Genome Medicine

REVIEW

Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics

Alfonso Valencia* and Manuel Hidalgo

Abstract

Progress in genomics has raised expectations in many fields, and particularly in personalized cancer research. The new technologies available make it possible to combine information about potential disease markers, altered function and accessible drug targets, which, coupled with pathological and medical information, will help produce more appropriate clinical decisions. The accessibility of such experimental techniques makes it all the more necessary to improve and adapt computational strategies to the new challenges. This review focuses on the critical issues associated with the standard pipeline, which includes: DNA sequencing analysis; analysis of mutations in coding regions; the study of genome rearrangements; extrapolating information on mutations to the functional and signaling level; and predicting the effects of therapies using mouse tumor models. We describe the possibilities, limitations and future challenges of current bioinformatics strategies for each of these issues. Furthermore, we emphasize the need for the collaboration between the bioinformaticians who implement the software and use the data resources, the computational biologists who develop the analytical methods, and the clinicians, the systems' end users and those ultimately responsible for taking medical decisions. Finally, the different steps in cancer genome analysis are illustrated through examples of applications in cancer genome analysis.

The cancer genome challenge and the importance of analytical pipelines

Recent progress in incorporating genomic information into clinical practice means that it is becoming increasingly feasible to personalize treatment according to the composition of the patient's genome [1]. Indeed, biomedicine seems to be moving rapidly in this direction [2]. Current estimates predict that the cost of sequencing will drop to below US$1,000 per genome and that when sequencing 1 million bases costs less than $1 it will become economically feasible to systematically implement this type of clinical approach [3-6]. The full implications of massive sequencing in a clinical setting have been discussed extensively [7-10], including discussion of some of the economic considerations, which are of considerable general interest [11].

There are already a number of exciting examples of the application of whole-genome sequencing to the study of Mendelian diseases. For example, in one family with four siblings affected by Charcot-Marie-Tooth disease (a peripheral polyneuropathy), a direct relationship between a specific gene locus and this disease was demonstrated [12]. Moreover, analyses of individual genomes have also now been published [13-17], including the first complete individual high-throughput approach [18].

Cancer is a general class of diseases that may benefit from the application of personalized therapeutic approaches, particularly given the wide spectrum of mutations that must be analyzed and the complexity of cancer-related genome variation: germline susceptibility, somatic single nucleotide and small insertion/deletion mutations, copy number alterations, structural variants and complex epigenetic regulation.

Initial whole-genome sequencing studies have included the sequencing of the genome of a patient with chronic lymphocytic leukemia, in which novel somatic mutations were identified by comparing the variations in the tumor with both control tissue and the available database information [19]. Alternative approaches involve the sequencing of coding regions alone (exomes), with the implicit reduction in the cost and effort required. Such analyses have also led to significant advances in our understanding of several types of cancer (see, for example, [20-24]).

Our work in this area is strongly motivated by the case of a patient with advanced pancreatic cancer who responded dramatically to mitomycin C treatment [25]. The molecular basis for this response, the inactivation of

*Correspondence: valencia@cnio.es
Spanish National Cancer Research Centre (CNIO), Calle Melchor Fernández Almagro, 3, E-28029 Madrid, Spain

the *PALB2* gene, was discovered by sequencing almost all the coding genes in the cancer cells from this patient [26]. Approximately 70 specific variations were detected in the tumor tissue and they were analyzed manually to search for mutations that might be related to the onset of the disease and, more importantly from a clinical point of view, that could be targeted with an existing drug. In this case, the mutation in the *PALB2* gene was linked to a deficiency in the DNA repair mechanism [27] and this could be targeted by mitomycin C.

The obvious challenge in relation to this approach is to develop a systematic form of analysis in which a bioinformatics-assisted pipeline can rapidly and effectively analyze genomic data, thereby identifying targets and treatment options. An ideal scenario for personalized cancer treatment would require performing the sequencing and analysis steps before deciding on new treatments.

Unfortunately, there are still several scientific and technical limitations that make the direct implementation of such a strategy unfeasible. Although pipelines to analyze next-generation sequencing (NGS) data have become commonplace, the systematic analysis of mutations requires more time and effort than is available in routine hospital practice. A further challenge is to predict the functional impact of the variations discovered by sequencing, which presents serious obstacles in terms of the reliability of current bioinformatics methods. These difficulties are particularly relevant in terms of protein structure and function prediction, the analysis of non-coding regions, functional analyses at the cellular and sub-cellular levels, and the gathering of information about the relationships between mutations and drug interactions.

