Hum Genet (2009) 126:481–498
DOI 10.1007/s00439-009-0677-y

REVIEW ARTICLE

Analytical methods for inferring functional effects of single base pair substitutions in human cancers

William Lee · Peng Yue · Zemin Zhang

Received: 7 April 2009 / Accepted: 29 April 2009 / Published online: 12 May 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract Cancer is a genetic disease that results from a variety of genomic alterations. Identification of some of these causal genetic events has enabled the development of targeted therapeutics and spurred efforts to discover the key genes that drive cancer formation. Rapidly improving sequencing and genotyping technology continues to generate increasingly large datasets that require analytical methods to identify functional alterations that deserve additional investigation. This review examines statistical and computational approaches for the identification of functional changes among sets of single-nucleotide substitutions. Frequency-based methods identify the most highly mutated genes in large-scale cancer sequencing efforts while bioinformatics approaches are effective for independent evaluation of both non-synonymous mutations and polymorphisms. We also review current knowledge and tools that can be utilized for analysis of alterations in non-protein-coding genomic sequence.

Introduction

Cancer is a complex genetic disease and understanding the myriad genetic factors involved in oncogenesis is an important step towards prevention and treatment. During the stepwise process of tumorigenesis, cells acquire a series of somatic mutations that lead to the excessive cell growth and eventually lead to the development of cancer. The progression to cancer can be accelerated when the individual also carries a germ-line mutation in a cancer susceptibility gene (Knudson 1971). According to the Cancer Gene Census (Futreal et al. 2004), the majority of known cancer mutations are somatic mutations, but some germline polymorphisms with a connection to cancer have also been identified. Identification of these mutations and polymorphisms can lead to the discovery of the genes that control cancer development and, therefore, also serve as attractive therapeutic targets.

The importance of a targeted approach towards cancer treatment has been emphasized by a number of successful therapies brought to market in recent years. Novartis' Gleevec is an example of a drug that resulted from the identification of a cancer-causing genetic abnormality (Druker et al. 2001). A chromosomal translocation resulting in the constitutively active protein tyrosine kinase *bcr-abl* was identified as the casual event in development of chronic myelogenous leukemia (Lugo et al. 1990). A small molecule compound was discovered through high-throughput screening as a potent inhibitor of *bcr-abl* and it was then developed into Gleevec, a commercial therapy for inhibiting *bcr-abl* to block tumor growth while having minimal effect upon normal cells. Several other drugs have also been developed to target specific proteins that are commonly mutated in cancers. For example, Genentech's Herceptin is a HER2-specific antibody, which is effective in treating breast cancers that overexpress the gene HER2, and AstraZeneca's Iressa was the first of several EGFR inhibitors to treat carcinomas that have excess EGFR activity (Ciardiello et al. 2000; Vogel et al. 2002).

Rapid improvements in genomic technologies have allowed for large-scale genotyping and sequencing of cancer tissues and normal genomes as well. This influx of sequence

W. Lee and P. Yue contributed equally to this work.

This work was performed at Genentech, Inc., South San Francisco, CA, USA.

W. Lee · P. Yue · Z. Zhang (✉)
Department of Bioinformatics, Genentech, Inc., 1 DNA Way, M.S. 93, South San Francisco, CA 94080, USA
e-mail: zemin@gene.com

data has revealed a vast array of genetic variations present in cancer, with a large portion of both somatic mutations and naturally occurring variations in the form of single-nucleotide substitutions. Among these single-nucleotide changes, missense mutations in which a single-nucleotide change within a gene results in an amino acid substitution in the protein product are the most investigated (Ding et al. 2008; Forbes et al. 2008; Greenman et al. 2007; Jones et al. 2008; TCGA 2008; Parsons et al. 2008; Sjoblom et al. 2006; Wood et al. 2007). The primary question facing the interpretation of this wealth of data is the delineation of functional mutations from those that are simply the result of the genetic instability inherent in cancer genomes.

The most common ways of analyzing missense mutations are focused on two distinct but related goals. In the case of recently published large-scale sequencing efforts, the analysis is gene-centric and attempts to identify highly mutated genes that are, therefore, likely to be important in the development of a specific cancer (Ding et al. 2008; Greenman et al. 2007; Jones et al. 2008; TCGA 2008; Parsons et al. 2008; Sjoblom et al. 2006; Wood et al. 2007). The premise behind this frequency-based approach is that genes that are mutated significantly more often than would be expected by chance probably function to favor tumor growth when mutated. This methodology requires a large dataset to provide sufficient statistical power and its strength lies in the identification of important genes in the condition of interest. Complementary to this approach is a mutation-centric view that removes a given mutation from the disease context in which it was observed and attempts to predict its functionality based solely on the substitution itself. These methods have the benefit of being able to potentially identify the actual causal mutation, as opposed to just the causal gene. Identification of specific functional mutations could give additional insight into the biological mechanisms of the disease.

Although the majority of large-scale sequencing efforts to date have focused on protein-coding regions, next generation sequencing technologies are beginning to allow for whole-genome sequencing of individual samples (Ley et al. 2008; Wheeler et al. 2008). This will bring in a wealth of information on mutations occurring in non-genic genomic regions, which will in turn require different analysis techniques. Single-nucleotide polymorphism (SNP) analysis has shown that alterations in non-coding sequences can have significant functional effects and contributions towards disease (Chorley et al. 2008; Srebrow and Kornblihtt 2006), and so making full use of whole-genome sequencing data will require analysis of mutations found outside of genes. The tools for predicting the functionality of a non-coding mutation are limited, but there exist a number of methods and databases that attempt to map the various non-coding functional regions to the genome through sequence analysis (Cartegni et al. 2003; Conde et al. 2006; Enright et al. 2003; Freimuth et al. 2005; Griffiths-Jones et al. 2008; Hallikas et al. 2006; Kim et al. 2008; Lambert et al. 2004; Matys et al. 2003; Palin et al. 2006; Ponomarenko et al. 2002, 2003; Riva and Kohane 2002; Sandelin et al. 2004; Tabaska and Zhang 1999; Thierry-Mieg and Thierry-Mieg 2006; Wang 2008). These tools model specific sequence elements such as transcription factor-binding sites based on the experimentally verified regions and provide the most effective means of large-scale prediction of where functional sites lie. They can also help in functional prediction by first identifying mutations that lie in functional elements, and if so, which ones may perturb the functionality of that element. Most such tools give a quantitative measure of how likely a given sequence is to be a functional element of interest (e.g. transcription factor-binding site), and so simply examining the difference in scores between a mutated sequence and its original sequence can give an idea of the functionality of the mutation.

Distinct from the study of somatic mutations in cancer is the investigation of naturally occurring human germline mutations that could contribute to the risk of cancer and other genetic diseases. A large number of SNPs have been identified, but functional information is still sparse (Sherry et al. 2001). Large-scale systematic genotyping projects (Frazer et al. 2007) have employed high-throughput genotyping technologies that enable investigations into associations between variation and disease risk. Genome-wide association studies (GWAS) have discovered SNPs that contribute to the risk of cancer development, but many of the identified risk alleles require additional analysis to validate and understand (Amos et al. 2008; Broderick et al. 2007; Easton et al. 2007; Eeles et al. 2008; Gold et al. 2008; Gudmundsson et al. 2007; Hunter et al. 2007; Kingsmore et al. 2008; Tenesa et al. 2008; Thomas et al. 2008; Tomlinson et al. 2008; Zanke et al. 2007). SNPs are distributed throughout the human genome in both coding and non-coding regions, but many methods used for analyzing mutations are equally valuable for evaluating the potential functionality of SNPs, which could be a valuable step towards the interpretation and utilization of GWAS data.

This review focuses on the analysis of single-nucleotide substitutions in the context of cancer, with a particular spotlight on recent large-scale cancer genome sequencing projects. We examine the methods by which cancer sequencing efforts can leverage their data to identify disease-driving genes and provide an overview of amino acid-change-based bioinformatics analysis methods, many of which are also applicable to non-cancer inherited diseases (Karchin 2009; Mooney 2005; Ng and Henikoff 2006; Steward et al. 2003). We also review the current knowledge of functional mutations that act in ways other than alteration of protein

sequence, such as mutations that alter gene expression or splicing.

Mutation frequency-based analysis

Several large-scale cancer genome exon re-sequencing projects have recently been published by four publicly funded consortiums including groups at John Hopkins University (JHU) (Jones et al. 2008; Parsons et al. 2008; Sjoblom et al. 2006; Wood et al. 2007), Sanger Institute (Greenman et al. 2007), the Cancer Genome Atlas (TCGA 2008), and the Tumor Sequencing Project (TSP) (Ding et al. 2008) (Table 1). The JHU group focused on sequencing nearly the complete human transcriptome with a limited number of samples (11–24 samples in each cancer type), whereas the other groups sequenced a smaller number of candidate genes and, therefore, could afford to cover a larger number of samples in a single cancer type (as high as 188 lung adenocarcinoma samples in the case of the TSP study). Although the first strategy allows for the detection of novel cancer genes due to the larger search space, the latter strategy enables a broader survey of possible mutations in genes that are already known to be involved in cancer.

