Computational approaches to identify functional genetic variants in cancer genomes

Abel Gonzalez-Perez[1,23], Ville Mustonen[2,23], Boris Reva[3,23], Graham R S Ritchie[2,4,23], Pau Creixell[5], Rachel Karchin[6], Miguel Vazquez[7], J Lynn Fink[8], Karin S Kassahn[8], John V Pearson[8], Gary D Bader[9], Paul C Boutros[10–12], Lakshmi Muthuswamy[10,11], B F Francis Ouellette[10,13], Jüri Reimand[9], Rune Linding[5], Tatsuhiro Shibata[14], Alfonso Valencia[7,15], Adam Butler[2], Serge Dronov[2], Paul Flicek[4], Nick B Shannon[16], Hannah Carter[6], Li Ding[17,18], Chris Sander[3], Josh M Stuart[19,20], Lincoln D Stein[9,21] & Nuria Lopez-Bigas[1,22] for the International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group

The International Cancer Genome Consortium (ICGC) aims to catalog genomic abnormalities in tumors from 50 different cancer types. Genome sequencing reveals hundreds to thousands of somatic mutations in each tumor but only a minority of these drive tumor progression. We present the result of discussions within the ICGC on how to address the challenge of identifying mutations that contribute to oncogenesis, tumor maintenance or response to therapy, and recommend computational techniques to annotate somatic variants and predict their impact on cancer phenotype.

Large-scale sequencing of cancer genomes often reveals many thousands of somatic missense (amino acid–changing) mutations in proteins. However, not all cancer mutations provide a selective ('driving') advantage to cancer cells[1,2]. Many mutations are so-called 'passengers' because their impact on protein function is either minor or the affected protein is not important for tumor progression. The important practical problem is to determine which mutations are likely drivers. Although the carcinogenicity of a particular mutation depends on concurrent genomic alterations in the cell, one can considerably decrease the number of potential driver candidates by determining the functional impact of each mutation. Thus, a key challenge is to distinguish between functional and nonfunctional mutations, and by extension between those that contribute to tumorigenesis (drivers) and those that do not (passengers) (see **Box 1** for definitions).

Cancer has been likened to an evolutionary process by which tumor cells gain a fitness advantage over their neighboring cells[2]. The process creates cells with altered abilities such as the circumvention of apoptosis and senescence, deregulated cell division and failed responses to external cues such as contact-contact inhibition and ligand-mediated cell signaling[3,4]. Normal cells are reprogrammed by changes in the genome that are subsequently selected and

[1]Research Unit on Biomedical Informatics, University Pompeu Fabra, Barcelona, Spain. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [4]Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [5]Cellular Signal Integration Group, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. [6]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [7]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre, Madrid, Spain. [8]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane, Queensland, Australia. [9]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [10]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [11]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [12]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario, Canada. [13]Department of Cell and Systems Biology, University of Toronto, Toronto, Ontario, Canada. [14]Division of Cancer Genomics, National Cancer Center, Chuo-ku, Tokyo, Japan. [15]Spanish National Bioinformatics Institute, Madrid, Spain. [16]Cambridge Research Institute, Cambridge, UK. [17]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA. [18]Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, Missouri, USA. [19]Department of Biomolecular Engineering, University of California, Santa Cruz, California, USA. [20]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California, USA. [21]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [22]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [23]These authors contributed equally to this work. Correspondence should be addressed to N.L.-B. (nuria.lopez@upf.edu) or L.S. (lincoln.stein@oicr.on.ca).

RECEIVED 25 JANUARY; ACCEPTED 7 JUNE; PUBLISHED ONLINE 30 JULY 2013; DOI:10.1038/NMETH.2562

BOX 1 DEFINITIONS

We define a functional variant as a genomic variant that affects the molecular function of a protein (as a gain, loss or switch of function). A nonfunctional variant does not appreciably affect the molecular function of a protein. A driver variant confers a selective advantage to a particular tumor cell, whereas a passenger variant does not. It is important to distinguish between functional versus nonfunctional and driver versus passenger as they describe different concepts. For example, a substitution might dramatically affect the function of a protein without providing any selective advantage to the tumor (it is a functional passenger variant).

