingly terminates seven codons downstream of junction 4, a phenomenon closely associated to childhood hepatoblastoma [78].

Variation in non-coding RNA (ncRNA) sequence and expression is another potential component of cancer progression. The first important offenders in this class were miRNAs. Single nucleotide variations in miRNA sequences or in their mRNA target sites lead to alteration of binding specificity, thus affecting expression and/or translation of target mRNAs [79–83]. For instance, SNPs in mRNAs of the CEP cell division family alter mRNA/miRNA interactions, greatly affecting mRNA expression, disrupting the cell cycle and contributing to initiate cancer [81]. Overall, more than 236 miRNAs have been associated to 79 human cancers either as potential oncogenes or tumor suppressors [84].

Long non-coding RNA is the most recent class of regulatory ncRNA to be associated to cancer. According to a recent study [18], over 68% (58,648) of expressed genes in human tumors are lncRNAs, 7942 of them lineage- or cancer-specific. Through gene regulation or other mechanisms, lncRNAs may act as proto-oncogenes, tumor

Table 2
Summary of computational approaches for scoring non-coding mutations.

	Based on	Machine learning	Cancer-specific	Web server, references
RegulomeDB	Overlap with functional elements	Empirical scoring systems	No	http://www.regulomedb.org/ [104]
Funseq	Negative selection in general population, recurrent cancer mutations	Empirical scoring systems	Yes	http://funseq.gersteinlab.org/ [21]
Funseq2	Negative selection in general population, recurrence in cancer mutations	Empirical scoring systems	Yes	http://funseq2.gersteinlab.org/ [105]
GWAVA	HGMD regulatory mutations, integrated genome annotation	Random forest	No	https://www.sanger.ac.uk/sanger/StatGen_Gwava [106]
CADD	Deleteriousness, diverse genome annotation	Support vector machine	No	http://cadd.gs.washington.edu/ [107]
SPANR	RNA splicing model	Bayesian machine learning	No	http://tools.genes.toronto.edu/ [108]
FATHMM-MKL	HGMD mutations, ten feature annotations (6 from ENCODE)	Support vector machine	No	http://fathmm.biocompute.org.uk [109]

suppressor genes or drivers of metastatic transformation. For instance, the HOTAIR lncRNA is highly expressed in primary breast tumors and metastases, as well as in gastric cancer, and its repression inhibits xenograft tumor growth and metastasis in mouse models [85,86]. MALAT1 is another lncRNA whose expression is correlated with metastasis and survival in lung cancer [87]. Knockout of MALAT1 greatly impairs the migration and formation of tumor nodules of MALAT1-deficient A549 cells in a mouse xenograft [88]. Jin et al. [89] observed that among a set of 33 SNPs independently associated with elevated prostate cancer (PCa) risk, eight were located in lncRNAs. Moreover, lncRNA loci showed a five-fold enrichment of PCa risk-related SNPs in comparison with the entire genome. SNPs in the lncRNA PRNCR1 were proposed to be related to colorectal cancer (CRC) risk [90].

In spite of these recent advances, the list of cancer-driving elements in the non-coding genome remains extremely short with respect to the size of the regions involved. A major avenue in identifying new potentially relevant loci involves exploring chromatin states. Indeed, regions where chromatin is open or active in a given cell type are the most likely to contain key regulatory elements. For instance, DNase I hypersensitive sites (DHSs), *i.e.* DNA regions sensitive to the DNase I enzyme, harbor many regulatory elements such as enhancers, promoters and silencers [91,92]. Moreover, DHSs are associated with elevated levels of nearby gene expression, at least in certain cells [92]. Other important functional hallmarks are provided by histone modifications such as acylation and methylation, which control chromatin states and are thus important regulators of gene expression [93]. Specific histone marks suggest different types of regulatory elements: H3K4me3 generally marks promoters and transcription start sites. Putative enhancers tend to be marked with H3K4me1 alone or in combination with H3K27ac or H3K27me3 [94,95]. Conversely, major repressive marks, such as H3K9me3 and H3K27me3, are associated with constitutive heterochromatin and repetitive elements, repressive domains and silent developmental genes [96] and are therefore less likely to harbor cancer drivers.

Prioritizing non-coding variants

Although the number of cancer-associated non-coding mutations is increasing, finding cancer-driving mutations in the non-coding genome remains a huge challenge. A major bottleneck lies in identifying functional domains while trying to explore the consequences of the variations. Functional interpretation of non-coding variations is now turning into a realistic goal through the completion of major high-throughput studies such as the Encyclopedia of DNA Elements (ENCODE) [96], the "29 Mammals" Project [97], the Health Roadmap Epigenomics project [98] and other large scale regulatory data collections [99–103]. Particularly, The ENCODE Project has provided researchers with genome-wide mapping of histone modification, Dnase I hypersensitive sites, FAIRE sites (formaldehyde-detected nucleosome-depleted elements), transcription factor binding sites, RNA-seq expression data and replication timing across multiple cell lines [96]. These extensive data form a major stepping-stone toward the functional annotation of non-coding variants. More and more studies are taking advantage of these annotations to explore and prioritize non-coding variants implicated in cancer and other diseases. Table 2 presents seven systems that are currently available for scoring non-coding variants. We distinguish below two families of such methods, based either on empirical scoring systems or on machine learning.