Our own strategy is focused on testing the drugs and treatments proposed by the computational analysis of genomic information in animal models as a key clinical element. The use of xenografts, in which nude mice are used to grow tumors seeded by implanting fragments of the patient's tissue, may be the most practical model of real human tumors. Despite their limitations, including the mixture of human and animal cells and the possible differences in the evolution of the tumors with respect to their human counterparts, such 'avatar' models provide valuable information about the possible treatment options. Importantly, such xenografts allow putative drugs or treatments for individual tumors to be assayed before applying them in clinical practice [25].

A summary of the elements that are required in an ideal data analysis pipeline is depicted in Figure 1, including: the analysis of genomic information; prediction of the consequences of specific mutations, particularly in protein coding regions; interpretation of the variation at the gene/protein network level; and the basic approaches in pharmacogenomic analysis to identify potential drugs related to the predicted genetic alterations. Finally, the pipeline includes the interfaces necessary to integrate the genomic information with other resources required by teams of clinicians, genome experts and bioinformaticians to analyze the information.

In this review, we outline the possibilities and limitations of a comprehensive pipeline and the future developments that will be required to generate it, including a brief description of the approaches currently available to cover each stage. We begin by examining the bioinformatics required for genome analysis, before focusing on how mutation and variation data can be interpreted, then explore network analysis and the downstream applications available for selecting appropriate drugs and treatments.

Genome analysis

Array technologies are relied on heavily to analyze disease-related tissue samples, including expression arrays and single nucleotide polymorphism (SNP) arrays to analyze point mutations and structural variations. However, personalized medicine platforms are now ready to benefit from the transition from these array-based approaches towards NGS technology [28].

The detection of somatic mutations by analyzing sequence data involves a number of steps to filter out technical errors. The first series of filters are directly related to the sequencing data and they vary depending on the technical set-up. In general, this takes into consideration the base-calling quality of the variants in the context of the corresponding regions. It also considers the regions covered by sequencing and their representativeness or uniqueness at the genome level.

As the sequencing and software analysis technologies are not fully integrated, errors are not infrequent and, in practice, thousands of false positives are detected when the results move on to the validation phase. In many cases, this is due to the non-unique placement of the sequencing reads in the genome or the poor quality of alignments. In other cases, variants can be missed because of insufficient coverage of the genomic regions.

The analysis of tumors is further complicated by their heterogeneous cellular composition. New experimental approaches are being made available to address the heterogeneity of normal and disease cells in tumors, including single-cell sequencing [29,30]. Other intrinsic difficulties include the strong mosaicism recently discovered [31-33], and thus greater sequencing quality and coverage is necessary and more stringent sample selection criteria must be applied. These requirements place additional pressure on the need to acquire samples in sufficient quantity and of appropriate purity, inevitably increasing the cost of such experiments.