Nonsynonymous mutations are those that alter the amino acid sequence of a protein.

clonally expanded. In a similar manner to the way germline mutations can leave behind patterns indicative of negative or positive selection over millions of years, somatic mutations that engender increases in tumor fitness also can leave telltale signs in the protein sequence. The analysis of a given protein can thus reveal a pattern of alterations that recurrently result in its loss of function, as in classic tumor suppressors, such as *TP53*, *RB1* or *PTEN*[5].

Mutation events collected across several patient samples can also reveal signs of clustering in the peptide sequence or the three-dimensional protein structure that indicate a critical domain has been modulated. In the extreme case, the presence of the same amino acid change in the same position in different individuals can be a strong indicator of such gain-of-function or oncogenomic events, as is the case with the *KRAS*[6] or *BRAF*[7] oncogenes. Such patterns can be leveraged by informatics tools to predict whether a particular mutational event induces a selectable phenotype.

Here we review the computational analyses that are commonly carried out after the detection of somatic mutations across a cohort of cancer samples to identify likely functional mutations and likely driver mutations (**Fig. 1**). Our focus will be on single-nucleotide variants and small insertions and/or deletions (operationally defined here as variants shorter than 50 base pairs) that change the amino acid sequence or affect regulatory regions. The output of these analyses consists of prioritized lists of mutations, genes and pathways that may be analyzed in follow-up experiments to test their actual role in cancer.

We divide the process of identifying functional and driver variants into three independent, but related, approaches (**Fig. 1**). The first consists of mapping mutations to annotated functional genomic features, identifying their consequences and determining whether these mutations have been previously reported. The second uses computational methods to predict the nature and magnitude of the functional impact of mutation in particular elements (for example, proteins or regulatory regions). The third relies on statistical methods to find signs of positive selection across the cohort. In **Figure 1**, we list a subset of the computational tools used in each of the approaches. In the sections that follow, we review the rationale and tools for each approach and conclude by presenting some of the unsolved challenges and future perspectives in the field.