Empirical scoring systems

The RegulomeDB database and software [104] assigns functions to non-coding variants based on the principle that a variant impacting a regulatory element likely results in functional consequence. Non-coding variants are classified into different functional categories according to their overlap with functional elements such as transcription factor binding, histone modifications, DNase I hypersensitive sites, FAIRE sites and eQTLs (expression Quantitative Trait Loci, that is loci likely to affect expression of target genes). Application of this tool to the annotation of non-coding variants from 69 full sequenced genomes [110] identified thousands of potential functional variants.

The FunSeq tool [21] predicts non-coding drivers by scoring the deleterious potential of variants, based on two assumptions. First, somatic variants in non-coding elements containing a high fraction of rare variants (derived allele frequency < 0.5% in the general population) are considered as under negative selection and thus are most likely to be cancer drivers. Second, driver mutations should be recurrent in the same genomic element across multiple cancer samples. Application of this workflow to 90 cancer genomes yielded nearly a hundred non-coding drivers candidates. An improved algorithm, FunSeq2 [105] exploits large-scale genome data from 1000 Genomes and ENCODE into a scoring pipeline that combines functional features such as sequence conservation, transcription-factor binding sites, enhancer-gene linkages, network centrality and recurrence across samples. In this model, features are weighted by their probability of overlapping a natural polymorphism in the 1000 Genome database, which is a negative indicator of selection strength. Application of FunSeq2 to germline pathogenic regulatory variants successfully distinguished HGMD (Human Gene Mutation Database) and GWAS non-coding pathogenic variants from neutral ones. The method also effectively scored COSMIC recurrent variants higher than non-recurrent variants.

Machine-learning models

While the RegulomeDB and FunSeq systems prioritize functional genetic variations using empirical models, recent methods aim to integrate functionally predictive features automatically using machine learning [106,107,109]. One of these models, GWAVA [106]

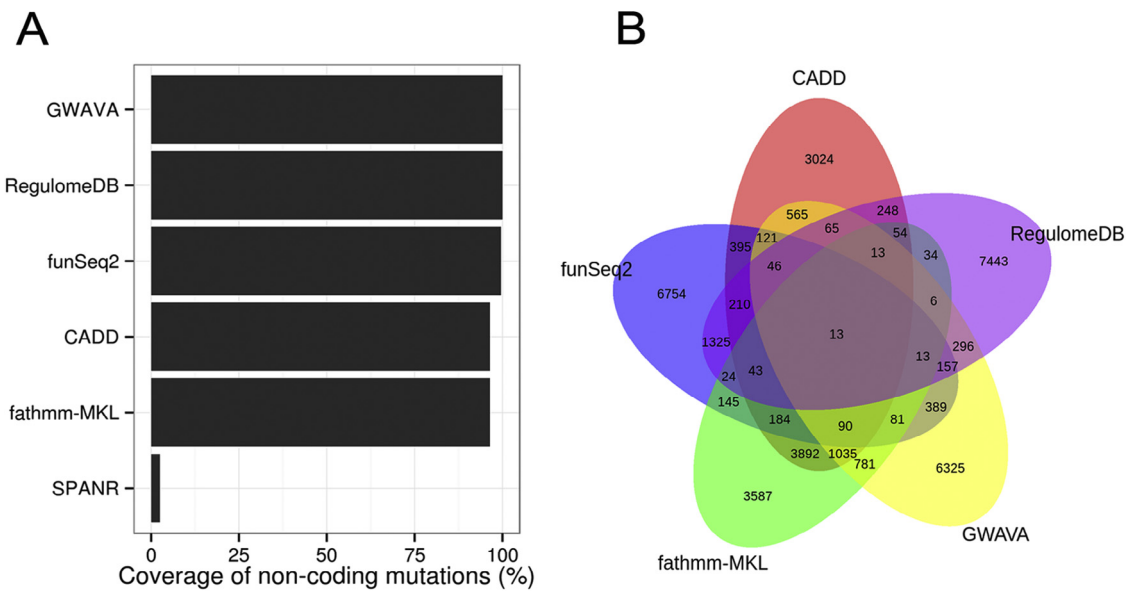

Fig. 1. Comparison of six non-coding mutation scoring tools. A. Fraction of positions covered by each tool in a set of 874,325 non-coding variants. B. Overlap of the 10,000 top-scoring variants, using the 5 scoring tools with the larger prediction coverage (CADD, Fathmm-MKL, FunSeq2, GWAVA and RegulomeDB), from the 841,402 variants common to their prediction coverage.

uses regulatory mutations annotated in the HGMD database as a training set for non-coding variants of medical importance. These variants are predicted using a random forest classifier based on a combination of regulatory features, genic context and genome-wide properties such as DNase I hypersensitivity sites, FAIRE sites, Transcription factor binding sites, Histone modifications, RNA polymerase binding sites, complex epigenetic states, CpG islands, sequence conservation, allele frequency of variants and gene annotation. The model was able to effectively discriminate a set of disease-relevant variations of the ClinVar [111] and GWAS Central databases from control variants. More importantly, recurrent cancer mutations from the COSMIC database were scored significantly higher than non-recurrent mutations, suggesting that this approach might be useful in prioritizing cancer driver mutations.