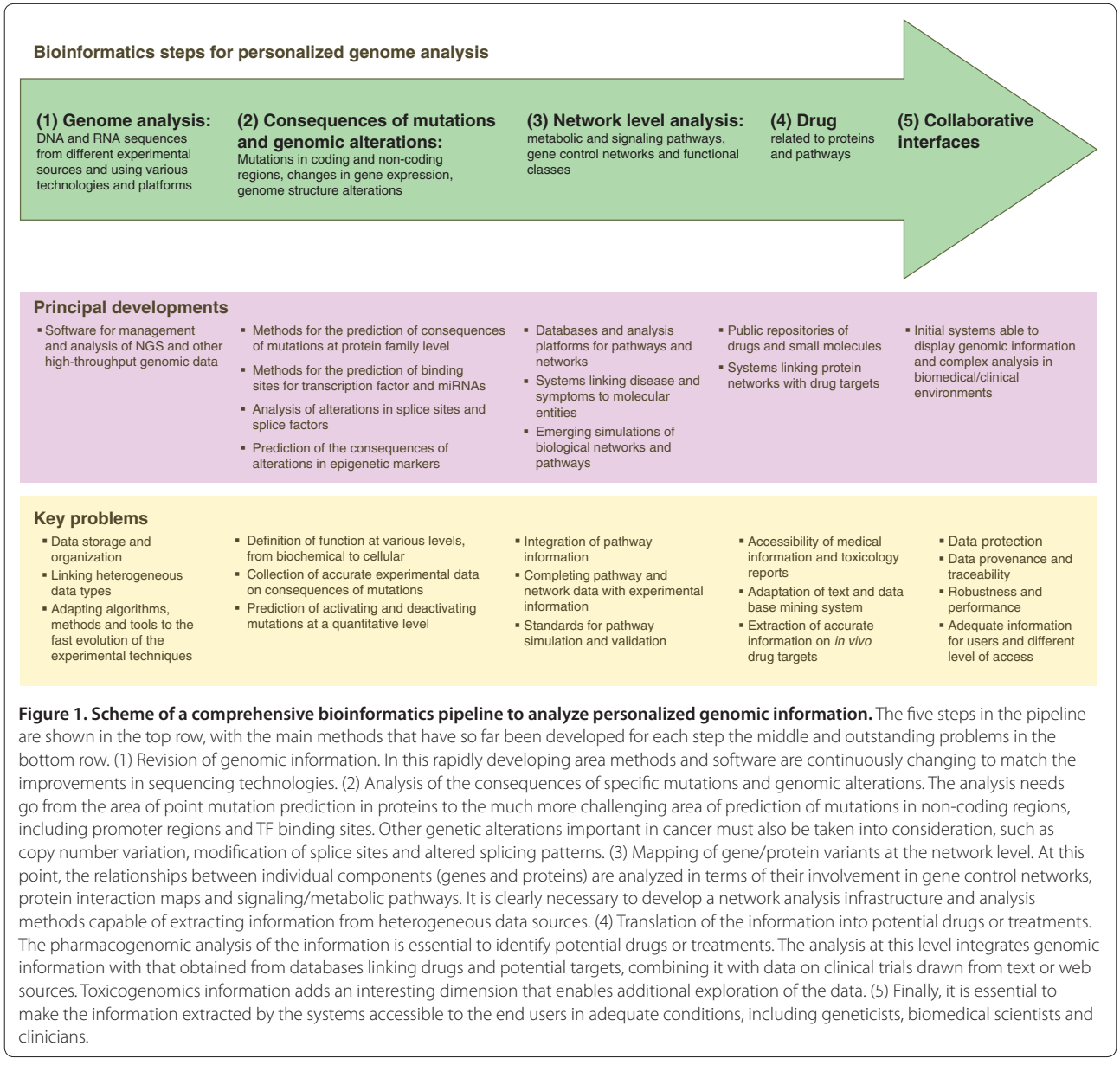

Figure 1. Scheme of a comprehensive bioinformatics pipeline to analyze personalized genomic information. The five steps in the pipeline are shown in the top row, with the main methods that have so far been developed for each step the middle and outstanding problems in the bottom row. (1) Revision of genomic information. In this rapidly developing area methods and software are continuously changing to match the improvements in sequencing technologies. (2) Analysis of the consequences of specific mutations and genomic alterations. The analysis needs go from the area of point mutation prediction in proteins to the much more challenging area of prediction of mutations in non-coding regions, including promoter regions and TF binding sites. Other genetic alterations important in cancer must also be taken into consideration, such as copy number variation, modification of splice sites and altered splicing patterns. (3) Mapping of gene/protein variants at the network level. At this point, the relationships between individual components (genes and proteins) are analyzed in terms of their involvement in gene control networks, protein interaction maps and signaling/metabolic pathways. It is clearly necessary to develop a network analysis infrastructure and analysis methods capable of extracting information from heterogeneous data sources. (4) Translation of the information into potential drugs or treatments. The pharmacogenomic analysis of the information is essential to identify potential drugs or treatments. The analysis at this level integrates genomic information with that obtained from databases linking drugs and potential targets, combining it with data on clinical trials drawn from text or web sources. Toxicogenomics information adds an interesting dimension that enables additional exploration of the data. (5) Finally, it is essential to make the information extracted by the systems accessible to the end users in adequate conditions, including geneticists, biomedical scientists and clinicians.

After analyzing the sequence data, putative mutations must be compared with normal tissue from the same individual, as well as with other known genetic variants, to identify true somatic mutations related to the specific cancer. This step involves comparing the data obtained with information regarding variation and with complete genomes, which can be obtained from various databases (see below), as well as with information on rare variants [34,35]. For most applications, including the possible use in a clinical setup, a subsequent validation step is necessary, which is normally carried out by PCR sequencing of the variants or, where possible, by sequencing biological replicates.