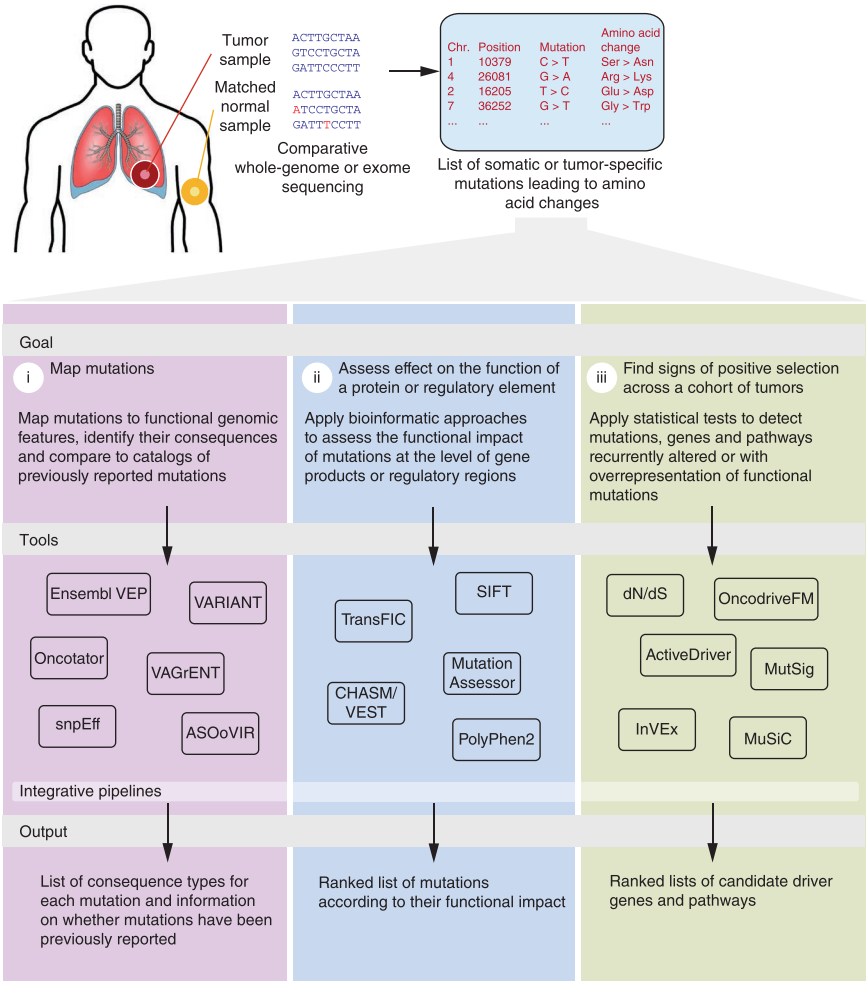

Figure 1 | Scheme depicting the three main approaches routinely used in the analysis of cancer somatic mutations. Although there are important relationships of precedence between elements from different approaches, they do not necessarily correspond to sequential steps. Tools used in each of the approaches are shown in the middle; those tools are defined in **Supplementary Tables 2–4**. Integrative pipelines refer to tools that facilitate the use of methods across all approaches (for example, IntOGen-mutations pipeline).

Approach 1: mutation mapping and annotation

The first step in determining the possible functional consequences of somatic mutations is to identify annotated genomic features that may be affected by them. Features that are more likely to encode genomic functions include protein-coding and noncoding transcripts, transcription factor binding sites and other potential regulatory regions. Less well-characterized features, such as highly conserved regions or regions of open chromatin, may also be of interest. There are a variety of software tools that infer the consequences of mutations, but frequently these use different terms and different definitions for the effect itself[8–10] (**Supplementary Table 1**).

A large consortium such as the ICGC requires a common set of terms describing mutation consequences to facilitate the comparison of results among different groups. We have developed a standard set of 'consequence terms' drawn from the Sequence Ontology[11] (**Supplementary Table 2**). This list will be extended and updated as the projects of the consortium unfold. Along with the Sequence Ontology term used to describe the effect of a mutation, we also identify a minimal set of ancillary information that annotation tools should provide for each relevant consequence term, such as coding DNA sequence, protein relative coordinates and predicted amino acid substitutions. Several of these annotations will depend on the specific transcript the mutation falls within, and so we recommend that a transcript identifier always be included. Note that this caveat means that a single mutation can be, and frequently will be, assigned multiple consequences on multiple transcripts.

We recommend using tools that can output mutation descriptions in the format defined by Human Genome Variation Society at all relevant levels (for example, DNA level for all mutations, and RNA and protein level descriptions where applicable). This nomenclature provides a succinct and feature-centric format for description of variants, and some of the tools listed in **Supplementary Table 1** (for example, the Ensembl Variant Effect Predictor (VEP)) have options to produce output in this format. We propose a common ranking scheme for the term set that summarizes the effects of a mutation that falls in multiple genomic features, such as multiple transcripts (**Supplementary Table 2**). In addition, the ranking may be used to prioritize mutations for follow-up analysis.

When assigning consequence terms to variants, one must note the source of all underlying annotations, such as gene models and regulatory elements, to clearly document the event. In the context of ICGC, we recommend using the GENCODE[12] comprehensive set of gene models for all gene-associated annotations and identifying the specific release that was used. We advocate the use of GENCODE annotations because of the detailed and frequently updated annotation of splice variants, pseudogenes and noncoding RNA loci, and the ready accessibility of all data for automated annotation via the Ensembl genome browser[13] and the University of California Santa Cruz (UCSC) genome browser[14]. Using the same gene models as the Encyclopedia of DNA Elements (ENCODE) project[15] will also allow additional integration of somatic mutation data and the wider set of ENCODE annotations.

Comparing the list of mutations to catalogs of known variants. An obvious step in determining the implication of detected variants is to identify those that have been observed previously in other cancers, that are involved in other diseases or that exist as germline polymorphisms. The growing collection of somatic variants detected in different ICGC projects is a useful source of information, as are databases such as dbSNP[16], 1000 Genomes[17], Catalogue of Somatic Mutations in Cancer (COSMIC)[18] and databases of variants associated with hereditary diseases, such as The Human Gene Mutation Database[19] and Exome Variant Server (http://evs.gs.washington.edu/EVS/). Several of the tools listed in **Supplementary Table 2** automatically report whether the variant is already known. As none of these sources are definitive, the ICGC recommends that, at a minimum, projects report matches to variants known in dbSNP, Online Mendelian Inheritance in Man (OMIM)[20], 1000 Genomes and COSMIC along with the version number of the database. Although dbSNP historically contained germline variants for the most part, many somatic mutations including mutational hotspots are also present in newer releases, for example in *JAK2*, *KRAS* and *BRAF*. Thus, projects should make sure that the origin of dbSNP variants is taken into account, especially if the database is used to filter out somatic mutations.