Another tool, FATHMM-MKL, implements multiple kernel learning to weight different ENCODE feature annotations based on their relevance. The program builds a Support Vector Machine classifier based on a positive training set of non-coding pathogenic variants annotated in HGMD and a negative set of common single-nucleotide variants with allele frequency above 1% within 1-Kb surrounding disease-causing variants. The model uses for prediction a kernel matrix of 10 annotation features, including transcription factor binding sites, evolutionary conservation, DNase I hypersensitive sites and histone modifications [109]. A possible limitation in GWAVA and FATHMM-MKL is the methods highly rely on a set of promoter proximal, pathogenic mutations that are well characterized and thus are subject to ascertainment bias.

Instead of building a classifier using limited curated pathogenic variants, the CADD system [107] contrasts the annotations of fixed derived alleles in humans with those of de novo simulated variants. Here fixed (or nearly fixed) alleles are used as models for deleterious variants. The CADD system is trained to recognize such variants using a support vector machine classifier based on a combination of 63 tracks of annotations, including conservation, regulatory information, transcript information, protein-level score produced by SIFT, Polyphen or Grantham [112]. CADD successfully differentiated 14.7 million high-frequency human-derived alleles (observed variants) from 14.7 million simulated variants (half simulated de novo mutations).

To conclude this section, we mention SPANR (splicing-based analysis of variants) [108], a program that combines a Bayesian machine learning algorithm and a regulatory model of gene splicing to detect and score disease-associated genetic variants. The RNA splicing model integrates regulatory elements and splicing levels generated from RNA-seq data of healthy human tissues. SPANR is capable of a precise classification of both intronic disease-related variants and deleterious disease mutations within exons, from common variants in the dbSNP database. Analyses using SPANR have generated a large body of splice-disruptive mutations involved in autism, familial colorectal cancer and spinal muscular atrophy, which are known for RNA-splicing deregulation.

Comparing non-coding variant scoring tools

To illustrate the divergence of predictions by different non-coding mutation scoring systems, we selected six tools from the current literature (CADD, FunSeq2, GWAVA, RegulomeDB, Fathmm-MKL and SPANR) and used them to score 874,325 non-coding variants (both substitutions and short indels) from the whole genome sequencing of 88 liver cancer samples [8]. First, we should note that all tools are not applicable to the entire set of somatic mutation (Fig. 1A). GWAVA, RegulomeDB, and funSeq2 were able to score over 99% of variants, while SPANR provided scores for only 2.48% of variants due to its specificity for splicing regulation. Due to this different scope, we excluded SPANR from further comparison. We scored the 841,402 somatic mutations covered by the other 5 tools and collected the 10,000 highest scoring variants from each tool. The Venn diagram in Fig. 1B shows the overlapping of predictions. Strikingly, only 13 variants are commonly predicted as high scoring by all five tools, illustrating the remarkable divergence of non-coding variant prioritization strategies. While a full benchmark of the different prediction algorithms is beyond the scope of this review, we may refer to two studies that assessed the performances of various non-coding variant prioritization tools in classifying sets of known deleterious HGMD variants. Each study compared a specific program developed by the authors to leading "state-of-the-art" algorithms. Fu et al. [105] showed that FunSeq2 has a better average prediction power compared to GWAVA and CADD, while Shihab et al. [109]

showed that FATHMM-MKL outperformed GWAVA and CADD in terms of accuracy. Due to the substantial number of recently developed methods, a full scale and independent comparative study would be valuable to provide consistent results and objectively identify the strengths and weakness of each tool.

Conclusion

The search for cancer drivers requires a reliable functional annotation of variants and adapted tools for analyzing the recurrence of deleterious variants across patients. The former requisite is particularly challenging in the non-coding genome. An active research community is developing tools for non-coding variant annotation and prioritization using a variety of methods ranging from empirical scoring scheme to machine-learning and elaborate hybrid frameworks. Due to the heterogeneity and complexities of these scoring tools, objective comparisons based on proper benchmarks using different sets of validated or probable disease-causing variants are strongly required. Among multiple sources of possible improvement, the success of hybrid methods for scoring coding variants, and the widely divergent predictions by the non-coding tools suggest that combining outputs from different tools will significantly increase scoring accuracy for non-coding variants. A further challenge is to jointly consider this "functional" score and the heterogeneity of cancer specific mutation constraints in different genome areas. These potential enhancements suggest we can expect important reliability gains in non-coding variant prioritization in the near future.

Acknowledgement

This work was funded in part by ITMO Cancer "Plan Cancer – Systems Biology" grant #bio2014-04.

Conflict of interest

The authors declare no conflict of interest.