Exome sequencing

The cost of whole-genome sequencing still remains high. Furthermore, when mutations associated with diseases are mapped in genome-wide association studies (GWASs) [36], they tend to map in regulatory and functional elements but not necessarily in the conserved coding regions, which actually represent a very small fraction of the genome. This highlights the importance of studying mutations in non-coding regions and the need for more experimental information on regulatory elements, including promoters, enhancers and microRNAs (miRNAs; see below). Despite all these considerations, the current alternative for economic and technical

reasons is often to limit sequencing to the coding regions in the genome (exome sequencing), which can be performed for less than $2,000. Indeed, sequencing all the exons in a genome has already provided useful data for disease diagnosis, such as in identifying the genes responsible for Mendelian disorders in studies of a small number of affected individuals. Such proof-of-concept studies have correctly identified the genes previously known to underlie diseases such as Freeman-Sheldon syndrome [37] and Miller syndrome [38].

A key step in exome sequencing is the use of the appropriate capturing technology to enrich the DNA samples to be sequenced with the exons desired. There has been considerable progress in developing and commercializing arrays to capture specific exons (for example, see [39]), which has facilitated the standardization and systematization of such approaches, thereby increasing the feasibility of applying these techniques in clinical settings.

Despite the current practical advantages offered by exome sequencing, it is possible that technological advances will soon mean that it will be replaced by whole-genome sequencing, which will be cheaper in practice and requires less experimental manipulation. However, such a scenario will certainly increase the complexity of the bioinformatic analysis (see, for example, [40] for an approach using whole-genome sequencing, or [19] for the combined use of whole-genome sequencing as a discovery system, followed by exome sequencing validation in a larger cohort).

Sequencing to study genome organization and expression

NGS can provide sequence information complementary to DNA sequencing that will be important for cancer diagnosis, prognosis and treatment. The main applications include RNA sequencing (RNA-seq), miRNAs and epigenetics.

NGS-based approaches can also be used to detect structural genomic variants, and these techniques are likely to provide better resolution than previous array technologies (see [41] for an initial example). Cancer research is an obvious area in which this technology will be applied, as chromosomal gains and losses are very common in cancer. Further improvements in this sequencing technology, and in the related computational methods, will enable more information to be obtained at a lower cost [42] (see also a recent application in [43] and the evolution of computational approaches from [44-46] to [47]).

RNA-seq

DNA sequencing data, particularly data from non-coding regions (see below), can be better understood when accompanied by gene expression data. Direct sequencing of RNA samples already provides an alternative to the use of expression arrays, and it promises to increase the accessible dynamic range and limits of sensitivity [48-50]. RNA-seq could be used to provide a comprehensive view of the differences in transcription between normal and diseased samples but also to correlate alterations in structure and copy number that may affect gene expression, thereby helping to interpret the consequences of mutations in gene control regions. Furthermore, RNA sequencing data can be used to explore the capacity of the genome to produce alternative splice variants [51-55]. Indeed, the prevalence of splice variants at the genomic level has been assessed, suggesting a potential role for the regulation of alternative splicing in different stages of disease, and particularly in cancer [56,57]. Recent evidence clearly points to the importance of mutations in splicing factors and RNA transport machinery in cancer [24,58].

miRNAs

NGS data on miRNAs can also complement sequencing data. This is particularly important in cancer research given the rapidly expanding roles proposed for miRNAs in cancer biology [59]. For example, interactions have been demonstrated between miRNA overexpression and the well-characterized Sonic hedgehog/Patched signaling pathway in medulloblastoma [60]. Moreover, novel miRNAs and miRNAs with altered expression have also been detected in ovarian and breast cancers [61,62].

Epigenetics

NGS can provide invaluable data on DNA methylation (methyl-seq) and the epigenetic modification of histones - for example, through chromatin immunoprecipitation sequencing (ChIP-seq) with antibodies corresponding to the various modifications. Epigenetic mechanisms have been linked to disease [63,64] (reviewed in [65]).

The wealth of information provided by all these NGS-based approaches will substantially increase our capacity to understand the complete genomic landscape of the disease, although it will also increase the complexity of the analysis at all levels, from basic data handling to problems related to data linking to interpretation. There will also be complications in areas in which our knowledge of the basic biological processes is developing at the same rhythm as the analytical technology (for a good example of the intrinsic association between new discoveries in biology and the development of analytical technologies, see recent references on chromothripsis [66-68]). Furthermore, it is important to keep in mind that, from the point of view of clinical applications, most if not all drugs available target proteins. Thus, even if it is essential to have complete genomic information to understand a disease and to detect disease markers and

stratification, as well as to design clinical trials, the identification of potential drugs and treatments will still be mainly based on the analysis of alterations in coding regions.