It has been estimated that most of the observed cancer mutations are functionally neutral and are, thus, often referred to as passenger mutations, while a smaller set of driver mutations will actually confer growth advantages to tumors (Greenman et al. 2007; Sjoblom et al. 2006). Driver mutations increase the fitness of cells that they reside in and are assumed to be under positive selection during the multistage neoplastic progression. This selection should result in driver mutations occurring more frequently in tumor samples; hence, the most common approach for identifying driver mutations is based on calculating mutation frequencies with the assumption that a higher prevalence implies functionality. This method of frequency-based analysis typically requires an estimation of the non-synonymous background mutation rate (nsBMR) followed by calculation of the statistical likelihood of observing a certain number of mutations based on the nsBMR. For example, a gene harboring a significantly greater number of mutations than expected by chance would be considered a driver, since it is likely that mutations in that gene are selected for during oncogenesis. In the following sections, we will discuss how mutation data were analyzed in the recently published cancer genome studies with regard to the above-referenced steps.

It is crucial to determine a valid nsBMR, which, if underestimated, would overstate the significance of the observed mutations. The nsBMR can be estimated empirically from presumed passenger mutations as shown in the studies conducted by Jones et al. (2008); Parsons et al. (2008). These studies estimated that the nsBMR from the set of genes remaining after the most highly mutated previously known driver genes were removed from the dataset. More commonly, nsBMR is indirectly estimated as the product of the mutation rate of synonymous mutations and the expected ratio of the number of passenger non-synonymous mutations to the number of synonymous mutations (NS/S). With rare exceptions, a synonymous mutation is not likely to change the function of the protein and is, therefore, usually considered functionally neutral and not subject to selective pressure (Greenman et al. 2006). The passenger NS/S ratio is obtained by dividing the total number of possible non-synonymous changes by the total number of possible synonymous changes within the sequenced nucleotides (Ding et al. 2008; TCGA 2008; Wood et al. 2007). This ratio, ranging from 2 to 3, may result in an overestimation of the nsBMR because some of the possible non-synonymous mutations may be detrimental to the growth of the tumors and are, thus, under negative selection. A different approach is to use the observed NS/S in human population SNPs, approximately 1 (Jones et al. 2008; Parsons et al. 2008; Wood et al. 2007), which may result in an underestimation due to a greater selective pressure on germline mutations. Because these approaches, respectively, delineate an upper bound and lower bound for the NS/S ratio, the average of the two values is used by the Jones et al. (2008); Parsons et al. (2008) studies.

Estimation of a background mutation rate can be significantly affected by mutation rate heterogeneity across different DNA contexts. For example, CpG dinucleotides have a much higher mutation rate (up to 6.44-fold higher than the overall mutation frequency in one colorectal dataset (Sjoblom et al. 2006)) compared with other DNA contexts. Owing to this context-dependence, it can be beneficial to partition mutations into multiple types to account for such variations (Ding et al. 2008; Greenman et al. 2007; Jones et al. 2008; TCGA 2008; Parsons et al. 2008; Sjoblom et al. 2006; Stephens et al. 2005; Wood et al. 2007). The relative mutation rates of the different DNA contexts are usually measured directly using mutations that are presumably non-functional, such as synonymous mutations or mutations observed on the least frequently mutated genes (TCGA 2008). DNA context groups can be defined either based on prior knowledge, such as the high mutation rate at CpG dinucleotides, or using data-driven methods, which may better capture the heterogeneity of the mutation rates across different nucleotide contexts (Ding et al. 2008; Jones et al. 2008; TCGA 2008; Parsons et al. 2008; Sjoblom et al. 2006; Wood et al. 2007). In a recent study of lung carcinoma, Ding et al. 2008 partitioned all of the observed mutations into 192 categories with consideration of all 12 possible mutation changes within 16 possible flanking dinucleotides (5′ and 3′). Observed mutation rates for each

category were calculated and low frequency categories were then collapsed if they did not show statistically distinct mutation rates ($P < 0.05$, Fisher exact test). This process resulted in 18 distinct categories. Recently, the JHU group and the TCGA group each published a glioblastoma study in which they reported two quite different background mutation rates: 0.38–1.02 (estimated lower and upper bounds) and 3.70 ± 0.57, respectively (TCGA 2008; Parsons et al. 2008). This discrepancy is likely due to the heterogeneity between the two sample populations and gene sets, as evidenced by the difference in the observed synonymous mutation rate (0.37 and 1.29). Further study might suggest that the most effective method is to use gene-specific nsBMRs to reduce the disparities between separate studies, but a large amount of data is necessary to use this method effectively.

Several studies identify novel putative cancer genes using statistical methods, some of which have stirred controversies (Forrest and Cavet 2007; Getz et al. 2007; Rubin and Green 2007). Most of these studies applied the one-tailed binomial test to identify significantly mutated genes, followed by a false discovery rate procedure to control for multiple testing (Benjamini and Hochberg 1995). As mentioned above, a single fixed background mutation rate was used in the simplest version while a more complex approach took into account the DNA context of each mutation to adjust for the heterogeneity of mutation rates under different DNA contexts. As shown in TCGAs analysis, the context-specific method is more sensitive when a smaller set of samples is analyzed (TCGA 2008). It is also worth noting that a custom method was used by the JHU group to incorporate their unique two-stage experimental design (discovery and validation screens) into the analysis (Jones et al. 2008; Parsons et al. 2008; Wood et al. 2007). Greenman et al. (2007) used a different strategy of directly modeling the NS/S ratio based on the rationale that if non-synonymous mutations yield amino acid changes with a selective advantage, a higher ratio of NS/S may be observed. Therefore, the significance of the results can be measured as a function of the degree of deviation from the expected 2:1 ratio, which had been used in several early studies (Bardelli et al. 2003; Samuels et al. 2004; Wang et al. 2004). With this model, the selective pressure can be estimated for various gene sets via maximum likelihood by considering the deviation from the expected ratio of non-synonymous to synonymous mutations. The selection pressure can be calculated on different levels, from a single gene to the whole mutation dataset, from which candidate driver genes can be predicted and the number of driver mutations can be estimated.

Another factor that could greatly affect the result of a frequency-based study is the sample size present in the study. A wide range of sample sizes has been processed in the published large-scale sequencing efforts, ranging from 11 in breast and colon cancers to about 200 in lung cancer. The authors of the TCGA study have examined the effects of sample size by randomly selecting subsets of the original 72 samples (TCGA 2008). They found that with as few as 48 samples, all eight of the cancer genes that were identified in their complete set of 72 samples could still be discovered. When the sample size was further reduced, only a fraction of the eight genes could still be identified as significant.

Somatic point mutations may account for only a fraction of the genetic alterations required for tumorigenesis. Integration with other genomic data would greatly enhance the possibility of identifying genes and biological pathways involved in tumor development. In the glioblastoma study, Parsons et al. (2008) integrated mutation analysis with genomic copy number analysis and identified three major signaling pathways with critical genes mutated in a majority of the studied tumors. They also found a mutually exclusive pattern for the alterations within each pathway. The same pattern was also reported in TCGAs glioblastoma paper (TCGA 2008). Ding et al. (2008) found that mutations in known tumor suppressor genes such as PTEN, APC and TP53 were correlated with copy number loss and mutations in proto-oncogenes, such as EGFR, HCK, KRAS and EPHB1. Therefore, an integrative approach to analyze all types of genetic alterations in a pathway context could provide greater insight into the genetic mechanisms of cancer development.

Many of the somatic mutations identified in the recent cancer exon re-sequencing studies are novel and rare mutations, often observed in only a single sample. This implies that a large number of samples are required to establish the statistical significance of potential cancer driver genes. The rapid development of sequencing technology will eventually allow us to expand beyond the current focus on coding regions to the whole human genome and therefore make it possible to identify all of the genetic alterations underlying the individual cancers. In the meantime, it also presents an even bigger statistical challenge since many more mutations need to be analyzed. Despite these challenges, the current large-scale cancer studies have successfully identified many novel cancer genes and provide more insight into the complex genetic basis of cancer (Table 1) .

