Computational approaches to identify functional genetic variants in cancer genomes

Abel Gonzalez-Perez[1,*], **Ville Mustonen**[2,*], **Boris Reva**[3,*], **Graham R.S. Ritchie**[2,4,*], **Pau Creixell**[5], **Rachel Karchin**[6], **Miguel Vazquez**[7], **J. Lynn Fink**[8], **Karin S. Kassahn**[8], **John V. Pearson**[8], **Gary Bader**[13], **Paul C. Boutros**[9,10,11], **Lakshmi Muthuswamy**[9,10], **B.F. Francis Ouellette**[9,12], **Jüri Reimand**[13], **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], and **the ICGC Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group**

[1]Research Unit on Biomedical Informatics, University Pompeu Fabra, Barcelona, 08003, Spain

[2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

[3]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA

[4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[5]Cellular Signal Integration Group, Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800 Lyngby, Denmark

[6]Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA

[7]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain

[8]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, University of Queensland, St. Lucia, Brisbane, Queensland 4072, Australia

[9]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada

[10]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G 2M9, Canada

[11]Department of Pharmacology & Toxicology, University of Toronto, Toronto, ON M5S 1A8, Canada

[12]Department of Cell & Systems Biology, University of Toronto, Toronto, ON M5S 3G4, Canada

[13]The Donnelly Centre, University of Toronto, Toronto, Canada

[14]Division of Cancer Genomics, National Cancer Center, Chuo-ku, Tokyo, 104-0045, Japan

[15]Spanish National Bioinformatics Institute, Madrid 28029, Spain

[16]Cambridge Research Institute, Cambridge CB2 0RE, UK

[17]The Genome Institute, Washington University School of Medicine, St. Louis, MO 63110, USA

[18]Department of Medicine, Division of Oncology, Washington University School of Medicine, St. Louis, MO 63108, USA

[19]Biomolecular Engineering Department, University of California, Santa Cruz, CA 95064, USA

[20]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA

[21]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada

[22]Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain

Abstract

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 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.

Introduction

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 insignificant 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 significantly reduce the number of potential driver candidates by determining the functional impact of each mutation. Thus, a key challenge is to distinguish between functional and non-functional mutations, and by extension between those that contribute to tumorigenesis (drivers) and those that do not (passengers) (see Box 1 for definitions).

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 non-functional variant does not significantly affect the molecular function of a protein. A driver variant confers a selective advantage to a particular tumor cell, while a passenger variant does not. It is important to distinguish between functional versus non-functional and driver versus passenger as they describe different concepts. For example, a mutation might dramatically affect the function of a protein without providing any selective advantage to the tumor (it is a functional passenger variant). Non-synonymous mutations are those that alter the amino acid sequence of a protein.

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 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, like *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 indicates 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 if a particular mutational event induces a selectable phenotype.

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 and likely driver mutations (Fig. 1). Our focus will be on single nucleotide variants (SNVs) and small indels (operationally defined here as variants shorter than 50 bp) 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 undergo follow-up experiments to demonstrate 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 if they have been previously reported. The second uses computational methods to predict the nature and magnitude of the functional impact of mutation in particular elements (*e.g.*, proteins or regulatory regions). The third employs statistical methods to find signs of positive selection across the cohort. Figure 1 lists a subset of the computational tools employed in each of the approaches. In the sections that follow, we review the rationale and tools of each approach and conclude by presenting some of the unsolved challenges and future perspectives in the field.

Approach 1: Mutation mapping, annotation and comparison to known variants

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 non-coding 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 project 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] (see Supplementary Table 2). This list will be extended and updated as the project unfolds. 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 (CDS), 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, 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 (HGVS) at all relevant levels (e.g. DNA-level for all mutations, and RNA and protein level descriptions where applicable). HGVS nomenclature provides a succinct and feature-centric format for variant descriptions, and some of the tools in Supplementary Table 1 (e.g. the Ensembl 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 (see Supplementary Table 2). In addition, the ranking may be used for prioritizing mutations for follow-up analysis.