Interpreting mutation and variation data

The growing number of large-scale studies has led to a rapid increase in the number of potential disease-associated genes and mutations (Table 1). An overview of these studies can be found in [69] and the associated web catalog of GWASs [70].

Interpreting the causal relationship between the mutations considered to be significant in GWASs and the corresponding disease phenotypes is clearly complicated, and serious concerns about the efficacy of GWASs have been much discussed [71,72]. In the case of cancer research, the interpretation of mutations is additionally complicated by the dynamic nature of tumor progression, and also the need to distinguish between mutations associated with the initiation of the cancer and others that accumulate as the tumors evolve. In this field, the potential cancer initiators are known as 'drivers' and those that accumulate during tumor growth as 'passengers' (terminology taken from [73], referring metaphorically to the role of certain viruses in either causing or merely being passengers in infected cells).

In practice, the classification of mutations as drivers and passengers is based on their location at positions considered to be important because of their evolutionary conservation, and on observations in other experimental datasets (for a review of the methods used to classify driver mutations and the role of tumor progression models, see [74]). Ultimately, more realistic biological models of tumor development and a more comprehensive understanding of the relationship between individual mutations will be necessary to classify mutations according to their role in the underlying process of tumor progression (reviewed in [75]).

Despite the considerable advances in database development, it will take additional time and effort to fully consolidate all the information available in the scientific literature into databases and annotated repositories. To alleviate this problem, efforts have been made to extract mutations directly from the literature by systematically mapping them to the corresponding protein sequences. For example, CJO Baker and D Rebholz-Schuhmann organize a biennial workshop focusing on this particular approach (the ECCB Workshop: Annotation, Interpretation and Management of Mutations; the corresponding publication is [76]).

In the case of protein kinases, one of the most important families of proteins for cancer research, many mutations have been detected that are not currently stored in databases and that have been mapped to their corresponding positions in protein sequences [77]. However, for a large proportion of the mutations in kinases already introduced into databases, text mining provides additional links to stored information and mentions of the mutations in the literature.

These automated approaches, when applied not only to protein kinases but to any protein family [78-84], should be viewed as a means of facilitating rapid access to information, although they are not aimed at replacing databases, as the text mining results require detailed manual curation. Therefore, in the quest to identify and interpret mutations, it is important to bear in mind that text mining can provide additional information complementary to that retrieved in standard database searches.

Information about protein function

Accurately defining protein function is an essential step in analyzing mutations and predicting their possible consequences. Databases are annotated by extrapolating the functions of the small number of proteins on which detailed experiments have been carried out (estimated to be less than 3% of the proteins annotated in the UniProt database). The protocols for these extrapolations have been developed over the past 20 years and they are continually adjusted to incorporate additional filters and information sources [85-87]. Interestingly, several ongoing community-based efforts aim to evaluate the methods used to predict and extract information regarding protein function, such as Biocreative in the field of text mining [88,89], CASP for predicting function and binding sites [90], and challenge in function prediction organized by Iddo Friedberg and Predrag Radivojac [91].

Protein function at the residue level

The analysis of disease-associated mutations naturally focuses on key regions of proteins that are directly related to their activity. The identification of binding sites and active sites in proteins is therefore an important aid to interpreting the effects of mutations. In this case, and as in other areas of bioinformatics, the availability of large and well-annotated repositories is essential. The annotations of binding sites and active sites in Swiss-Prot [92], the main database with hand-curated annotations of protein characteristics, provide a combination of experimental information and patterns of conservation of key regions. For example, the well-characterized GTP binding site of the Ras family of small GTPases is divided into four small sequence regions. This definition is based on the conservation of these sequences, despite the fact that they include residues that do not directly contact GTP or participate in the catalytic mechanism. Obviously, the ambiguity of this type of definition tends to complicate the interpretation of mutations in such regions.