References

[1] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491–498, doi:10.1038/ng.806.
[2] V. Bansal, A statistical method for the detection of variants from next-generation resequencing of DNA pools, Bioinformatics 26 (2010) 318–324, doi:10.1093/bioinformatics/btq214.
[3] A. Wilm, P.P.K. Aw, D. Bertrand, G.H.T. Yeo, S.H. Ong, C.H. Wong, et al., LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets, Nucleic Acids Res. 40 (2012) 11189–11201, doi:10.1093/nar/gks918.
[4] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, et al., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. (2012) 568–576, doi:10.1101/gr.129684.111.
[5] Z. Wei, W. Wang, P. Hu, G.J. Lyon, H. Hakonarson, SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data, Nucleic Acids Res. 39 (2011) 1–13, doi:10.1093/nar/gkr599.
[6] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, et al., A survey of tools for variant analysis of next-generation genome sequencing data, Brief. Bioinform. 15 (2014) 256–278, doi:10.1093/bib/bbs086.
[7] H.W. Huang, J.C. Mullikin, N.F. Hansen, Evaluation of variant detection software for pooled next-generation sequence data, BMC Bioinformatics 16 (2015) 235, doi:10.1186/s12859-015-0624-y.
[8] M.S. Lawrence, P. Stojanov, P. Polak, G.V. Kryukov, K. Cibulskis, A. Sivachenko, et al., Mutational heterogeneity in cancer and the search for new cancer-associated genes, Nature 499 (2013) 214–218, doi:10.1038/nature12213.
[9] Y. Moreau, L.-C. Tranchevent, Computational tools for prioritizing candidate genes: boosting disease gene discovery, Nat. Rev. Genet. 13 (2012) 523–536, doi:10.1038/nrg3253.
[10] D.G. MacArthur, T.A. Manolio, D.P. Dimmock, H.L. Rehm, J. Shendure, G.R. Abecasis, et al., Guidelines for investigating causality of sequence variants in human disease, Nature 508 (2014) 469–476, doi:10.1038/nature13127.
[11] L. Ding, G. Getz, D.A. Wheeler, E.R. Mardis, M.D. McLellan, K. Cibulskis, et al., Somatic mutations affect key pathways in lung adenocarcinoma, Nature 455 (2008) 1069–1075, doi:10.1038/nature07423.
[12] M.A. Chapman, M.S. Lawrence, J.J. Keats, K. Cibulskis, C. Sougnez, A.C. Schinzel, et al., Initial genome sequencing and analysis of multiple myeloma, Nature 471 (2011) 467–472, doi:10.1038/nature09837.
[13] Y. Gui, G. Guo, Y. Huang, X. Hu, A. Tang, S. Gao, et al., Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder, Nat. Genet. 43 (2011) 875–878, doi:10.1038/ng.907.
[14] K. Wang, J. Kan, S.T. Yuen, S.T. Shi, K.M. Chu, S. Law, et al., Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer, Nat. Genet. 43 (2011) 1219–1223, doi:10.1038/ng.982.
[15] X. Wei, V. Walia, J.C. Lin, J.K. Teer, T.D. Prickett, J. Gartner, et al., Exome sequencing identifies GRIN2A as frequently mutated in melanoma, Nat. Genet. 43 (2011) 442–446, doi:10.1038/ng.810.
[16] D. Vitkup, C. Sander, G.M. Church, The amino-acid mutational spectrum of human genetic disease, Genome Biol. 4 (2003) R72, doi:10.1186/gb-2003-4-11-r72.
[17] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, et al., COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer, Nucleic Acids Res. 39 (2011) 945–950, doi:10.1093/nar/gkq929.
[18] M.K. Iyer, Y.S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, et al., The landscape of long noncoding RNAs in the human transcriptome, Nat. Genet. 47 (2015) doi:10.1038/ng.3192.
[19] T. Beck, R.K. Hastings, S. Gollapudi, R.C. Free, A.J. Brookes, GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies, Eur. J. Hum. Genet. 22 (2014) 949–952, doi:10.1038/ejhg.2013.274.