Approach 2: assessing the functional impact of mutations

For many variants, no additional assessment other than determining the functional element they affect can be made about their potential impact on cell operation. Nevertheless, for the specific subset of mutations that affect either protein-coding sequences or known regulatory sites, one can make computational predictions about their potential effects. Here we describe computational analyses that may shed light on the possible functions of these variants.

Mutations affecting protein-coding sequence. Several computational methods have been developed to differentiate 'functional' or 'disease-associated' nonsynonymous mutations from 'nonfunctional' or polymorphic variants[21–26] (**Supplementary Table 3**). Some of these are specifically designed for cancer variants[27–30]. As a general rule, these approaches use evolutionary information (multiple sequence alignments), secondary and tertiary structure features, physicochemical properties of amino acids as well as information about the role of amino acid side chains in the three-dimensional structure of proteins, such as protein surface placement in interaction sites.

Methods aimed at assessing the functional effect of nonsynonymous mutations can be classified as based on machine learning and direct. Machine learning–based methods use relevant properties of the original and mutant residues (for example, size and polarity), structural information (for example, surface accessibility and hydrogen bonding) and/or evolutionary conservation and other features. These methods are then trained to distinguish between positive sets of disease-associated variants and negative control sets of presumably nonfunctional or passenger variants. In contrast, direct methods assess the effect of a mutation through a computed phenomenological score based on a particular theoretical model that does not require training sets.

Most of these computational approaches have been benchmarked on variants with pronounced phenotypic effects[31] (for example, functionally deleterious and Mendelian disease-associated variants) and appropriate negative control sets, resulting in

reporting accuracies close to ~80%. Although these approaches were not originally designed for this purpose, some of them have been widely used to rank cancer somatic mutations for their likelihood to be drivers, without previously benchmarking their performance on this problem.

One of the main challenges to produce such benchmarking is the difficulty of collecting well-curated sets of driver and passenger mutations. A recent effort to circumvent this problem used various data sets of likely driver mutations and likely passenger mutations[27]. Under the assumption that each proxy data set is incomplete in nonoverlapping ways, this study compared the performance of three well-known methods and their impact scores, transformed to account for the baseline tolerance across several data sets rather than on individual data sets[27]. In the future, when many more cancer genomes have been sequenced and we understand better the implication of genetic variants on cancer phenotype, it may be possible to collect gold-standard data sets to perform more accurate validation.

Given the high-throughput nature of cancer genome projects, one important aspect to consider for tool selection is the computational efficiency of the tools when thousands of variants are analyzed. Precomputation of functional impact scores for all possible mutations in the human proteome is a useful remedy (as done by some tools presented in **Supplementary Table 3**). There is also at least one database (database of nonsynonymous single-nucleotide polymorphisms, dbNSFP[32]) devoted to collecting and integrating such precomputed functional impact scores from different tools. In some cases it may be useful to visualize the location of substitutions in protein three-dimensional structure, if available, to assess their potential role with respect to protein stability and/or function, for instance using MuPIT Interactive[33] or the MutationAssessor web server[24].

The output of any computational method should be interpreted as a ranked list of candidate driver variants based on the user-submitted mutations, with the vast majority not likely to be true positives. The purpose of this ranking is to prioritize mutations for experimental testing. Using a combination of methods based on different theoretical principles (and hence independent error models) may help mitigate false positive and negative rates suffered by any one method alone, thus resulting in a cleaner list of candidates for experimental validation.

Mutations affecting regulatory sites. Only very recently has it become feasible to identify and characterize somatic noncoding mutations that affect putative regulatory sites. Predicting the functional effects of regulatory variants typically starts either by purely statistical approaches, such as the application of machine-learning methods to learn motif models from the regulatory sequences or by modeling the transcription factor to DNA binding biophysics aided by experimental data such as those obtained from microfluidics or protein-binding experiments[34,35]. Both approaches result in predictions of binding sites for different transcription factors in regulatory sequences. There are several tools for making such predictions, such as The Meme Suite[36], and the ENCODE project catalogs relevant experimental data sets[15]. Furthermore, RegulomeDB provides an integrated approach to analyze regulatory variants[37]. It uses data sets from ENCODE[15] and other sources and also uses motif models (for example, from JASPAR[38]).