Bioinformatics analysis of amino acid substitutions

The frequency-based approaches reviewed above are contingent upon either an assumption of a background mutation rate or the availability of a large number of mutations in the dataset to calculate an empirical background mutation rate for the sample of interest. Furthermore, these

Table 1 List of large-scale cancer re-sequencing projects

References	Team	Gene subset	Cancer	No. of samples	No. of genes	Total length (Mbp)	No. of mutations	No. of driver genes	nsBMR (per Mb)
Bardelli et al. (2003)	JHU	Tyrosine kinases	Colon	35	130	4	14	NA	1
Samuels et al. (2004)	JHU	Lipid kinases	Multiple	35	20	1	1	NA	1
Wang et al. (2004)	JHU	Tyrosine phosphatases	Colon	18	87	3.3	6	NA	1
Stephens et al. (2005)	Sanger Center	Kinases	Breast	25	518	31	65	NA	NA
Parsons et al. (2005)	JHU	Ser/Thr kinases	Colon	24	340	1	8	NA	NA
Davies et al. (2005)	Sanger Center	Kinases	Lung	26	518	34	141	NA	NA
Sjoblom et al. (2006)	JHU	CCDS	Breast/colon	22	13,023	462	1,307	191	1.2
Greenman et al. (2007)	Sanger Center	Kinases	Multiple	210	518	274	1,007	119	NA
Wood et al. (2007)	JHU	RefSeq	Breast/colon	11	18,191	645	2,185	280	Colon 0.99–2.35[a] Breast 1.40–3.62[a]
Jones et al. (2008)	JHU	RefSeq + Ensembl	Pancreas	24	20,661	753	1,562	83	0.54–1.38[a]
Parsons et al. (2008)	JHU	RefSeq + Ensembl	Glioblastoma	22	20,661	689	993	42	0.38–1.02[a]
(TCGA 2008)	TCGA Team	Candidate genes	Glioblastoma	91	601	97	453	8	3.7
Ding et al. (2008)	TSP	Candidate genes	Lung	188	623	247	1,013	26	3.3
Ley et al. (2008)	Wash U.	Complete genome	AML	1	25,000	3,000	8	NA	NA

[a] These ranges refer to the lower and upper bounds for calculated non-synonymous background mutation rate in the discovery screen stage of the study

methods cannot be used for independently evaluating the potential function of an individual mutation. Because many genes are infrequently mutated and large disease-specific datasets are often not available, other approaches may be more suitable for identification of functional mutations on an individual gene basis. A set of methods for predicting functions of specific amino acid substitution can fill this niche (Table 2). These methods look at the actual amino acid change occurring in missense mutations and can be used for the analysis of mutations on a case-by-case basis. Such methods have also been effective in the study of natural human genetic variation. Bioinformatics methods can help prioritize which of the greater than 60,000 estimated non-synonymous single SNPs (Livingston et al. 2004) in the human population are likely to have a function impact and warrant additional investigation. Furthermore, the ability of such methods to evaluate the functional impact of individual changes makes them useful for directing mutagenesis efforts, so that mutations that are most likely to produce a phenotype can be examined first (Henikoff and Comai 2003). We will review general features of substitution-based methods and focus specifically on their application towards cancer mutation research.

Amino acid-change-based prediction methods are primarily based on an observation that functional mutations appear to be distributed in a non-random manner across protein sequences and structures (Miller and Kumar 2001; Sunyaev et al. 2000; Wang and Moult 2001). Based only on sequence analysis, Miller and Kumar (2001) observed that disease-associated mutations in seven genes were particularly concentrated in conserved amino acid positions. This observation is consistent with the notion that conserved residues are more likely to be functional, since the conversation at that position is likely due to purifying selection throughout evolution. By adding structural data to their analysis, Sunyaev et al. (2000) found that ~70% of disease-related mutations they studied were located in structural sites more likely to be functionally important, such as active sites, interaction sites, or positions buried within the protein and inaccessible to solvent. In a similar manner, Wang and Moult (2001) modeled the effects of disease-associated SNPs on protein stability and found that 83% of such substitutions were found to affect protein stability.

Given the observations concerning disease-related mutations, algorithms that predict the functionality of amino acid substitutions do so based on sequence information, structure information, or a combination of the two. In sequence-based methods, a substitution will be evaluated based on its sequence context. The widely used SIFT algorithm (Sorting Intolerant From Tolerant) employs a multiple sequence alignment of homologous proteins to identify conserved regions in the protein of interest, each possible substitution can then be scored according to the conservation observed at each position (Ng and Henikoff 2001). Similarly, Clifford et al. describe a tool that takes advantage of known Pfam protein motifs to identify conserved regions in protein domains (Clifford et al. 2004; Finn et al. 2006). Jiang et al. (2007) developed a method consisting of 20 modules, each of which was optimized using a subset of sequence features specific to a particular starting residue. This method was shown to outperform other general methods such as SIFT. More recently, Hon et al. (2009) examined mutations within signal peptide regions and used outputs from the SignalP program to identify mutations that could affect signal peptide function. The authors found that combining SIFT with specific signal peptide information could accurately identify functional mutations within signal peptides. This context-specific approach was also adopted by Radivojac et al. who developed a model using the output of phosphorylation-site predictor DisPhos to assess the probability of losing or gaining phosphorylation sites due to mutation. The application of this model onto a cancer genome dataset (Greenman et al. 2007) revealed that cancer somatic mutations are enriched for mutations that affect phosphorylation sites (Radivojac et al. 2008).

Structure-based amino acid substitution prediction methods rely on an ability to map mutations of interest to a structure. In these methods, the first step is to find a suitable structure for the protein of interest and then identify the possible structural effects that the given amino acid substitution may have. For example, changes that affect solvent accessibility or at sites of protein–protein interactions are more likely to be functional. PolyPhen is a rule-based system that uses structural information and annotation data to identify functionally important sites to predict the potential function of a substitution (Sunyaev et al. 2001).

Just as frequency-based methods are heavily dependent upon available mutation information, substitution-based algorithms are limited by available sequence and structure information. For instance, three-dimensional structures are only available for a small fraction of all proteins, and not in all functional conformations. In this case, applying a structure-based algorithm to a protein with only limited structural information would not produce accurate results (Chasman and Adams 2001; Yue et al. 2005; Yue and Moult 2006). In an analogous manner, sequence-based methods can be limited by the number of available sequences homologous to the protein of interest. Predictions will be less accurate in cases where an inadequate number of sequences are used to identify conserved residues. Even with a large set of homologous sequences, however, it has been shown that many disease-causing mutations are in positions that are not highly conserved across species and could therefore be subject to less accurate analysis by sequence-based methods (Torkamani and Schork 2007b).

Table 2 List of selected amino acid substitution prediction tools

Method	Description
SIFT (Ng and Henikoff 2001, 2002, 2003) http://blocks.fhcrc.org/sift/SIFT.html	Sequence homology based; scores use position-specific scoring matrices with Dirichlet priors
PolyPhen (Sunyaev et al. 2001) http://genetics.bwh.harvard.edu/pph/	Based on sequence homology, structure, and Swissprot annotation. Classification uses rule-based integration of output of multiple subroutines
SNPs3D (Yue et al. 2005, 2006; Yue and Moult 2006) http://www.snps3d.org	Structure-based method based on the Support Vector Machine
PANTHER subPSEC (Thomas et al. 2003) http://www.pantherdb.org/tools/csnpScoreForm.jsp	Sequence homology based; scores use PANTHER-derived Hidden Markov Models
LS-SNP (Karchin 2005) http://alto.compbio.ucsf.edu/LS-SNP/	Based on structure, sequence, and annotation; scores use a Support Vector Machine
TopoSNP (Stitziel et al. 2004) http://gila.bioengr.uic.edu/snp/toposnp	Uses alpha shape method from computational geometry to characterize the structural locations of substitutions
CanPredict (Kaminker et al. 2007a, b) http://www.cgl.ucsf.edu/Research/genentech/canpredict/	Based on sequence homology and Gene Ontology annotation; scores use a Random Forest Classifier
Signal peptide-specific tool (Hon et al. 2009)	Based on the outputs of SignalP; assess the effects of an amino acid change within the signal peptide
Protein phosporylation-specific tool (Radivojac et al. 2008)	Based on the outputs of DisPhos; asses the probability of losing or gaining of a phosphorylation site resulted by a mutation
Kinase-specific tool (Torkamani and Schork 2007a, 2008)	A kinase-specific prediction method; take use of kinase-specific features
MSRV (Jiang et al. 2007)	Sequence-based method consisting of 20 modules, each of which was optimized using a subset of sequence features specific to a particular starting residue

The previously mentioned methods can be generally applied towards analysis of amino acid substitutions. However, cancer driver mutations have particular characteristics that Kaminker et al. (2007a) exploited with the CanPredict algorithm to distinguish cancer mutations from others based on annotation and sequence features. CanPredict applied specific knowledge of cancer mutations to distinguish them from other disease mutations in a manner similar to how disease mutations can be distinguished from non-functional mutations. In particular, CanPredict incorporates sequence-based predictions from two previously mentioned methods (SIFT and Pfam-based LogR.E-value) (Clifford et al. 2004; Ng and Henikoff 2001) as well as annotation information from Gene Ontology (Ashburner et al. 2000) into a random forest classifier (Breiman 2001). This classifier quantifies the differences in these features between cancer-related mutations and others and is then able to provide a call for whether or not a given mutation is likely to be a causal mutation in cancer.

Subsequently, Torkamani and Schork (2008) reported an SVM-based classifier to distinguish cancer driver mutations. A collection of sequence and structure features were used in their model, including sequence conservation measured with the SubPSEC score (Thomas et al. 2003; Thomas and Kejariwal 2004), the wild-type and mutant amino acid identity (Torkamani and Schork 2007b), changes in five amino acid metrics (Atchley et al. 2005), changes in hydropathy, water/octanol partition energy, hydrophobicity, polarity, charge and volume, protein domain information, protein secondary structure, amino acid solvent accessibility and structure flexibility predicted by Wiggle (Gu et al. 2006). Only mutations within protein kinase families are analyzed in this study, which allows the incorporation of two protein kinase-specific features into the model. The subgroup annotation of the specific protein kinase was used as the first feature since the distributions of disease and non-disease mutations within different protein kinase groups are significantly different. The second feature is the subdomain predictor of whether a given mutation falls within the N-terminal or the C-terminal lobe, since disease mutations have a tendency to cluster within the C-terminal lobe rather than the N-terminal lobe. The authors showed that this context-specific method outperforms CanPredict on the kinase mutations. In a different study, Torkamani et al. reported that their method is also superior to other popular methods (SIFT, Polyphen, Pmut, and SNPs3D) applied to germline variants within protein kinases (Torkamani and Schork 2007a). They attributed much of their success to the context-specific training data, where specific protein kinase features such as the group membership can be used.