When assigning consequence terms to variants, the source of all underlying annotations, such as gene models and regulatory elements, must be noted 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 because of the detailed and frequently updated annotation of splice variants, pseudogenes and non-coding RNA loci, and the ready accessibility of all data for automated annotation via Ensembl and UCSC. Using the same gene models as the ENCODE project[13] will also allow further integration of somatic mutation data and the wider set of ENCODE annotations.

Comparing the list of mutations to catalogues 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 within the different ICGC projects is a useful source of information, as are databases such as dbSNP[14], 1000 Genomes[15], Catalogue of Somatic Mutations in Cancer (COSMIC)[16] and databases of variants associated with hereditary diseases[17,18]. Several of the tools listed in Table 1 automatically report if the variant is already known. Since none of these sources are definitive, the ICGC recommends that, at a minimum, projects report matches to variants known in dbSNP, OMIM, 1000 Genomes and COSMIC along with the version number of the database. Although dbSNP has sometimes been used to filter for somatic mutations, historically it contained primarily germline variants. However, in newer releases, many somatic mutations including mutational hotspots are also present, for example in *JAK2*, *KRAS* and *BRAF*. Thus, although we recommend reporting matches in dbSNP we do not recommend using it to filter out somatic mutations.

Approach 2: Assessing the functional impact of mutations

For many variants, no further assessment 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. In this section we describe computational analyses that may shed light on the possible functions of these variants.

Mutations affecting protein coding sequence

A number of computational methods have been developed to differentiate “functional” or “disease-associated” non-synonymous mutations from “non-functional” or polymorphic

variants[19–24] (Supplementary Table 3). Some of these are specifically designed for cancer variants[25–28]. As a general rule, these approaches use evolutionary information (multiple sequence alignments), secondary and tertiary structure features, physico-chemical properties of amino acids, as well as information about the role of amino acid side chains in the 3D structure of proteins, such as protein surface placement in interaction sites.

Methods aimed at assessing the functional effect of non-synonymous mutations can be classified as “machine learning” and “direct”. Machine learning methods use relevant properties of the original and mutant residues (*e.g.*, size, polarity), structural information (*e.g.*, surface accessibility, 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 non-functional 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[29] (*e.g.*, functionally deleterious and Mendelian disease-associated variants) and appropriate negative control sets, reporting accuracies close to ~80%. Although not originally designed for this purpose, some of them have been widely employed 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 employed various datasets of likely driver and likely passenger mutations[25]. Under the assumption that each proxy dataset is incomplete in non-overlapping ways, this study compared the performance of three well-known methods and their impact scores transformed to account for the baseline tolerance across several datasets rather than on individual datasets[25]. 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 datasets to perform more accurate validation.

Given the high-throughput nature of cancer genome projects, one important aspect to consider for tool selection is their computational efficiency 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 (dbNSFP[30]) 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 mutations in protein 3D structure, if available, to further assess their potential role with respect to protein stability and/or function, for instance using MuPIT Interactive[31] or the MutationAssessor web server[22].

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 further 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 (TF) to DNA binding biophysics aided by experimental data such as those obtained from micro-fluidics or protein binding experiments[32,33]. Both approaches result in predictions of binding sites for different TFs within regulatory sequences. There are several tools for making such predictions, such as The Meme Suite[34], and the ENCODE project catalogues a number of relevant experimental data sets[13]. Furthermore, RegulomeDB provides an integrated approach to analyze regulatory variants[35]. It uses datasets from ENCODE[13] and other sources and also uses motif models (eg. from JASPAR[36]).

When a somatic mutation falls within a TF binding site, it is possible to score its effect in multiple ways. Perhaps the simplest is to take the relevant binding site motif model[36] 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[37]. However, the interpretation of this particular score is not straightforward because the actual binding probability of TF to DNA depends strongly on the factor concentration within 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 TF function, and clarifying this will require studying the particular promoter and gene in more detail.