When a somatic mutation falls within a transcription factor binding site, it is possible to score its effect in multiple ways. Perhaps the simplest is to take the relevant binding site motif model[38] and evaluate the score difference that the variant causes in that binding site's match to the model. This is close in spirit to scores that are derived from multiple alignments, such as Pfam log *E* value[39]. However, the interpretation of this particular score is not straightforward because the actual probability of a transcription factor binding to DNA depends strongly on the factor concentration in the cell and the presence of other protein-binding factors, and may thus vary across cell types. Furthermore, it is not clear in general whether stronger or weaker predicted binding is better or worse for transcription factor function, and clarifying this will require studying the particular promoter and gene in more detail.

Pleasance *et al.*[40] used a specific tool[41] to address the functionality of mutations in promoters in a lung cancer cell line. Although somatic mutations did not differ from the null expectation as a set, individual variants were predicted to have considerable disruptive effects on potential binding motifs. More recently, systematic analyses integrating transcription factor binding, histone marks and other epigenomic data were used to identify pathways disrupted in a genome-wide association study at the regulatory level[42].

In addition to considering effects of mutations in promoters and enhancers, it is also important to consider possible effects of mutations on splicing, especially now that the connection between splicing and cancer is becoming increasingly clear[43]. Consequences of mutations in splicing regulatory elements are still difficult to predict, but including additional experimental data, such as RNA sequencing data, may lead to improvements in this area.

Given that the majority of somatic mutations reside in noncoding sequence, the need to computationally prioritize them for follow-up functional validation is clear. The recent discovery of melanoma driver mutations in the promoter sequence of telomerase reverse transcriptase (*TERT*) gene highlights the potential of regulatory variation to drive tumorigenesis[44,45]. As cancer genome projects are moving toward sequencing whole genomes, more noncoding driving mutations will likely be discovered. To facilitate such discoveries, more computational method development to score regulatory variants is needed.

Approach 3: finding signs of positive selection across a cohort

Independent of whether or not a functional consequence can be predicted for a given mutation, one can assess to what extent a given mutation has been observed at a higher frequency than expected. The rationale for assessing mutation frequency is that driver mutations provide an adaptive advantage to cancer cells (**Box 1**; for example, the BRAF V600E substitution found in melanoma[7]) and should thus be positively selected during the clonal evolution of tumors. Provided that similar selective pressures act on different patient tumors and that the same mutation is positively selected, one should be able to trace driver mutations by noting their higher frequency, a common trace of positive selection.

In principle, exploiting this fact to find driver genes is straightforward: it is simply a statistical comparison between the mutation rate observed in a gene versus what is expected under a

BOX 2 CURRENT CHALLENGES

1. Assess the functional impact of sets of mutations. Most current methods cannot accurately predict changes in protein and cellular function because changes in tumor phenotype typically result from multiple genetic alterations.

2. Complement the identification of functional and driver mutations by the prediction of how mutations affect protein and cellular function. There is a need for methods that not only identify functional or driver mutations but also predict the likely cellular outcome resulting from mutations, such as gain, loss or switch of function, and how mutations might affect cellular networks.

3. Apply predictive tools to biologically relevant questions such as drug resistance. The ideal method should not only predict the effect of multiple mutations in an integrative manner and how they affect protein and cellular outcome but also tackle translational clinical challenges such as drug resistance.

neutral model. However, in practice this approach involves difficult choices with respect to the selection of appropriate models for neutral evolution. For example, germline variation should not be used to calibrate a null model for somatic mutation analysis[28] because this reflects evolutionary pressures and mutation processes during species evolution rather than during the development of cancer. In addition, many cancers have defects in DNA repair processes that change the neutral mutation rate, which have different regional impacts[40,46,47], and local mutation rate is variable depending on other factors such as replication timing[48].

To accurately identify genes with more mutations than expected, gene-specific mutation rates should thus be computed. This can be done using synonymous mutations[49] and/or mutations in introns and untranslated sequences (for example, 'Introns versus Exons'; InVEx)[50]; these approaches, however, can only be effectively used in tumors with very high mutation rates. In other cases, gene-specific mutation rates must be estimated, taking into account factors known to affect mutation rate such as mutation context, replication timing and expression levels (for example, 'Mutational Significance in Cancer' (MuSiC)[51] and 'Mutational Significance' (MutSig)[52]).