The protein sequence and structure-based analysis have also been applied in the recent large-scale cancer genome projects in order to help prioritize genes and somatic mutations for further validation (Ding et al. 2008; Jones et al. 2008; TCGA 2008; Parsons et al. 2008; Wood et al. 2007). Ding et al. (2008) used SIFT and PolyPhen to evaluate the potential impact on protein function for 811 missense mutations. SIFT predicted 430 missense mutations as deleterious while PolyPhen predicted 438 mutations as probably/possibly damaging. Taken together, 579 mutations were identified as likely to affect protein function. Wood et al. (2007) used two sequence analysis tools, SIFT and logR.E to prioritize mutations for further analysis. After projecting mutations onto protein structures, they observed that some somatic mutations showed clustering of mutations around active sites of proteins or near an interface residue. In the glioblastoma and pancreatic cancer studies by the same research group, a machine learning classifier using a random forest algorithm, LSMUT, was developed to predict the functional impact of the non-synonymous mutations (Jones et al. 2008; Parsons et al. 2008). Fifty-eight features based on the sequence and structural information of amino acids involved in the alterations were used as the predictive features for the classifier. The classifier was trained on common SNPs as the negative dataset (common SNPs are assumed to be tolerated and therefore not disease-causing) and cancer mutations in the COSMIC database as the positive dataset. The distribution of LSMUT scores of the missense mutations in the top-ranked CAN genes is significantly different from the scores in a set of randomly generated mutations. Approximately 15 and 17.3% of the missense mutations that can be predicted by the classifier were predicted to affect protein function in the glioblastoma and pancreatic cancer datasets, respectively. Furthermore, using protein structure information, they discovered that over 10% mutations (35 in glioblastoma and 55 in pancreatic cancer) are located close to a domain interface or substrate-binding site and thus are likely to affect protein functions.

The frequency-based analysis approaches and amino acid substitution prediction methods have been compared in a few papers. The two methods were found to produce the results that correlate well with each other. Indeed, mutations in genes that have a high CaMP score tend to be also classified as cancer-associated using CanPredict (Hon et al. 2008). Similarly, functional scores computed by combining the predictions of PolyPhen, PMut, SIFT and SNPs3D correlated with the odd ratios identified in association studies (Zhu et al. 2008). More importantly, these two approaches may work together in a complementary nature. For example, CanPredict can capture the functional effect of known driver mutation BRAF V600E (Kaminker et al. 2007a), which was missed by frequency analysis due to the lack of enough samples (Sjoblom et al. 2006). Many prediction tools have been developed in recent years to identify the

potential functionality of amino acid substitutions (Table 2). However, these tools often produce inconsistent results, making interpretation of the individual results more difficult. Chan et al. (2007) show that when the predictions of four different methods are in agreement the prediction accuracy is significantly improved. Others have proposed a metaserver to enable end users to more easily access consensus prediction results from different prediction servers (Karchin 2009; Ng and Henikoff 2006). With high-throughput, next generation sequencing technology becoming increasingly reliable and affordable, future cancer genome studies are likely to be sequencing the entire genome in large collections of cancer samples. Therefore, the power of frequency-based analysis will be increased accordingly. In the meantime, the amino acid-based bioinformatics analysis will become even more critical as more rare mutations are identified.

It is well known that there are some recurrent cancer "hotspot" mutations that can be observed in many samples, for example BRAF V600E is reported in over 4,000 samples according to COSMIC database. Moreover, cancer mutations are also found to be located at the locations that are analogous to other known mutations. For example, the T790M mutation in EGFR occurs at the same residue in the kinase domain as other known mutations in BCR-ABL, PDGFRA and KIT (Kobayashi et al. 2005). Marks et al. (2007) reported a novel mutation in the kinase domain of FGFR4, which is located at an analogous location to a known cancer mutation in ERBB2, which lead to the development of the "Mutagrator" web site (http://cbio.mskcc.org/mutagrator/) to capture a few other analogous mutation clusters in the protein kinase domain. Wood et al. also discovered a number of cancer mutations that occurred at locations identical to those of genes involved in human germline diseases. Based on this concept, one may develop a mutation cluster analysis tool to identify the analogous mutation clusters between cancer and germline disease mutations. Such a tool would be a valuable in addition to the existing bioinformatics tools for identifying functional mutations.

Analysis and significance of non-coding functional variants

We have thus far focused exclusively on reviewing the analysis of non-synonymous coding mutations, but there are several ways in which single-nucleotide mutations may give rise to abnormal function and therefore potentially lead to a disease phenotype. In addition to amino acid substitutions, single-nucleotide changes may also result in irregular gene expression through modification of regions of the genome important for regulation of transcription, such as transcription factor-binding sites (Knight 2005; Pastinen and Hudson 2004). In addition, most human genes require post-transcriptional processing before resulting in a mature mRNA, so modifications in splice sites or polyadenylation may also lead to altered protein function. Finally, mutations in any of the many regulatory RNAs that the genome encodes could result in an undesired phenotype through abnormal regulation of gene expression. Here, we will review experimental evidence demonstrating that mutations in these non-translated regions can have an effect on human health and give an overview of methods and resources that can be applied towards large-scale functional characterization of these non-coding features.

Regulatory SNPs

Gene expression is a tightly regulated cellular process and so mutations that affect gene expression can have a profound phenotypic effect or dramatically increase disease risk. Depending on the gene in question, variations that increase or decrease the expression of a gene can both have deleterious effects. One example of overexpression increasing disease susceptibility is in the MDM2 gene where a single nucleotide change known as SNP309 alters transcription factor binding (Bond et al. 2004). Individuals with a T=>G mutation show a substantially increased risk for developing colorectal cancer. MDM2 acts as an inhibitor of the p53 tumor suppressor pathway, and the evidence suggests that a guanine at SNP309 results in overexpression of MDM2, which in turn leads to increased suppression of the p53 pathway and increased risk of cancer development (Bond et al. 2005; Bond and Levine 2007). Another regulatory SNP of interest to cancer researchers is the 938C/A polymorphism present in the promoter region of the BCL-2 anti-apoptosis gene. The alanine variant of this polymorphism is thought to reduce the expression of BCL-2 relative to wild-type expression, which would in turn provide low risk of cancer development. In fact, studies have shown that BCL-2 938A is associated with decreased risk in prostate cancer and squamous cell carcinoma (Chen et al. 2007; Kidd et al. 2006). However, an additional study with a small patient set could not find association between protein levels of BCL-2 and any laboratory or clinical features of chronic lymphocytic leukemia (Majid et al. 2008). In one further example that is distinct from the MDM2 and BCL-2 SNPs above, rs6983267 is an SNP in an intergenic region of 8q24 which is hundreds of kilobases away from the nearest functional gene. Multiple independent studies have shown that the guanine allele at this position is associated with several cancers, most prominently colorectal cancer (Haiman et al. 2007; Tomlinson et al. 2007; Tuupanen et al. 2008; Zanke et al. 2007). In this case, it is unclear whether rs6983267 is functional by itself through disruption of a

long-range enhancer element or if it is tightly linked to another functional variant.

In a manner analogous to GWAS to discover SNPs correlated with complex traits, several studies have attempted large-scale efforts to associate SNPs with gene expression phenotypes (Cheung et al. 2003; Cheung and Spielman 2002; Cheung et al. 2005; Morley et al. 2004; Spielman et al. 2007). Rather than thinking of a disease condition as a phenotype, these studies utilize specific gene expression levels as the phenotype of interest and they discover a large number of genetic markers that are tightly associated with expression phenotypes. The results of these studies effectively comprise a list of candidate regulatory SNPs. Since with more traditional GWAS, it is not apparent without additional experimentation which of the associated SNPs are actually causal, as opposed to simply correlated. Even with the identification of many associated SNPs, it is likely that experiments are needed to filter for the genotypes that directly contribute to the gene expression phenotype. For example, although association of rs6983267 to cancer has been found, it would require detailed experimentation to resolve whether or not the SNP is causal, and if so whether it is tissue-specific. Such experiments are often performed with reporter assays where promoters containing the candidate alleles are used to drive expression of a reporter gene such as Luciferase (Cheung et al. 2005; Ogasawara et al. 2008), so that allele-specific expression can be quantified. Even so, however, identifying mutations that alter expression of key genes may still not result in discovering the causal factor in disease since it would have to be shown that the modified expression results in the observed phenotype. In cases such as rs6983267, the causal factor ends up being extremely difficult to detect since it is not near any gene and so even if it is affecting expression, identification of the genes that it is directly affecting can be a difficult problem.

In one case where significant experimental evidence has provided a strong theory for a regulatory SNP being responsible for driving disease, De Gobbi et al. (2006) characterized an SNP associated with the blood disorder α-thalassemia. This SNP does not alter protein function directly, but instead it lies in an intergenic region within the α-globin gene cluster, which has been associated with α-thalassemia onset. The disease-associated variant is in fact a gain-of-function mutation that results in a new transcriptional promoter for the GATA-1 transcription factor being created in the midst of the gene cluster. Activation of transcription at this new promoter appears to result in suppressed expression of downstream α-globin genes, which leads to α-thalassemia (De Gobbi et al. 2006).