[20] R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, et al., An atlas of active enhancers across human cell types and tissues, Nature 507 (2014) 455–461, doi:10.1038/nature12787.
[21] E. Khurana, Y. Fu, V. Colonna, X.J. Mu, H.M. Kang, T. Lappalainen, et al., Integrative annotation of variants from 1092 humans: application to cancer genomics, Science 342 (2013) 1235587, doi:10.1126/science.1m235587.
[22] J. Wegert, N. Ishaque, R. Vardapour, C. Geörg, Z. Gu, M. Bieg, et al., Mutations in the SIX1/2 pathway and the DROSHA/DGCR8 miRNA microprocessor complex underlie high-risk blastemal type wilms tumors, Cancer Cell 27 (2015) 298–311, doi:10.1016/j.ccell.2015.01.002.
[23] P. Chaluvally-Raghavan, F. Zhang, S. Pradeep, M.P. Hamilton, X. Zhao, R. Rupaimoole, et al., Copy number gain of hsa-miR-569 at 3q26.2 leads to loss of TP53INP1 and aggressiveness of epithelial cancers, Cancer Cell 26 (2014) 863–879, doi:10.1016/j.ccell.2014.10.010.
[24] Y.-Y. Tseng, B.S. Moriarity, W. Gong, R. Akiyama, A. Tiwari, H. Kawakami, et al., PVT1 dependence in cancer with MYC copy-number increase, Nature 82 (2014) doi:10.1038/nature13311.
[25] W. Kwanhian, D. Lenze, J. Alles, N. Motsch, S. Barth, C. Döll, et al., MicroRNA-142 is mutated in about 20% of diffuse large B-cell lymphoma, Cancer Med. 1 (2012) 141–155, doi:10.1002/cam4.29.
[26] H. Ling, R. Spizzo, Y. Atlasi, M. Nicoloso, M. Shimizu, R.S. Redis, et al., CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer, Genome Res. 23 (2013) 1446–1461, doi:10.1101/gr.152942.112.
[27] S. Ren, Z. Peng, J.-H. Mao, Y. Yu, C. Yin, X. Gao, et al., RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings, Cell Res. 22 (2012) 806–821, doi:10.1038/cr.2012.30.
[28] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814, doi:10.1093/nar/gkg509.
[29] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al., A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249, doi:10.1038/nmeth0410-248.
[30] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucleic Acids Res. 39 (2011) 1–14, doi:10.1093/nar/gkr407.
[31] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V.E. Velculescu, K.W. Kinzler, et al., Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations, Cancer Res. 69 (2009) 6660–6667, doi:10.1158/0008-5472.CAN-09-1133.
[32] P. Yue, W.F. Forrest, J.S. Kaminker, S. Lohr, Z. Zhang, G. Cavet, Inferring the functional effects of mutation through clusters of mutations in homologous proteins, Hum. Mutat. 31 (2010) 264–271, doi:10.1002/humu.21194.
[33] Y. Bromberg, B. Rost, SNAP: predict effect of non-synonymous polymorphisms on function, Nucleic Acids Res. 35 (2007) 3823–3835, doi:10.1093/nar/gkm238.
[34] J.S. Kaminker, Y. Zhang, C. Watanabe, Z. Zhang, CanPredict: a computational tool for predicting cancer-associated missense mutations, Nucleic Acids Res. 35 (2007) 595–598, doi:10.1093/nar/gkm405.
[35] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, et al., Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (2009) 2744–2750, doi:10.1093/bioinformatics/btp528.
[36] A. González-Pérez, N. López-Bigas, Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel, Am. J. Hum. Genet. 88 (2011) 440–449, doi:10.1016/j.ajhg.2011.03.004.
[37] K. Frousios, C.S. Iliopoulos, T. Schlitt, M.A. Simpson, Predicting the functional consequences of non-synonymous DNA sequence variants – evaluation of bioinformatics tools and development of a consensus strategy, Genomics 102 (2013) 223–228, doi:10.1016/j.ygeno.2013.06.005.