Pleasance *et al.* (ref. 38) used a specific tool[39] to address the functionality of mutations within promoters in a lung cancer cell line. Although somatic mutations did not differ significantly from the null expectation as a set, individual variants were predicted to have significant disruptive effects on potential binding motifs. More recently, systematic analyses integrating TF binding, histone marks, and other epigenomic data were used to identify pathways disrupted by Genome Wide Association Study (GWAS) at the regulatory level[40].

In addition to promoters and enhancers, it is also important to consider possible effects of mutations in splicing, especially now that the connection between splicing and cancer is becoming increasingly clear (e.g., ref 41). Consequences of mutations in splicing regulatory elements are still difficult to predict but including additional experimental data, such as RNA-Seq, may lead to improvements in this area.

Given that the majority of somatic mutations reside in non-coding 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[43,44]. As cancer genome projects are moving toward sequencing whole genomes, more non-coding 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, *e.g.*, *BRAF* V600E mutation 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 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[26] 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[38,45,46], and local mutation rate is variable depending on other factors such as replication timing[47].

To accurately identify significantly mutated genes, gene-specific mutation rates should thus be computed. This can be done using synonymous mutations[48] and/or mutations in introns and UTR sequences (eg. InVex)[49]; however, these approaches 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 (eg. MuSiC[50] and MutSig[51]).

Given the difficulties that are intrinsic to recurrence-based methods, new methods have been developed that try to infer signs of positive selection using alternative means. One such approach, OncodriveFM[52], consists of detecting genes that exhibit a significant bias towards the accumulation of somatic mutations with high functional impact. This method employs well-known metrics of the functional impact of individual mutations (those in Supplementary Table 3) to detect genes and pathways with this functional impact bias[52]. Another novel approach, ActiveDriver[53], involves the discovery of genes significantly 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 wide-spread downstream effect and lead to a phenotypic advantage for tumor cells[53].

Supplementary Table 4 lists several statistical approaches recently developed to identify candidate driver genes with signs of positive selection in a cohort of tumors[46,48–50,52–54]. 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 inter-tumor heterogeneity, as different tumors may acquire the same hallmark of cancer by different means (known as analogous mutations[55]). This heterogeneity is clearly underestimated in the current driver/passenger model.

Challenges and future perspectives

Cancer genome sequencing is a rapidly expanding field, and consequently computational methods used to interpret these data are evolving. We have presented a review of classes of practical tools currently available for analysis of a subset of genetic variation data. Because of the rapid evolution of the field, we have purposely avoided recommending particular tools or methods. Instead we present general guidelines to assist in making educated choices of methods that can address particular research problems. A number of pipelines facilitate the user-friendly application of various tools presented here. For instance, CRAVAT[56] maps mutations to their consequences on protein coding genes and it predicts their implication in cancer and disease using CHASM[26] and VEST[57]. IntOGen-mutations[58] provides a way to

apply tools of the three approaches, including mapping mutations using Ensembl VEP[8], reporting their functional impact on proteins (using MutationAssessor[22], SIFT[20], PolyPhen2[59] and TransFIC[25]) and identifying genes with signs of positive selection across a cohort using OncodriveFM[52].

It is important to emphasize the limited capacity of these approaches to directly identify the causative mutations of tumor development. Rather, they are intended to prioritize candidates for follow-up experiments that may demonstrate their actual implication in the cancer phenotype. Reporting back the results of these rounds of validation experiments to the method's 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 datasets for the development of computational prediction tools.

There are three key challenges in the field of cancer mutation analysis (Box 2). The first 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[60], and the use of pathway and network analyses in the interpretation of cancer genomes[22,61,62].

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.