Given the difficulties that are intrinsic to recurrence-based methods, new methods have been developed that try to infer signs of positive selection using other means. One such approach, OncodriveFM[53], consists of detecting genes that exhibit a bias toward the accumulation of somatic mutations with high functional impact. This method relies on well-known metrics of the functional impact of individual mutations (those listed in **Supplementary Table 3**) to detect genes and pathways with this functional impact bias[53]. Another approach, ActiveDriver[54], involves the discovery of genes enriched for somatic mutations that alter 'active sites' in proteins, such as signaling sites, regulatory domains or linear motifs, assuming that such active mutations are more likely to have a widespread downstream effect and lead to a phenotypic advantage for tumor cells[54].

In **Supplementary Table 4** we list several statistical approaches recently developed to identify candidate driver genes with signs of positive selection in a cohort of tumors[47,49–51,53–55]. As some of these methods are based on different theoretical principles, we recommend applying multiple complementary methods and comparing their results.

Despite these recent advances, future methods will need to capture the high degree of intertumor heterogeneity, as different tumors may acquire the same hallmark of cancer by different means (known as analogous mutations[56]). This heterogeneity is clearly underestimated in the current driver versus passenger model.

Challenges and future perspectives

Sequencing of cancer genomes is a rapidly expanding field, and consequently computational methods used to interpret these data are evolving. We have described classes of practical tools currently available for analysis of a subset of genetic variation data. Because of the rapid evolution of the field, we purposely avoided recommending particular tools or methods. Instead, we presented general guidelines to assist in making educated choices of methods that can be used to address particular research problems. Several pipelines facilitate the user-friendly application of various tools presented here. For instance the Cancer-Related Analysis of Variants Toolkit (CRAVAT)[57] maps mutations to their consequences on protein-coding genes and it predicts their implication in cancer and disease using Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM)[28] and the Variant Effect Scoring Tool (VEST)[58]. IntOGen-mutations[59] provides a way to apply tools of the three approaches, including mapping mutations using the Ensembl variant effect predictor[8], reporting their functional impact on proteins using MutationAssessor[24], Sorting Intolerant from Tolerant (SIFT)[22], PolyPhen2 (ref. 60) and TransFIC[27], and identifying genes with signs of positive selection across a cohort using OncodriveFM[53].

It is important to emphasize the limited capacity of these approaches to directly identify the causative mutations for tumor development. Rather, they are intended to prioritize candidates for follow-up experiments that may demonstrate the actual implication of those mutations in the cancer phenotype. Reporting the results of these rounds of validation experiments to the methods' authors could in principle help them improve their approaches. The current relative scarcity of established spaces for this information exchange should be specifically addressed as part of the development of this field. Furthermore, these validation experiments will contribute to expand the catalogs of well-characterized driver and passenger mutations, thus creating appropriate data sets for the development of computational prediction tools.

There are three key challenges in the field of cancer mutation analysis (**Box 2**). The first challenge is to improve the accuracy of prediction of the functional impact of a mutation. Because mutations do not occur in isolation, but coexist with other somatic alterations that work together to alter cellular processes, separate gene-by-gene analyses are error-prone. A promising direction is the integration of multiple sources of biological information[61], and the use of pathway and network analyses in the interpretation of cancer genomes[24,62,63].

The second challenge is to develop reliable computational methods for the classification of mutations by functional impact type: loss of function, gain of function or switch of function[24,62,63]. The computational classification of mutations by type as well as strength of impact will contribute to the more complete understanding of functional alterations in a cancer genome. The rich information encoded in the three-dimensional structure of proteins, which is not yet well used by current approaches, can be particularly useful for deducing both the functional type and cellular consequences of mutations.

Lastly, there is the practical challenge of identifying mutations that confer resistance or sensitivity to a particular form of therapy[64,65]. We look forward to the day when functional prediction methods support personalized therapeutics, in which the patient's therapy is informed by analysis of the specific genetic alteration profile in an individual tumor. The development of better approaches for analysis of functional and driver mutations will help to facilitate this process and in so doing will support the future development of personalized cancer medicine.