[38] M.C. Lopes, C. Joyce, G.R.S. Ritchie, S.L. John, F. Cunningham, J. Asimit, et al., A combined functional annotation score for non-synonymous variants, Hum. Hered. 73 (2012) 47–51, doi:10.1159/000334984.
[39] L. Bao, M. Zhou, Y. Cui, nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms, Nucleic Acids Res. 33 (2005) 480–482, doi:10.1093/nar/gki372.
[40] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, et al., PANTHER: a library of protein families and subfamilies indexed by function, Genome Res. 13 (2003) 2129–2141, doi:10.1101/gr.772403.
[41] E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics 22 (2006) 2729–2734, doi:10.1093/bioinformatics/btl423.
[42] R. Calabrese, E. Capriotti, P. Fariselli, P.L. Martelli, R. Casadio, Functional annotations improve the predictive score of human disease-related mutations in proteins, Hum. Mutat. 30 (2009) 1237–1244, doi:10.1002/humu.21047.
[43] E.A. Stone, A. Sidow, Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity, Genome Res. (2005) 978–986, doi:10.1101/gr.3804205.
[44] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M.P. Schroeder, A. Jene-Sanz, et al., IntOGen-mutations identifies cancer drivers across tumor types, Nat. Methods 10 (2013) 1081–1082, doi:10.1038/nmeth.2642.
[45] T. Kawabata, M. Ota, K. Nishikawa, The protein mutant database, Nucleic Acids Res. 27 (1999) 355–357, doi:10.1093/nar/27.1.355.
[46] T. Sjöblom, L.D. Wood, D.W. Parsons, J. Lin, T.D. Barber, D. Mandelker, et al., The consensus coding sequences of human breast and colorectal cancers, Science 314 (2006) 268–274, doi:10.1126/science.1133427.
[47] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, et al., Patterns of somatic mutation in human cancer genomes, Nature 446 (2007) 153–158, doi:10.1038/nature05610.
[48] P.D. Stenson, M. Mort, E.V. Ball, K. Howells, A.D. Phillips, N.S. Thomas, et al., The Human Gene Mutation Database: 2008 update, Genome Med. 1 (2009) 13, doi:10.1186/gm13.
[49] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370, doi:10.1093/nar/gkg095.
[50] R.J. Clifford, M.N. Edmonson, C. Nguyen, K.H. Buetow, Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms, Bioinformatics 20 (2004) 1006–1014, doi:10.1093/bioinformatics/bth029.
[51] J.S. Kaminker, Y. Zhang, A. Waugh, P.M. Haverty, B. Peters, D. Sebisanovic, et al., Distinguishing cancer-associated missense mutations from common polymorphisms, Cancer Res. 67 (2007) 465–473, doi:10.1158/0008-5472.CAN-06-1736.
[52] J. Thusberg, A. Olatubosun, M. Vihinen, Performance of mutation pathogenicity prediction methods on missense variants, Hum. Mutat. 32 (2011) 358–368, doi:10.1002/humu.21445.
[53] B. Giardine, C. Riemer, T. Hefferon, D. Thomas, F. Hsu, J. Zielenski, et al., PhenCode: connecting ENCODE data with mutations and phenotype, Hum. Mutat. 28 (2007) 554–562, doi:10.1002/humu.20484.
[54] Y. Chen, F. Cunningham, D. Rios, W.M. McLaren, J. Smith, B. Pritchard, et al., Ensemble variation resources, BMC Genomics 11 (2010) 293, doi:10.1186/1471-2164-11-293.
[55] A. Gonzalez-Perez, N. Lopez-Bigas, Functional impact bias reveals cancer drivers, Nucleic Acids Res. 40 (2012) 1–10, doi:10.1093/nar/gks743.
[56] D. Tamborero, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes, Bioinformatics 29 (2013) 2238–2244, doi:10.1093/bioinformatics/btt395.
[57] N.D. Dees, Q. Zhang, C. Kandoth, M.C. Wendl, W. Schierding, D.C. Koboldt, et al., MuSiC: identifying mutational significance in cancer genomes, Genome Res. 22 (2012) 1589–1598, doi:10.1101/gr.134635.111.
[58] E. Hodis, I.R. Watson, G.V. Kryukov, S.T. Arold, M. Imielinski, J.P. Theurillat, et al., A landscape of driver mutations in melanoma, Cell 150 (2012) 251–263, doi:10.1016/j.cell.2012.06.024.
[59] L.D. Wood, D.W. Parsons, S. Jones, J. Lin, T. Sjöblom, R.J. Leary, et al., The genomic landscapes of human breast and colorectal cancers, Science 318 (2007) 1108–1113, doi:10.1126/science.1145720.
[60] P.J. Killela, Z.J. Reitman, Y. Jiao, C. Bettegowda, N. Agrawal, L.A. Diaz, et al., TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) 6021–6026, doi:10.1073/pnas.1303607110.
[61] F.W. Huang, E. Hodis, M.J. Xu, G.V. Kryukov, L. Chin, L.A. Garraway, Highly recurrent TERT promoter mutations in human melanoma, Science 339 (2013) 957–959, doi:10.1126/science.1229259.
[62] S. Horn, A. Figl, P.S. Rachakonda, C. Fischer, A. Sucker, A. Gast, et al., TERT promoter mutations in familial and sporadic melanoma, Science 339 (2013) 959–961, doi:10.1126/science.1230062.
[63] Y. Sato, T. Yoshizato, Y. Shiraishi, S. Maekawa, Y. Okuno, T. Kamura, et al., Integrated molecular analysis of clear-cell renal cell carcinoma, Nat. Genet. 45 (2013) 860–867, doi:10.1038/ng.2699.
[64] M. Kasowski, F. Grubert, C. Heffelfinger, M. Hariharan, A. Asabere, S.M. Waszak, et al., Variation in transcription factor binding among humans, Science 328 (2010) 232–235, doi:10.1126/science.1183621.
[65] R. McDaniell, B. Lee, L. Song, Z. Liu, A.P. Boyle, M.R. Erdos, et al., Heritable individual-specific, Science 328 (2010) 235–240.
[66] W. Zheng, H. Zhao, E. Mancera, L.M. Steinmetz, M. Snyder, Genetic analysis of variation in transcription factor binding in yeast, Nature 464 (2010) 1187–1191, doi:10.1038/nature08934.