Table 1. Some of the main data repositories of genetic variation associated with human phenotypes and disease

Name	Description	URL	Reference
dbSNP	General catalog of polymorphisms	http://www.ncbi.nlm.nih.gov/SNP	[142]
Ensembl	Maps known mutations and SNPs in the human genome from other databases	http://www.ensembl.org	[143]
OMIM	Online Mendelian Inheritance in Man; a large collection of disease annotations, often for monogenetic diseases	http://www.ncbi.nlm.nih.gov/omim/	[144]
COSMIC	Catalog of somatic mutations in cancer	http://www.sanger.ac.uk/genetics/CGP/cosmic/	[145]
CGC	Cancer Gene Census	http://www.sanger.ac.uk/genetics/CGP/Census/	[146]

Various tools have been designed to provide validated annotations of binding sites (residues in direct contact with biologically relevant compounds) in proteins of known structure; these include FireDB and FireStar [93]. This information is organized according to protein families so as to help analyze the conservation of the compounds bound and the corresponding binding residues. Other resources, such as the Catalytic Site Atlas [94], provide detailed information about protein residues directly involved in the catalysis of biochemical reactions by enzymes. In addition to substrate binding sites, it is also important to interpret the possible incidence of mutations at sites of interaction between proteins. Indeed, there are a number of databases that store and annotate such interaction sites [95].

Given that there are still relatively few proteins for which binding sites can be deduced from their corresponding structures, it is particularly interesting to be able to predict substrate binding sites and regions of interaction with other protein effectors. Several methods are currently available for this purpose [96-98]; for example, a recently published method [99] automatically classifies protein families into functional subfamilies, and detects residues that may functionally differentiate between subfamilies (for a user-friendly visualization environment, see [100]).

Prediction of the consequences of point mutations

Several methods are currently used to predict the functional consequences of individual mutations. In general, they involve a combination of parameters related to the structure and stability of proteins, interference from known functional sites, and considerations about the evolutionary importance of sites. These parameters are calculated for a number of mutations known to be linked to diseases and in the majority of systems they are extrapolated to new cases using machine learning techniques (support vector machines, neural networks, decision trees and others; for a basic reference in the field, see [101]).

The process of predicting the consequences of mutations is hampered by numerous inherent limitations, such as those listed below.

(1) Most of the known mutations used to calibrate the system are only weakly associated with the corresponding disease. In some cases the relationship is indirect or even non-existent (for example, mutations derived from GWASs; see above).
(2) The prediction of the structural consequences of mutations is a new area of research, and thus the risks of misinterpretation are considerable, particularly given the flexibility of proteins and our limited knowledge of protein folding.
(3) The consequences of mutations in protein structures should ideally be interpreted in quantitative terms, taking energies and entropies into account. This requires biophysical data that are not yet available for most proteins.
(4) Predictions are made on the assumption that proteins act alone when, in reality, specific constraints and interactions within the cellular or tissue environment can considerably attenuate or enhance the effects of a mutation.
(5) The current knowledge of binding sites, active sites and interaction sites is limited (see above). The accuracy of predictions regarding the effects of mutations at these sites is thus similarly limited.

Despite such limitations, these approaches are very useful and they currently represent the only means of linking mutations with protein function (Table 2). Many of these methods are user-friendly and well documented, with their limitations emphasized to ensure careful analysis of the results. Indeed, an initial movement to assess prediction methods has been organized (a recent evaluation of such methods can be found in [102]).

For example, the PMUT method [103] (Table 2) is based on neural networks calibrated using known mutations, integrating several sequence and structural parameters (multiple sequence alignments generated with PSI-BLAST and PHD scores for secondary structure, conservation and surface exposure). The input required is the sequence or alignment, and the output consists of a list of the mutations with a corresponding disease prediction presented as a pathogenicity index that ranges from 0 to 1. The scores corresponding to the neural network's internal parameters are interpreted in terms of

Table 2. Methods for predicting the consequences of point mutations

Name	URL	How it works
SIFT	http://sift.jcvi.org	Uses sequence homology scores that are calculated using position-specific scoring matrices with Dirichlet priors
Polyphen 2	http://genetics.bwh.harvard.edu/pph2/	Uses sequence conservation, structure and Swiss-Prot annotations
PMUT	http://mmb2.pcb.ub.es:8080/PMut/	Formulates predictions with neural networks, using internal databases, secondary structure prediction and sequence conservation
SNPs3D	http://www.snps3d.org/	Based on a support vector machine that uses structural or sequence conservation parameters
PantherPSEC19	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Uses sequence homology scores calculated using PANTHER hidden Markov model families
Mutationassessor	http://mutationassessor.org	Provides predictions using additional information based on the specific patterns of conservation of protein families
VEP (Variant Effect Predictor)	http://www.ensembl.org/info/docs/variation/vep	This system categorizes Ensembl genomic variants in known transcripts by their potential effect
KinMut	http://kinmut.bioinfo.cnio.es	Prediction of the consequences of mutations in protein kinases; the system was trained with specific information about the kinase subfamilies, and together with the predictions provides general information about the corresponding proteins, a comparison with other predictors and links to the related literature

the level of confidence in the prediction. The system also provides pre-calculated results for large groups of proteins, thereby offering a fast and accessible web resource [103].