A number of computational tools have been developed to help with the analysis of SNPs (Mooney 2005), and many of them have specific features towards the identification of regulatory SNPs. A recent review on methods for annotating SNPs provides a detailed description of many of these resources (Karchin 2009). With regards to functional analysis of regulatory SNPs, the general paradigm is to utilize existing databases of transcription factor-binding sites, such as TRANSFAC and JASPAR to identify SNPs that may map to important regulatory regions (Matys et al. 2003; Sandelin et al. 2004). Transcription factor-binding site mapping is primarily accomplished through identification of genomic elements that match known sequence motifs or positional weight matrices, so there can be a quantitative measure of how close a given sequence is to the canonical binding site. A method of predicting whether or not a nucleotide change will have an effect on gene expression could be to score both the wild-type and the variant sequences for transcription factor binding and look for differences (GuhaThakurta et al. 2006). Any perturbation of a binding site could result in a functional effect, since reducing transcription factor binding will cause deregulation of the target gene, whereas introducing a new site could cause undesired transcription (De Gobbi et al. 2006). For example, JASPAR sequence analysis of two breast-cancer susceptibility SNPs (rs7895676 and rs2981578) in the FGFR2 locus shows that they are likely to affect transcription factor binding. In each case, JASPAR scores one allele with a high similarity to the known transcription factor-binding site whereas the other receives a score below the default 0.80 relative profile score threshold. Detailed experimental evidence confirms that the minor allele C in rs7895676 disrupts binding of C/EBPβ while the minor allele G in rs2981578 increases binding affinity of Runx2 (Meyer et al. 2008).

Post-transcriptional processing SNPs

Single-nucleotide mutations can affect the cell in ways even beyond amino acid substitution and transcriptional regulation. There are several cases of SNPs affecting protein function through alterations in splicing rather than affecting protein structure through amino acid substitution (Pagani and Baralle 2004; Srebrow and Kornblihtt 2006). Furthermore, gene expression can be regulated through means such as micro RNAs (miRNAs) or polyadenylation rather than through genomic regulatory regions. These factors can also be influenced through mutation and result in abnormal phenotypes.

Splicing mechanisms in humans are relatively well known and with the genome tools available today, it has become possible to systematically identify genomic sites important in splicing, which in turn allows the identification of variants that may affect splicing. Several large-scale efforts have attempted to identify SNPs that may have an effect on splicing (ElSharawy et al. 2006; Hull et al. 2007; Nembaware et al. 2008) and a number of tools exist for

identification of splicing-related genomic elements (Cartegni et al. 2003; Thierry-Mieg and Thierry-Mieg 2006). There are multiple accounts of splicing SNPs with an impact in cancer risk, with many of these being identified in the breast-cancer susceptibility gene BRCA1 (Mazoyer et al. 1998; Pettigrew et al. 2005). Splicing mutations and polymorphisms can affect phenotype in several ways. Modification of a splicing donor or acceptor site could affect splicing efficiency, which can result in unwanted constitutive splicing or reduced splicing. Altering splicing enhancers or silencers could have a similar affect of affecting splicing efficiency. Alternative splicing is a well-known phenomenon that could also be affected by genetic variation. The large amount of transcript data available now suggests that alternative splicing is an extremely prevalent process in human, and misregulation of this process could produce undesirable phenotypes (Blencowe 2006). Several splicing-related mutations have been identified in *cis*-acting sequences that affect cancer-related genes and drive cancer formation. Li-Fraumeni syndrome and Peutz–Jeghers syndrome are hereditary genetic disorders that substantially increase risk and both have been linked to splicing-related polymorphisms (Hastings et al. 2005; Warneford et al. 1992).

In a non-cancer example related to post-transcriptional regulation, Uitte de Willige et al. (2007) found that a single SNP in the alternatively spliced fibrinogen gamma (FGG) gene leads to increased risk for deep-venous thrombosis. Linkage studies found that a particular haplotype of FGG was associated with increased disease risk and reduced protein levels. The C10034T SNP was discovered to be primarily responsible for this phenotype, and sequence evidence suggested that this SNP could be disrupting the normal polyadenylation signal and in fact increasing polyadenylation of one isoform relative to another and disrupting the ratios of protein production. This phenomenon is similar to rSNP variations resulting in misregulated gene expression except that it occurs at the post-transcriptional level, demonstrating that disruption of proper protein production at any stage could lead to deleterious effects.

In another example, where disruption of post-transcriptional gene expression regulation results in disease risk, there have been numerous studies linking mutations in the Hmga2 gene to cancer risk (Fedele et al. 2001; Lee and Dutta 2007; Mayr et al. 2007). Many of these mutations are a result of a truncation in the open-reading frame (ORF) of the Hmga2 gene, but a subset of these do not disrupt the ORF but instead only truncate the 3′ untranslated region (UTR). The Hmga2 3′ UTR contains several conserved binding sites for the let-7 miRNA and experimental evidence suggests that let-7 is involved in post-transcriptional repression of Hmga2 production. Further studies confirmed that, indeed, truncation of the Hmga2 3′ UTR was responsible for loss of let-7 repression, which in turn leads to oncogenesis (Lee and Dutta 2007; Mayr et al. 2007). In a manner similar to Hmga2/let-7, there is evidence that miRNA SNPs can also be involved in altering drug response. For example, the C829T SNP in the 3′ UTR of dihydrofolate reductase appears to affect miRNA-dependent regulation of DHFR expression. DHFR is the target of the commonly used chemotherapeutic agent methotrexate and studies have shown that SNP C829T causes loss of miR24 miRNA binding and results in DHFR overexpression which in turn drives resistance to methotrexate (Mishra et al. 2007).

There are many tools available for predicting the potential functional impact of SNPs, whether they affect the coding sequence of a protein, the sequences regulating the expression of a gene, or other aspects of protein expression. In addition, there are several databases available that catalog known SNPs and provide tools for selecting SNPs under specified criteria. A number of resources have been developed specifically to assist in analysis of potential regulatory SNPs (Table 3). For instance, rSNP_Guide (Ponomarenko et al. 2002, 2003) and SNP@Promoter (Kim et al. 2008) both store known SNPs, but specifically attempt to associate them with known transcription factor-binding sites for identification of potential regulatory SNPs. MAPPER is a companion tool to the SNPper retrieval system and database that locates computationally predicted transcription factor-binding sites (Riva and Kohane 2002). Both PolyMAPr (Freimuth et al. 2005) and PupaSNP Finder (Conde et al. 2006) use computational methods to find possible exon splicing enhancer sites that can then be mapped to SNP locations to find SNPs that potentially affect splicing. Other than simply attempting to map SNPs to potential transcription factor-binding sites or promoter regions to find SNPs that may have regulatory significance, the tools will also use functional information provided by databases, such as HGMD (Stenson et al. 2008) or OMIM (Amberger et al. 2009) to attempt to provide a functional annotation to some SNPs. Figure 1 shows a schematic of the various sorts of functional mutations that are possible and what tools are available for analysis of each type of functional region. The greatest number of methods exists for analysis of non-synonymous coding mutations, but there are tools available that attempt to identify each of the non-coding functional regions, such as exonic splicing enhancers (Cartegni et al. 2003), splice junctions (Stamm et al. 2006; Thierry-Mieg and Thierry-Mieg 2006), polyadenylation sites (Lambert et al. 2004; Tabaska and Zhang 1999) (http://www.imtech.res.in/raghava/polyapred/), transcription enhancers (Hallikas et al. 2006; Palin et al. 2006), micro-RNA-binding sites (Enright et al. 2003; Griffiths-Jones et al. 2008; Wang 2008), and transcription factor-binding sites (Matys et al. 2003; Sandelin et al. 2004).

Table 3 List of resources available for high-throughput SNP annotation and selection

Resource	Transcription factor-binding sites	Splicing	Annotation	SNP sources
rSNP_Guide (Ponomarenko et al. 2002, 2003)	Custom method based on experimentally verified sites from TRANSFAC, TRRD, COMPEL, ACTIVITY	No	HGMD, HGBASE, ALFRED, OMIM	dbSNP
SNPper and MAPPER (Riva and Kohane 2002)	TRANSFAC, JASPAR	No	GoldenPath, LocusLink, GO, Swiss-Prot	dbSNP
PolyMAPr (Freimuth et al. 2005)	JASPAR	Splice junctions (ASD) and exon splicing enhancer sites (ESEfinder)	GoldenPath, PolyPhen	dbSNP, CGAP, JSNP
PupaSuite (Conde et al. 2006)	TRANSFAC, JASPAR	Exon splicing enhancer sites (ESEfinder)	Ensembl, GO, OMIM	dbSNP
SNP@Promoter (Kim et al. 2008)	TRANSFAC	No	KEGG, GO, GAD, HGMD, OMIM	dbSNP
SNPnexus (Chelala et al. 2008)	TRANSFAC, FirstEF	AceView, PupaSuite	Ensembl, GAD, VEGA, AceView	dbSNP
SNPLogic (Pico et al. 2009)	Delta-MATCH, PupaSuite	PupaSuite	PolyPhen, SNP3D, OMIM, GO, KEGG, WikiPathways, BioCart, BioCyc	dbSNP

Furthermore, a recent study has shown that some regions of the genome have evolutionarily conserved three-dimensional DNA structures that correlate with non-coding functional genomic regions, thereby providing another method for identification of important substitutions (Parker et al. 2009). In each of these cases, even if there is not a tool that specifically predicts the potential functional impact of a sequence alteration, simply examining the difference in scores between the wild-type and mutated sequences is a method that can be universally applied. These algorithms could then provide a comprehensive means of functional prediction on all kinds of genetic variations.