[67] K. Sugimachi, A. Niida, K. Yamamoto, T. Shimamura, S. Imoto, H. Iinuma, et al., Allelic imbalance at an 8q24 oncogenic SNP is involved in activating MYC in human colorectal cancer, Ann. Surg. Oncol. (2014) 10434, doi:10.1245/s10434-013-3468-6.
[68] B.N. French, et al., Functional variants at the 11q13 breast cancer risk Loci regulate cyclin D1 expression through long-range enhancers, Am. J. Hum. Genet. 92 (2013) 78540526, doi:10.1016/j.ajhg.2013.01.002.
[69] J. Jiang, P. Jia, B. Shen, Z. Zhao, Top associated SNPs in prostate cancer are significantly enriched in cis -expression quantitative trait loci and at transcription factor binding sites, Oncotarget 5 (2012).
[70] V.C. Lin, C.-Y. Huang, Y.-C. Lee, C.-C. Yu, T.-Y. Chang, T.-L. Lu, et al., Genetic variations in TP53 binding sites are predictors of clinical outcomes in prostate cancer patients, Arch. Toxicol. 88 (2014) 901–911, doi:10.1007/s00204-014-1196-8.
[71] S.-P. Huang, V.C. Lin, Y.-C. Lee, C.-C. Yu, C.-Y. Huang, T.-Y. Chang, et al., Genetic variants in nuclear factor-kappa B binding sites are associated with clinical outcomes in prostate cancer patients, Eur. J. Cancer 49 (2013) 3729–3737, doi:10.1016/j.ejca.2013.07.012.
[72] T. Sterne-Weiler, J.R. Sanford, Exon identity crisis: disease-causing mutations that disrupt the splicing code, Genome Biol. 15 (2014) 201, doi:10.1186/gb4150.
[73] M. Krawczak, N.S.T. Thomas, B. Hundrieser, M. Mort, M. Wittig, J. Hampe, et al., Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing, Hum. Mutat. 28 (2007) 150–158, doi:10.1002/humu.20400.
[74] C. He, F. Zhou, Z. Zuo, H. Cheng, R. Zhou, A global view of cancer-specific transcript variants by subtractive transcriptome-wide analysis, PLoS ONE 4 (2009) doi:10.1371/journal.pone.0004732.
[75] J.P. Venables, R. Klinck, A. Bramard, L. Inkel, G. Dufresne-Martin, C. Koh, et al., Identification of alternative splicing markers for breast cancer, Cancer Res. 68 (2008) 9525–9531, doi:10.1158/0008-5472.CAN-08-1769.
[76] I.M. Shapiro, A.W. Cheng, N.C. Flytzanis, M. Balsamo, J.S. Condeelis, M.H. Oktay, et al., An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype, PLoS Genet. 7 (2011) doi:10.1371/journal.pgen.1002218.
[77] M.D. Taylor, N. Gokgoz, I.L. Andrulis, T.G. Mainprize, J.M. Drake, J.T. Rutka, Familial posterior fossa brain tumors of infancy secondary to germline mutation of the hSNF5 gene, Am. J. Hum. Genet. 66 (2000) 1403–1406, doi:10.1086/302833.
[78] H. Kurahashi, K. Takami, T. Oue, Biallelic inactivation of the APC gene in hepatoblastoma Binlie! ic inactivation of the APC gene in hepatoblastoma1, Cancer Res. 5 (1995) 5007–5011.
[79] M. Manikandan, G. Raksha, A.K. Munirajan, Haploinsufficiency of tumor suppressor genes is driven by the cumulative effect of micrornas, microrna binding site polymorphisms and microrna polymorphisms: an in silico approach, Cancer Inform. 11 (2012) 157–171, doi:10.4137/CIN.S10176.
[80] V. Vaishnavi, M. Manikandan, A.K. Munirajan, Mining the 3'UTR of autism-implicated genes for SNPs perturbing MicroRNA regulation, Genomics Proteomics Bioinformatics 12 (2014) 92–104, doi:10.1016/j.gpb.2014.01.003.
[81] B. Kamaraj, C. Gopalakrishnan, R. Purohit, In silico analysis of miRNA-mediated gene regulation in OCA and OA genes, Cell Biochem. Biophys. 70 (2014) 12013, doi:10.1007/s12013-014-0152-9.
[82] C. Gopalakrishnan, B. Kamaraj, R. Purohit, Mutations in microRNA binding sites of CEP genes involved in cancer, Cell Biochem. Biophys. 70 (2014) 1–10, doi:10.1007/s12013-014-0153-8.
[83] M. Manikandan, A.K. Munirajan, Single nucleotide polymorphisms in microRNA binding sites of oncogenes: implications in cancer and pharmacogenomics, OMICS 18 (2014) 142–154, doi:10.1089/omi.2013.0098.
[84] B. Xie, Q. Ding, H. Han, D. Wu, MiRCancer: a microRNA-cancer association database constructed by text mining on literature, Bioinformatics 29 (2013) 638–644, doi:10.1093/bioinformatics/btt014.
[85] R.A. Gupta, N. Shah, K.C. Wang, J. Kim, H.M. Horlings, D.J. Wong, et al., Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis, Nature 464 (2010) 1071–1076, doi:10.1038/nature08975.
[86] Y. Okugawa, Y. Toiyama, K. Hur, S. Toden, S. Saigusa, K. Tanaka, et al., Metastasis-associated long non-coding RNA drives gastric cancer development and promotes peritoneal metastasis, Carcinogenesis 35 (2014) 2731, doi:10.1093/carcin/bgu200.
[87] P. Ji, S. Diederichs, W. Wang, S. Böing, R. Metzger, P.M. Schneider, et al., MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer, Oncogene 22 (2003) 8031–8041, doi:10.1038/sj.onc.1206928.
[88] T. Gutschner, M. Hämmerle, M. Eißmann, J. Hsu, Y. Kim, G. Hung, et al., The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells, Cancer Res. 73 (2013) 1180–1189, doi:10.1158/0008-5472.CAN-12-2850.
[89] G. Jin, J. Sun, S.D. Isaacs, K.E. Wiley, S.T. Kim, L.W. Chu, et al., Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk, Carcinogenesis 32 (2011) 1655–1659, doi:10.1093/carcin/bgr187.
[90] L. Li, R. Sun, Y. Liang, X. Pan, Z. Li, P. Bai, et al., Association between polymorphisms in long non-coding RNA PRNCR1 in 8q24 and risk of colorectal