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[22,61,62]. The computational classification of mutations by type as well as strength of impact will contribute to the more complete elucidation of functional alterations in a cancer genome. The rich information encoded in the 3D structure of proteins, which is not yet well utilized 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 (see for example[63,64]). 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. ICGC et al. International network of cancer genome projects. Nature. 2010; 464:993–998. [PubMed: 20393554]
2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009; 458:719–724. [PubMed: 19360079]
3. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000; 100:57–70. [PubMed: 10647931]
4. Hanahan D, Weinberg R. a Hallmarks of cancer: the next generation. Cell. 2011; 144:646–74. [PubMed: 21376230]
5. Futreal PA, et al. A census of human cancer genes. Nature Reviews Cancer. 2004; 4:177–183.
6. Malumbres M, Barbacid M. RAS oncogenes: the first 30 years. Nature reviews Cancer. 2003; 3:459–65.
7. Davies H, et al. Mutations of the BRAF gene in human cancer. Nature. 2002; 417:949–954. [PubMed: 12068308]
8. McLaren W, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics (Oxford, England). 2010; 26:2069–70.
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)3; iso-2; iso-3. Fly. 2012; 6:80–92. [PubMed: 22728672]
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 research. 2012; 40:W54–8. [PubMed: 22693211]
11. Hoehndorf R, Kelso J, Herre H. The ontology of biological sequences. BMC Bioinformatics. 2009; 10:377. [PubMed: 19919720]
12. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research. 2012; 22:1760–74. [PubMed: 22955987]
13. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
14. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001; 29:308–11. [PubMed: 11125122]
15. Project G, Asia E, Africa S, Figs S, Tables S. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 135:0–9.
16. Forbes SA, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research. 2010; 39:D945–950. [PubMed: 20952405]
17. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Medicine. 2009; 1:13. [PubMed: 19348700]
18. NHLBI Exome Sequencing Project (ESP) Exome Variant Server.
19. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protocols. 2009; 4:1073–1081.
20. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research. 2003; 31:3812–3814. [PubMed: 12824425]
21. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. The American Journal of Human Genetics. 2011; 88:440–449.

22. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research. 2011; 39:e118. [PubMed: 21727090]
23. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinformatics (Oxford, England). 2009; 25:1431–2.
24. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Research. 2005; 15:978–986. [PubMed: 15965030]
25. Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome medicine. 2012; 4:89. [PubMed: 23181723]
26. Carter H, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research. 2009; 69:6660–7. [PubMed: 19654296]
27. Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Research. 2007; 35:W595–598. [PubMed: 17537827]
28. Capriotti E, Altman RB. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics. 2011; 98:310–7. [PubMed: 21763417]
29. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Human mutation. 2011; 32:358–68. [PubMed: 21412949]
30. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human mutation. 2011; 32:894–9. [PubMed: 21520341]
31. Niknafs N, et al. MuPIT Interactive: Webserver for mapping variant positions to annotated, interactive 3D structures. Human Genetics. 2013 In press.
32. Maerkl SJ, Quake SR. A systems approach to measuring the binding energy landscapes of transcription factors. Science (New York, NY). 2007; 315:233–7.
33. Badis G, et al. Diversity and complexity in DNA recognition by transcription factors. Science (New York, NY). 2009; 324:1720–3.
34. Bailey TL, et al. MEME SUITE: tools for motif discovery and searching. Nucleic acids research. 2009; 37:W202–8. [PubMed: 19458158]
35. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome research. 2012; 22:1790–7. [PubMed: 22955989]
36. Bryne JC, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic acids research. 2008; 36:D102–6. [PubMed: 18006571]
37. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics (Oxford, England). 2004; 20:1006–1014.
38. Pleasance ED, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. 2010; 463:184–190. [PubMed: 20016488]
39. Hoffman MM, Birney E. An effective model for natural selection in promoters. Genome research. 2010; 20:685–92. [PubMed: 20194951]
40. Cowper-Sal Lari R, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. Nature genetics. 2012; 44:1191–8. [PubMed: 23001124]
41. Quesada V, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nature Genetics. 2011; 44:47–52. [PubMed: 22158541]
42. Desmet FO, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic acids research. 2009; 37:e67. [PubMed: 19339519]
43. Horn S, et al. TERT Promoter Mutations in Familial and Sporadic Melanoma. Science (New York, NY). 2013; 339:959–61.
44. Huang FW, et al. Highly Recurrent TERT Promoter Mutations in Human Melanoma. Science (New York, NY). 2013; 339:957–9.
45. Pleasance ED, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2010; 463:191–196. [PubMed: 20016485]