Although the examples of functional non-coding SNPs presented above are all naturally occurring polymorphisms, it is likely that somatically gained mutations would have similar functional effects. The focus of most large-scale cancer sequencing projects to date has been on finding mutations in coding regions, with the majority of sequencing projects focusing on transcript sequence. Because of this data-generation bias, the amount of sequence data available for non-coding regions of the human genome is substantially smaller than for coding regions, but the extensive evidence from SNP data presented above implies that it is likely that some cancers may at least be partially driven by regulatory, splicing, or miRNA mutations. With the release of the first completely sequenced cancer genome and other cancer sequencing projects with coverage beyond that of just the coding regions (Ley et al. 2008), the data are becoming available to fully explore the extent of non-coding mutations in cancer.

Conclusion

The field of cancer mutation research is speeding up dramatically as the rate of data generation increases for advancements in sequencing technology. With the success of targeted cancer therapeutics, it appears that there is a significant benefit to be gained from continued efforts to identify genes and mutations that drive cancer development. The recent cancer genome sequencing projects are just the beginning of what will likely to be a continued flood of mutation data as next generation sequencing technologies continue to increase throughput and decrease cost, enabling the examination of both more regions of the genome as well as more samples. The combination of these factors emphasizes the need for robust computational pipelines for analysis of mutation data. Frequency-based methods are best equipped to leverage the statistical benefits of large datasets, but they may be subject to some weaknesses that amino acid-change-based methods can mitigate. Methods targeted to specific types of proteins (such as kinases and their targets) have also shown that they can be more effective

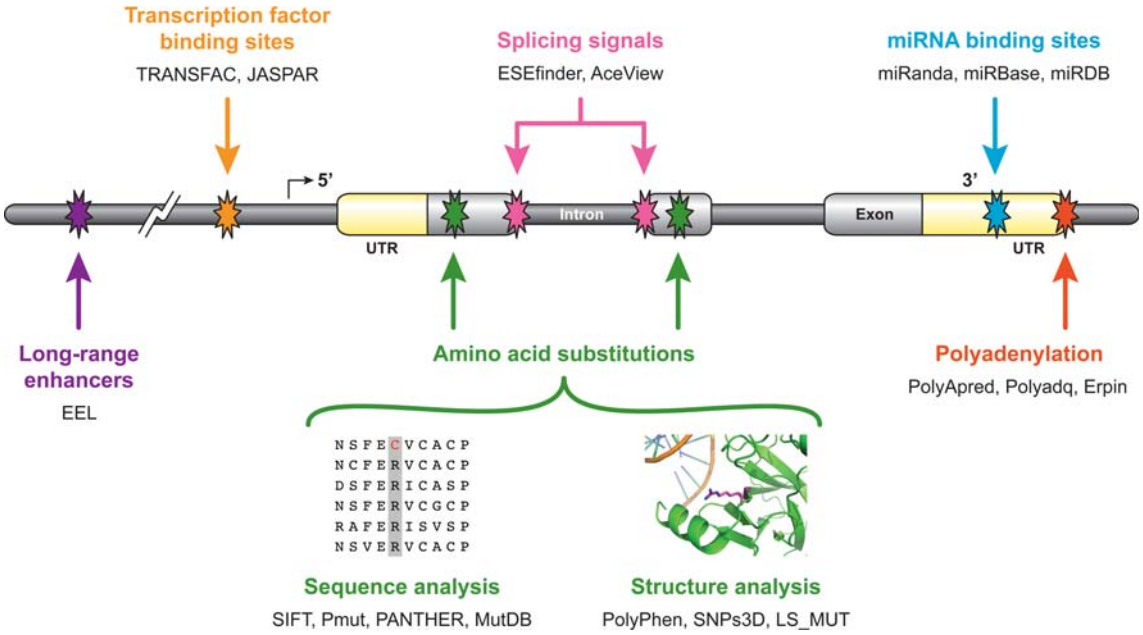

Fig. 1 Genomic regions, which are subject to functional alteration through single-nucleotide substitutions. Select computational tools that could be used for mapping or analysis of the various kinds of sequence elements are listed under each category. Methods for analysis of amino acid substitutions are roughly separated into those that incorporate protein structure information or those that are purely sequence based

than generalized tools due to their ability to incorporate more specific models and reduce noise through prior knowledge (Radivojac et al. 2008; Torkamani and Schork 2008). Integration of other datasets, such as genome-wide expression and copy number analysis will also be crucial in providing the best candidates for focused analysis (TCGA 2008).

Many of the issues resulting from small sample size or a candidate gene approach that were present in early studies will be mitigated through the decreasing cost of sequencing. However, the rapidly approaching dream of inexpensive complete genome sequencing will also bring about a new set of analysis challenges. Current cancer genome sequencing projects are able to constrain their analysis by focusing on protein-coding regions and non-synonymous-coding mutations since that is the majority of generated data. Next generation sequencing efforts will likely include the copious amounts of non-coding genomic sequence present in the human genome which have not yet been examined by most existing sequencing efforts. The results from GWAS and small-scale experiments have already demonstrated that non-coding mutations can have a significant impact on cellular function, but novel analysis methods are needed to leverage this new data.

The exponentially increasing ability to sequence has enabled experiments that were previously prohibitively expensive. This new technology leads to an exciting time in the field of cancer mutation research, but puts the burden on the computational tools to provide the greatest value from the generated data. Next generation tools will have to be both accurate and fast to process the large amounts of incoming data, and it will require multilateral efforts to fully mine each dataset.

Acknowledgments The authors would like to acknowledge Allison Bruce for design and production of graphics used in this manuscript.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37:6–D793

Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, Sullivan K, Matakidou A, Wang Y, Mills G, Doheny K, Tsai YY, Chen WV, Shete S, Spitz MR, Houlston RS (2008) Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 40:616–622

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology: The Gene Ontology Consortium. Nat Genet 25:25–29

Atchley WR, Zhao J, Fernandes AD, Druke T (2005) Solving the protein sequence metric problem. Proc Natl Acad Sci USA 102:6395–6400

Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, Markowitz S, Willson JK, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE (2003) Mutational analysis of the tyrosine kinome in colorectal cancers. Science 300:949

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B Methodol 57:289–300

Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126:37–47

Bond GL, Levine AJ (2007) A single nucleotide polymorphism in the p53 pathway interacts with gender, environmental stresses and tumor genetics to influence cancer in humans. Oncogene 26:1317–1323

Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, Arva NC, Bargonetti J, Bartel F, Taubert H, Wuerl P, Onel K, Yip L, Hwang SJ, Strong LC, Lozano G, Levine AJ (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119:591–602

Bond GL, Hu W, Levine A (2005) A single nucleotide polymorphism in the MDM2 gene: from a molecular and cellular explanation to clinical effect. Cancer Res 65:5481–5484

Breiman L (2001) Random forests. Mach Learn 45:5–32

Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, Rowan A, Lubbe S, Spain S, Sullivan K, Fielding S, Jaeger E, Vijayakrishnan J, Kemp Z, Gorman M, Chandler I, Papaemmanuil E, Penegar S, Wood W, Sellick G, Qureshi M, Teixeira A, Domingo E, Barclay E, Martin L, Sieber O, Kerr D, Gray R, Peto J, Cazier JB, Tomlinson I, Houlston RS (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 39:1315–1317

Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acids Res 31:3568–3571

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS (2007) Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat 28:683–693

Chasman D, Adams RM (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706

Chelala C, Khan A, Lemoine NR (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. Bioinformatics 25:655–661

Chen K, Hu Z, Wang LE, Sturgis EM, El-Naggar AK, Zhang W, Wei Q (2007) Single-nucleotide polymorphisms at the TP53-binding or responsive promoter regions of BAX and BCL2 genes and risk of squamous cell carcinoma of the head and neck. Carcinogenesis 28:2008–2012

Cheung VG, Spielman RS (2002) The genetics of variation in gene expression. Nat Genet 32 Suppl:522–525

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33:422–425

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365–1369

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA (2008) Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659:147–157

Ciardiello F, Caputo R, Bianco R, Damiano V, Pomatico G, De Placido S, Bianco AR, Tortora G (2000) Antitumor effect and potentiation of cytotoxic drugs activity in human cancer cells by ZD-1839 (Iressa), an epidermal growth factor receptor-selective tyrosine kinase inhibitor. Clin Cancer Res 6:2053–2063

Clifford RJ, Edmonson MN, Nguyen C, Buetow KH (2004) Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics 20:1006–1014

Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 34:W621–W625

Davies H, Hunter C, Smith R, Stephens P, Greenman C, Bignell G, Teague J, Butler A, Edkins S, Stevens C, Parker A, O'Meara S, Avis T, Barthorpe S, Brackenbury L, Buck G, Clements J, Cole J, Dicks E, Edwards K, Forbes S, Gorton M, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jones D, Kosmidou V, Laman R, Lugg R, Menzies A, Perry J, Petty R, Raine K, Shepherd R, Small A, Solomon H, Stephens Y, Tofts C, Varian J, Webb A, West S, Widaa S, Yates A, Brasseur F, Cooper CS, Flanagan AM, Green A, Knowles M, Leung SY, Looijenga LH, Malkowicz B, Pierotti MA, Teh BT, Yuen ST, Lakhani SR, Easton DF, Weber BL, Goldstraw P, Nicholson AG, Wooster R, Stratton MR, Futreal PA (2005) Somatic mutations of the protein kinase gene family in human lung cancer. Cancer Res 65:7591–7595

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312:1215–1217

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, Metcalf GA, Ng B, Milosavljevic A, Gonzalez-Garay ML, Osborne JR, Meyer R, Shi X, Tang Y, Koboldt DC, Lin L, Abbott R, Miner TL, Pohl C, Fewell G, Haipek C, Schmidt H, Dunford-Shore BH, Kraja A, Crosby SD, Sawyer CS, Vickery T, Sander S, Robinson J, Winckler W, Baldwin J, Chirieac LR, Dutt A, Fennell T, Hanna M, Johnson BE, Onofrio RC, Thomas RK, Tonon G, Weir BA, Zhao X, Ziaugra L, Zody MC, Giordano T, Orringer MB, Roth JA, Spitz MR, Wistuba II, Ozenberger B, Good PJ, Chang AC, Beer DG, Watson MA, Ladanyi M, Broderick S, Yoshizawa A, Travis WD, Pao W, Province MA, Weinstock GM, Varmus HE, Gabriel SB, Lander ES, Gibbs RA, Meyerson M, Wilson RK (2008) Somatic mutations affect key pathways in lung adenocarcinoma. Nature 455:1069–1075

Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med 344:1031–1037

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A,

Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093

Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaley DP, Muir KR, Smith C, Bagnato M, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF (2008) Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 40:316–321

ElSharawy A, Manaster C, Teuber M, Rosenstiel P, Kwiatkowski R, Huse K, Platzer M, Becker A, Nurnberg P, Schreiber S, Hampe J (2006) SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs. Hum Mutat 27:1129–1134

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS (2003) MicroRNA targets in *Drosophila*. Genome Biol 5:R1

Fedele M, Battista S, Manfioletti G, Croce CM, Giancotti V, Fusco A (2001) Role of the high mobility group A proteins in human lipomas. Carcinogenesis 22:1583–1591

Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34:D247–D251

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet Chapter 10: Unit 10 11

Forrest WF, Cavet G (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 317:1500 author reply 1500

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Freimuth RR, Stormo GD, McLeod HL (2005) PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat 25:110–117

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. Nat Rev Cancer 4:177–183

Getz G, Hofling H, Mesirov JP, Golub TR, Meyerson M, Tibshirani R, Lander ES (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 317:1500

Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, Kosarin K, Olsh A, Bergeron J, Ellis NA, Klein RJ, Clark AG, Norton L, Dean M, Boyd J, Offit K (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci USA 105:4340–4345

Greenman C, Wooster R, Futreal PA, Stratton M, Easton D (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics 173:2187–2198

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR (2007) Patterns of somatic mutation in human cancer genomes. Nature 446:153–158

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36:D154–D158

Gu J, Gribskov M, Bourne PE (2006) Wiggle-predicting functionally flexible regions from primary sequence. PLoS Comput Biol 2:e90

Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, Partin AW, Albers-Akkers MT, Godino-Ivan Marcos J, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. Nat Genet 39:631–637

GuhaThakurta D, Xie T, Anand M, Edwards SW, Li G, Wang SS, Schadt EE (2006) *Cis*-regulatory variations: a study of SNPs around genes showing *cis*-linkage in segregating mouse populations. BMC Genomics 7:235

Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X, Kolonel LN, Wu AH, Reich D, Henderson BE (2007) A common genetic risk factor for colorectal and prostate cancer. Nat Genet 39:954–956

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124:47–59

Hastings ML, Resta N, Traum D, Stella A, Guanti G, Krainer AR (2005) An LKB1 AT-AC intron mutation causes Peutz–Jeghers syndrome via splicing at noncanonical cryptic splice sites. Nat Struct Mol Biol 12:54–59

Henikoff S, Comai L (2003) Single-nucleotide mutations for plant functional genomics. Annu Rev Plant Biol 54:375–401

Hon LS, Kaminker JS, Zhang ZM (2008) Computational approaches for predicting causal missense mutations in cancer genome projects. Curr Bioinform 3:46–55

Hon LS, Zhang Y, Kaminker JS, Zhang Z (2009) Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. Hum Mutat 30:99–106

Hull J, Campino S, Rowlands K, Chan MS, Copley RR, Taylor MS, Rockett K, Elvidge G, Keating B, Knight J, Kwiatkowski D (2007) Identification of common genetic variation that modulates alternative splicing. PLoS Genet 3:e99

Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39:870–874

Jiang R, Yang H, Zhou L, Kuo CC, Sun F, Chen T (2007) Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations. Am J Hum Genet 81:346–360

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321:1801–1806

Kaminker JS, Zhang Y, Watanabe C, Zhang Z (2007a) CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 35:W595–W598

Kaminker JS, Zhang Y, Watanabe C, Zhang Z (2007a) CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res 35:W595–W598

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820

Karchin R (2009) Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52

Kidd LR, Coulibaly A, Templeton TM, Chen W, Long LO, Mason T, Bonilla C, Akereyeni F, Freeman V, Isaacs W, Ahaghotu C, Kittles RA (2006) Germline BCL-2 sequence variants and inherited predisposition to prostate cancer. Prostate Cancer Prostatic Dis 9:284–292

Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J (2008) SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinform 9(Suppl 1):S2

Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD (2008) Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov 7:221–230

Knight JC (2005) Regulatory polymorphisms underlying complex disease traits. J Mol Med 83:97–109

Knudson AG Jr (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 68:820–823

Kobayashi S, Boggon TJ, Dayaram T, Janne PA, Kocher O, Meyerson M, Johnson BE, Eck MJ, Tenen DG, Halmos B (2005) EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 352:786–792

Lambert A, Fontaine JF, Legendre M, Leclerc F, Permal E, Major F, Putzer H, Delfour O, Michot B, Gautheret D (2004) The ERPIN server: an interface to profile-based RNA motif identification. Nucleic Acids Res 32:W160–W165

Lee YS, Dutta A (2007) The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. Genes Dev 21:1025–1030

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456:66–72

Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, Rieder MJ, Gowrisankar S, Aronow BJ, Weiss RB, Nickerson DA (2004) Pattern of sequence variation across 213 environmental response genes. Genome Res 14:1821–1831

Lugo TG, Pendergast AM, Muller AJ, Witte ON (1990) Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. Science 247:1079–1082

Majid A, Tsoulakis O, Walewska R, Gesk S, Siebert R, Kennedy DB, Dyer MJ (2008) BCL2 expression in chronic lymphocytic leukemia: lack of association with the BCL2 938A>C promoter single nucleotide polymorphism. Blood 111:874–877

Marks JL, McLellan MD, Zakowski MF, Lash AE, Kasai Y, Broderick S, Sarkaria IS, Pham D, Singh B, Miner TL, Fewell GA, Fulton LL, Mardis ER, Wilson RK, Kris MG, Rusch VW, Varmus H, Pao W (2007) Mutational analysis of EGFR and related signaling pathway genes in lung Adenocarcinomas identifies a novel somatic kinase domain mutation in FGFR4. PLoS ONE 2:e426

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31:374–378

Mayr C, Hemann MT, Bartel DP (2007) Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. Science 315:1576–1579

Mazoyer S, Puget N, Perrin-Vidoz L, Lynch HT, Serova-Sinilnikova OM, Lenoir GM (1998) A BRCA1 nonsense mutation causes exon skipping. Am J Hum Genet 62:713–715

Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, Ponder BA (2008) Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. PLoS Biol 6:e108

Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328

Mishra PJ, Humeniuk R, Mishra PJ, Longo-Sorbello GS, Banerjee D, Bertino JR (2007) A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. Proc Natl Acad Sci USA 104:13513–13518

Mooney S (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6:44–56

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747

Nembaware V, Lupindo B, Schouest K, Spillane C, Scheffler K, Seoighe C (2008) Genome-wide survey of allele-specific splicing in humans. BMC Genomics 9:265

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874

Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12:436–446

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

Ng PC, Henikoff S (2006) Predicting the effects of amino Acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80

Ogasawara K, Terada T, Motohashi H, Asaka JI, Aoki M, Katsura T, Kamba T, Ogawa O, Inui KI (2008) Analysis of regulatory polymorphisms in organic ion transporter genes (SLC22A) in the kidney. J Hum Genet 53:607–614

Pagani F, Baralle FE (2004) Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5:389–396

Palin K, Taipale J, Ukkonen E (2006) Locating potential enhancer elements by comparative genomics using the EEL software. Nat Protoc 1:368–374

Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324:389–392

Parsons DW, Wang TL, Samuels Y, Bardelli A, Cummins JM, DeLong L, Silliman N, Ptak J, Szabo S, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Lengauer C, Velculescu VE (2005) Colorectal cancer: mutations in a signalling pathway. Nature 436:792