cancer, J. Exp. Clin. Cancer Res. 32 (2013) 1–7, http://dx.doi.org/10.1186/1756-9966-32-104.
[91] D.S. Gross, W.T. Garrard, Nuclease hypersensitive sites in chromatin, Annu. Rev. Biochem. 57 (1988) 159–197, doi:10.1146/annurev.biochem.57.1.159.
[92] Y. He, J.A. Carrillo, J. Luo, Y. Ding, F. Tian, I. Davidson, et al., Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells, Front. Genet. 5 (2014) 1–9, doi:10.3389/fgene.2014.00308.
[93] M.A. Dawson, T. Kouzarides, Cancer epigenetics: from mechanism to therapy, Cell 150 (2012) 12–27, doi:10.1016/j.cell.2012.06.013.
[94] G.E. Zentner, P.J. Tesar, P.C. Scacheri, Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions, Genome Res. 21 (2011) 1273–1283, doi:10.1101/gr.122382.111.
[95] A. Rada-Iglesias, R. Bajpai, T. Swigut, S.A. Brugmann, R.A. Flynn, J. Wysocka, A unique chromatin signature uncovers early developmental enhancers in humans, Nature 470 (2011) 279–283, doi:10.1038/nature09692.
[96] N.P. Access, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74, doi:10.1038/nature11247.An.
[97] C.B. Lowe, D. Haussler, 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome, PLoS ONE 7 (2012) doi:10.1371/journal.pone.0043128.
[98] B.E. Bernstein, J.A. Stamatoyannopoulos, J.F. Costello, B. Ren, A. Milosavljevic, A. Meissner, et al., The NIH roadmap epigenomics mapping consortium, Nat. Biotechnol. 28 (2010) 1045–1048, doi:10.1038/nbt1010-1045.
[99] S. Yu, K. Cui, R. Jothi, D.M. Zhao, X. Jing, K. Zhao, et al., GABP controls a critical transcription regulatory module that is essential for maintenance and differentiation of hematopoietic stem/progenitor cells, Blood 117 (2011) 2166–2178, doi:10.1182/blood-2010-09-306563.
[100] T. Zeller, P. Wild, S. Szymczak, M. Rotival, A. Schillert, R. Castagne, et al., Genetics and beyond – the transcriptome of human monocytes and disease susceptibility, PLoS ONE 5 (2010) doi:10.1371/journal.pone.0010693.
[101] H.S. Rhee, B.F. Pugh, Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution, Cell 147 (2011) 1408–1419, doi:10.1016/j.cell.2011.11.013.
[102] C.G. Palii, C. Perez-Iratxeta, Z. Yao, Y. Cao, F. Dai, J. Davison, et al., Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages, EMBO J. 30 (2011) 494–509, doi:10.1038/emboj.2010.342.
[103] J.F. Degner, A.A. Pai, R. Pique-Regi, J.-B. Veyrieras, D.J. Gaffney, J.K. Pickrell, et al., DNase I sensitivity QTLs are a major determinant of human expression variation, Nature 482 (2012) 390–394, doi:10.1038/nature10808.
[104] A.P. Boyle, E.L. Hong, M. Hariharan, Y. Cheng, M.A. Schaub, M. Kasowski, et al., Annotation of functional variation in personal genomes using RegulomeDB, Genome Res. 22 (2012) 1790–1797, doi:10.1101/gr.137323.112.
[105] Y. Fu, Z. Liu, S. Lou, J. Bedford, X.J. Mu, K.Y. Yip, et al., FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer, Genome Biol. 15 (2014) 1–15, doi:10.1186/s13059-014-0480-5.
[106] G.R.S. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Functional annotation of noncoding sequence variants, Nat. Methods 11 (2014) 294–296, doi:10.1038/nmeth.2832.
[107] M. Kircher, D.M. Witten, P. Jain, B.J. O'Roak, G.M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants, Nat. Genet. 46 (2014) 310–315, doi:10.1038/ng.2892.
[108] N. Hs, H. Tr, Q. Morris, Y. Barash, K. Ar, N. Jojic, et al., RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease, Science 347 (2015) 1254806, doi:10.1126/science.1254806.
[109] H.A. Shihab, M.F. Rogers, J. Gough, M. Mort, D.N. Cooper, I.N.M. Day, et al., An integrative approach to predicting the functional effects of non-coding and coding sequence variation, Bioinformatics 31 (2015) 1536–1543, doi:10.1093/bioinformatics/btv009.
[110] L. Clarke, X. Zheng-Bradley, R. Smith, E. Kulesha, C. Xiao, I. Toneva, et al., The 1000 Genomes Project: data management and community access, Nat. Methods 9 (2012) 459–462, doi:10.1038/nmeth.1974.
[111] M.J. Landrum, J.M. Lee, G.R. Riley, W. Jang, W.S. Rubinstein, D.M. Church, et al., ClinVar: public archive of relationships among sequence variation and human phenotype, Nucleic Acids Res. 42 (2014) 980–985, doi:10.1093/nar/gkt1113.
[112] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864, doi:10.1126/science.185.4154.862.