46. Lohr JG, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109:3879–84. [PubMed: 22343534]
47. Stamatoyannopoulos JA, et al. Human mutation rate associated with DNA replication timing. Nature genetics. 2009; 41:393–395. [PubMed: 19287383]
48. Greenman C, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007; 446:153–158. [PubMed: 17344846]
49. Hodis E, et al. A Landscape of Driver Mutations in Melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]
50. Dees ND, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Research. 2012; 22:1589–98. [PubMed: 22759861]
51. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 201310.1038/nature12213
52. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic acids research. 2012; 40:e169. [PubMed: 22904074]
53. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Molecular Systems Biology. 2013; 9:637. [PubMed: 23340843]
54. Sjöblom T, et al. The consensus coding sequences of human breast and colorectal cancers. Science (New York, NY). 2006; 314:268–274.
55. Creixell P, Schoof EM, Erler JT, Linding R. Navigating cancer network attractors for tumor-specific therapy. Nature Biotechnology. 2012; 30:842–848.
56. Douville C, et al. CRAVAT: Cancer-Related Analysis of VAriants Toolit. Bioinformatics. 2013
57. Carter H, et al. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics. 2013; 14(Supl 3):S3. [PubMed: 23819870]
58. Gonzalez-Perez, A.; Perez-Llamas, C.; Santos, A.; Deu-Pons, J.; Lopez-Bigas, N. IntOGen-mutations pipeline: To interpret catalogs of cancer somatic mutations. 2013. at <http://www.intogen.org/mutations/analysis>
59. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249. [PubMed: 20354512]
60. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer research. 2011; 71:4550–61. [PubMed: 21555372]
61. Lee W, Zhang Y, Mukhyala K, Lazarus RA, Zhang Z. Bi-Directional SIFT Predicts a Subset of Activating Mutations. PLOS one. 2009; 4:e8311. [PubMed: 20011534]
62. Ng S, et al. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics (Oxford, England). 2012; 28:i640–i646.
63. Iyer G, et al. Genome sequencing identifies a basis for everolimus sensitivity. Science (New York, NY). 2012; 338:221.
64. Valencia A, Hidalgo M. Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics. Genome medicine. 2012; 4:61. [PubMed: 22839973]
65. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010; 38:e164. [PubMed: 20601685]
66. Makarov V, et al. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. Bioinformatics (Oxford, England). 2012; 28:724–5.
67. Habegger L, et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics (Oxford, England). 2012; 28:2267–9.
68. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biology. 2007; 8:R232. [PubMed: 17976239]
69. Wong WC, et al. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics (Oxford, England). 2011; 27:2147–8.
70. Hartl, DL.; Clark, AG. Principles of Population Genetics. 4. Sinauer Associates, Inc; 2006. p. 545