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, Rasheed BA, Keir S, Nikolskaya T, Nikolsky Y, Busam DA, Tekleab H, Diaz LA Jr, Hartigan J, Smith DR, Strausberg RL, Marie SK, Shinjo SM, Yan H, Riggins GJ, Bigner DD, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW (2008) An integrated genomic analysis of human glioblastoma multiforme. Science 321:1807–1812

Pastinen T, Hudson TJ (2004) *Cis*-acting regulatory variation in the human genome. Science 306:647–650

Pettigrew C, Wayte N, Lovelock PK, Tavtigian SV, Chenevix-Trench G, Spurdle AB, Brown MA (2005) Evolutionary conservation analysis increases the colocalization of predicted exonic splicing enhancers in the BRCA1 gene with missense sequence changes and in-frame deletions, but not polymorphisms. Breast Cancer Res 7:R929–R939

Pico AR, Smirnov IV, Chang JS, Yeh RF, Wiemels JL, Wiencke JK, Tihan T, Conklin BR, Wrensch M (2009) SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. Nucleic Acids Res 37:D803–D809

Ponomarenko JV, Orlova GV, Merkulova TI, Gorshkova EV, Fokin ON, Vasiliev GV, Frolov AS, Ponomarenko MP (2002) rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites. Hum Mutat 20:239–248

Ponomarenko JV, Merkulova TI, Orlova GV, Fokin ON, Gorshkova EV, Frolov AS, Valuev VP, Ponomarenko MP (2003) rSNP_Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation. Nucleic Acids Res 31:118–121

Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, Mooney SD (2008) Gain and loss of phosphorylation sites in human cancer. Bioinformatics 24:i241–i247

Riva A, Kohane IS (2002) SNPper: retrieval and analysis of human SNPs. Bioinformatics 18:1681–1685

Rubin AF, Green P (2007) Comment on "The consensus coding sequences of human breast and colorectal cancers". Science 317:1500

Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell SM, Riggins GJ, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE (2004) High frequency of mutations of the PIK3CA gene in human cancers. Science 304:554

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32:D91–D94

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314:268–274

Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39:226–231

Srebrow A, Kornblihtt AR (2006) The connection between splicing and cancer. J Cell Sci 119:2635–2641

Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA (2006) ASD: a bioinformatics resource on alternative splicing. Nucleic Acids Res 34:D46–D55

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN (2008) Human gene mutation database: towards a comprehensive central mutation database. J Med Genet 45:124–126

Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, O'Meara S, Parker A, Tarpey P, Avis T, Barthorpe A, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Edwards K, Forbes S, Gorton M, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jones D, Kosmidou V, Laman R, Lugg R, Menzies A, Perry J, Petty R, Raine K, Shepherd R, Small A, Solomon H, Stephens Y, Tofts C, Varian J, Webb A, West S, Widaa S, Yates A, Brasseur F, Cooper CS, Flanagan AM, Green A, Knowles M, Leung SY, Looijenga LH, Malkowicz B, Pierotti MA, Teh B, Yuen ST, Nicholson AG, Lakhani S, Easton DF, Weber BL, Stratton MR, Futreal PA, Wooster R (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. Nat Genet 37:590–592

Steward RE, MacArthur MW, Laskowski RA, Thornton JM (2003) Molecular basis of inherited diseases: a structural perspective. Trends Genet 19:505–513

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32:D520–D522

Sunyaev S, Ramensky V, Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16:198–200

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P (2001) Prediction of deleterious human alleles. Hum Mol Genet 10:591–597

Tabaska JE, Zhang MQ (1999) Detection of polyadenylation signals in human DNA sequences. Gene 231:77–86

TCGA (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455:1061–1068

Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, Haq N, Barnetson RA, Theodoratou E, Cetnarskyj R, Cartwright N, Semple C, Clark AJ, Reid FJ, Smith LA, Kavoussanakis K, Koessler T, Pharoah PD, Buch S, Schafmayer C, Tepel J, Schreiber S, Volzke H, Schmidt CO, Hampe J, Chang-Claude J, Hoffmeister M, Brenner H, Wilkening S, Canzian F, Capella G, Moreno V, Deary IJ, Starr JM, Tomlinson IP, Kemp Z, Howarth K, Carvajal-Carmona L, Webb E, Broderick P, Vijayakrishnan J, Houlston RS, Rennert G, Ballinger D, Rozek L, Gruber SB, Matsuda K, Kidokoro T, Nakamura Y, Zanke BW, Greenwood CM, Rangrej J, Kustra R, Montpetit A, Hudson TJ, Gallinger S, Campbell H, Dunlop MG (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 40:631–637

Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 7(Suppl 1):S12 1–S1214

Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci USA 101:15398–15403

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141

Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats BJ, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt SI, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cussenot O, Valeri A, Andriole GL, Crawford ED, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hayes RB, Hunter DJ, Chanock SJ (2008) Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 40:310–315

Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, Spain S, Penegar S, Chandler I, Gorman M, Wood W, Barclay E, Lubbe S, Martin L, Sellick G, Jaeger E, Hubner R, Wild R, Rowan A, Fielding S, Howarth K, Silver A, Atkin W, Muir K, Logan R, Kerr D, Johnstone E, Sieber O, Gray R, Thomas H, Peto J, Cazier JB, Houlston R (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 39:984–988

Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, Pittman AM, Spain S, Lubbe S, Walther A, Sullivan K, Jaeger E, Fielding S, Rowan A, Vijayakrishnan J, Domingo E, Chandler I, Kemp Z, Qureshi M, Farrington SM, Tenesa A, Prendergast JG, Barnetson RA, Penegar S, Barclay E, Wood W, Martin L, Gorman M, Thomas H, Peto J, Bishop DT, Gray R, Maher ER, Lucassen A, Kerr D, Evans DG, Schafmayer C, Buch S, Volzke H, Hampe J, Schreiber S, John U, Koessler T, Pharoah P, van Wezel T, Morreau H, Wijnen JT, Hopper JL, Southey MC, Giles GG, Severi G, Castellvi-Bel S, Ruiz-Ponte C, Carracedo A, Castells A, Forsti A, Hemminki K, Vodicka P, Naccarati A, Lipton L, Ho JW, Cheng KK, Sham PC, Luk J, Agundez JA, Ladero JM, de la Hoya M, Caldes T, Niittymaki I, Tuupanen S, Karhu A, Aaltonen L, Cazier JB, Campbell H, Dunlop MG, Houlston RS (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 40:623–630

Torkamani A, Schork NJ (2007a) Accurate prediction of deleterious protein kinase polymorphisms. Bioinformatics 23:2918–2925

Torkamani A, Schork NJ (2007b) Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family. Genomics 90:49–58

Torkamani A, Schork NJ (2008) Prediction of cancer driver mutations in protein kinases. Cancer Res 68:1675–1682

Tuupanen S, Niittymaki I, Nousiainen K, Vanharanta S, Mecklin JP, Nuorva K, Jarvinen H, Hautaniemi S, Karhu A, Aaltonen LA (2008) Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution. Cancer Res 68:14–17

Uitte de Willige S, Rietveld IM, De Visser MC, Vos HL, Bertina RM (2007) Polymorphism 10034C>T is located in a region regulating polyadenylation of FGG transcripts and influences the fibrinogen gamma'/gammaA mRNA ratio. J Thromb Haemost 5:1243–1249

Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, Slamon DJ, Murphy M, Novotny WF, Burchmore M, Shak S, Stewart SJ, Press M (2002) Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol 20:719–726

Wang X (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. RNA 14:1012–1017

Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17:263–270

Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS, Parmigiani G, Yan H, Wang TL, Riggins G, Powell SM, Willson JK, Markowitz S, Kinzler KW, Vogelstein B, Velculescu VE (2004) Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science 304:1164–1166

Warneford SG, Witton LJ, Townsend ML, Rowe PB, Reddel RR, Dalla-Pozza L, Symonds G (1992) Germ-line splicing mutation of the p53 gene in a cancer-prone family. Cell Growth Differ 3:839–846

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE, Vogelstein B (2007) The genomic landscapes of human breast and colorectal cancers. Science 318:1108–1113

Yue P, Moult J (2006) Identification and analysis of deleterious human SNPs. J Mol Biol 356:1263–1274

Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473

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

Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, Farrington SM, Prendergast J, Olschwang S, Chiang T, Crowdy E, Ferretti V, Laflamme P, Sundararajan S, Roumy S, Olivier JF, Robidoux F, Sladek R, Montpetit A, Campbell P, Bezieau S, O'Shea AM, Zogopoulos G, Cotterchio M, Newcomb P, McLaughlin J, Younghusband B, Green R, Green J, Porteous ME, Campbell H, Blanche H, Sahbatou M, Tubacher E, Bonaiti-Pellie C, Buecher B, Riboli E, Kury S, Chanock SJ, Potter J, Thomas G, Gallinger S, Hudson TJ, Dunlop MG (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 39:989–994

Zhu Y, Hoffman A, Wu X, Zhang H, Zhang Y, Leaderer D, Zheng T (2008) Correlating observed odds ratios from lung cancer case-control studies to SNP functional scores predicted by bioinformatic tools. Mutat Res 639:80–88