Note: Supplementary information is available in the online version of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. International Cancer Genome Consortium. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
2. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458**, 719–724 (2009).
3. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
4. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
5. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
6. Malumbres, M. & Barbacid, M. RAS oncogenes: the first 30 years. *Nat. Rev. Cancer* **3**, 459–465 (2003).
7. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417**, 949–954 (2002).
8. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
9. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. *Fly* **6**, 80–92 (2012).
10. Medina, I. *et al.* VARIANT: command line, web service and web interface for fast and accurate functional characterization of variants found by next-generation sequencing. *Nucleic Acids Res.* **40**, W54–W58 (2012).
11. Hoehndorf, R., Kelso, J. & Herre, H. The ontology of biological sequences. *BMC Bioinformatics* **10**, 377 (2009).
12. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
13. Flicek, P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48–D55 (2013).
14. Karolchik, D., Hinrichs, A.S. & Kent, W.J. The UCSC Genome Browser. in *Current Protocols in Bioinformatics* (eds. Baxevanis, A.D. *et al.*) 1.4 (2012).
15. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
16. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
17. 1000 Genomes Project Consortium. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
18. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
19. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
20. Amberger, J., Bocchini, C.A., Scott, A.F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**, D793–D796 (2009).
21. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
22. Ng, P.C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
23. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011).
24. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118 (2011).
25. Ryan, M., Diekhans, M., Lien, S., Liu, Y. & Karchin, R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* **25**, 1431–1432 (2009).
26. Stone, E.A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
27. Gonzalez-Perez, A., Deu-Pons, J. & Lopez-Bigas, N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.* **4**, 89 (2012).
28. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
29. Kaminker, J.S., Zhang, Y., Watanabe, C. & Zhang, Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.* **35**, W595–W598 (2007).
30. Capriotti, E. & Altman, R.B. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. *Genomics* **98**, 310–317 (2011).
31. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358–368 (2011).
32. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. *Hum. Mutat.* **32**, 894–899 (2011).
33. Niknafs, N. *et al.* MuPIT Interactive: Webserver for mapping variant positions to annotated, interactive 3D structures. *Hum. Genet.* (in the press).
34. Maerkl, S.J. & Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science* **315**, 233–237 (2007).
35. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
36. Bailey, T.L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
37. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
38. Bryne, J.C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106 (2008).
39. Clifford, R.J., Edmonson, M.N., Nguyen, C. & Buetow, K.H. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* **20**, 1006–1014 (2004).
40. Pleasance, E.D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184–190 (2010).
41. Hoffman, M.M. & Birney, E. An effective model for natural selection in promoters. *Genome Res.* **20**, 685–692 (2010).
42. Cowper-Sal Lari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for *FOXA1* and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
43. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 47–52 (2011).
44. Horn, S. *et al.* *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
45. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
46. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
47. Lohr, J.G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 3879–3884 (2012).

48. Stamatoyannopoulos, J.A. *et al.* Human mutation rate associated with DNA replication timing. *Nat. Genet.* **41**, 393–395 (2009).
49. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
50. Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251–263 (2012).
51. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
52. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* advance online publication, doi:10.1038/nature12213 (16 June 2013).
53. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
54. Reimand, J. & Bader, G.D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
55. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
56. Creixell, P., Schoof, E.M., Erler, J.T. & Linding, R. Navigating cancer network attractors for tumor-specific therapy. *Nat. Biotechnol.* **30**, 842–848 (2012).
57. Douville, C. *et al.* CRAVAT: Cancer-Related Analysis of VAriants Toolit. *Bioinformatics* **29**, 647–648 (2013).
58. Carter, H. *et al.* Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14** (suppl. 3), S3 (2013).
59. Gundem, G. *et al.* IntOGen: integration and data-mining of multidimensional oncogenomic data. *Nat. Methods* **7**, 92–93 (2010).
60. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
61. Masica, D.L. & Karchin, R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* **71**, 4550–4561 (2011).
62. Lee, W., Zhang, Y., Mukhyala, K., Lazarus, R.A. & Zhang, Z. Bi-directional SIFT predicts a subset of activating mutations. *PLoS ONE* **4**, e8311 (2009).
63. Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640–i646 (2012).
64. Iyer, G. *et al.* Genome sequencing identifies a basis for everolimus sensitivity. *Science* **338**, 221 (2012).
65. Valencia, A. & Hidalgo, M. Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. *Genome Med.* **4**, 61 (2012).