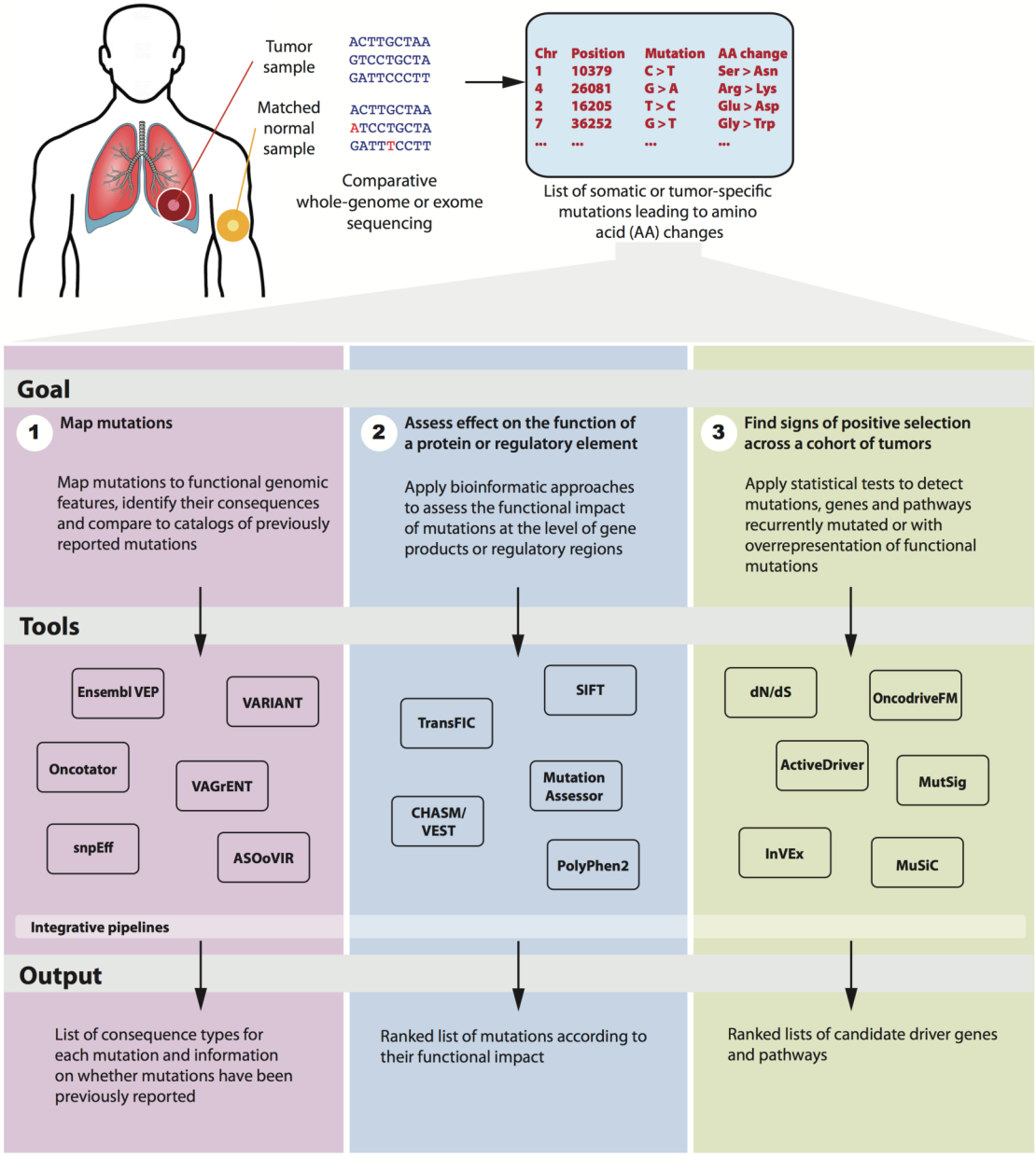

Figure 1.

Scheme depicting the three main approaches routinely employed in the analysis of cancer somatic mutations, as reviewed in this perspective. Although there are important relationships of precedence between elements from different approaches, they do not necessarily correspond to sequential steps. Tools employed in each of the approaches are shown in the middle. Integrative pipelines refer to tools that facilitate the use of methods across all approaches (*e.g.*, IntOGen-mutations pipeline).