Perhaps the most commonly used method in this area is SIFT [104] (Table 2), which compiles PSI-BLAST alignments and calculates the probabilities for all the 20 possible amino acids at that position. From this information it predicts to what degree substitutions will affect protein function. In its predictions, SIFT does not use structural information from the average diversity of the sequences in the multiple sequence alignments. The information provided about the variants in protein coding regions includes descriptions of the protein sequences and the families, the estimated evolutionary pressure and the frequency of SNPs at that position (if detected), as well as the association with diseases as found in the Online Mendelian Inheritance in Man (OMIM) database (Table 1).

In the light of the current situation, it is clearly necessary to move beyond the simple predictive methods that are currently available to fulfill the requirements for personalized cancer treatment. As in other fields of bioinformatics (see above), competitions and community-based evaluation efforts that openly compare systems are of great practical importance. In this case, Yana Bromberg and Emidio Capriotti are organizing an interesting workshop on the prediction of the consequences of point mutations [105], and Steven E Brenner, John Moult and Sadhna Rana organize the Critical Assessment of Genome Interpretation (CAGI) to assess computational methods for predicting the phenotypic impacts of genomic variation [106].

A key technical step in analyzing the consequences of mutations in protein structures is the ability to map the mutations described at the genome level onto the corresponding protein sequences and structures. The difficulty of translating information between coordinate systems (genomes and protein sequences and structures) is not trivial, and current methods only provide partial solutions to this problem. The protein structure classification database CATH [107] has addressed this issue using a system that allows the systematic transfer of DNA coordinates to positions in three-dimensional protein structures and models [108].

In addition to the general interpretation of the consequences of mutations, there is a large body of literature on the interpretation of mutations in specific protein families. By combining curated alignments and the detailed analysis of structures or models with sophisticated physical calculations, it is possible to gain additional insight into specific cases. For example, mutations in the protein kinase family have been analyzed, comparing the distribution of these mutations in terms of protein structure and their relationship with active sites and binding sites [109]. The conclusion of this study [109] was that putative cancer driver mutations tend to be more closely associated with key protein features than are other more common variants (non-synonymous SNPs) or somatic mutations (passengers) that are not directly linked to tumor progression. These driver-specific features include molecule binding sites, regions of specific binding to other proteins and positions conserved generally or in specific protein subfamilies at the sequence level. This observation fits well with the implication of altered protein kinase function in cancer

pathogenicity, and it supports the link between cancer-associated driver mutations and altered protein kinase structure and function.

Family-specific prediction methods based on the association of specific features in protein families [110], and on other methods that exploit family-specific information [111,112], pave the way to the development of a new generation of prediction methods that can assess all protein families using their specific characteristics.

Mutations do not only affect binding sites and functional sites but, in many cases, they also alter sites that are subject to post-translational modifications, potentially affecting the function of the corresponding proteins. Perhaps the largest and most effective resource to predict the mutational effects on sites subject to post-translational modification is that developed by Søren Brunak's group [113], which encompasses leucine-rich nuclear export signals, non-classical secretion of proteins, signal peptides and cleavage sites, arginine and lysine propeptide cleavage sites, generic and kinase-specific phosphorylation sites, c-mannosylation sites, glycation of ε amino groups of lysines, *N*-linked glycosylation sites, *O*-GalNAc (mucin type) glycosylation sites, amino-terminal acetylation, *O*-β-GlcNAc glycosylation and 'Yin-Yang' sites (intracellular/nuclear proteins). The output for each sequence predicts the potential of mutations to affect different sites. However, there is as yet no predictor capable of combining the output of this method and applying it to specific mutations. An example of a system to predict the consequences of mutations in an information rich environment is provided in Figure 2.

Mutations in non-coding regions

Predicting the consequences of mutations in non-coding regions presents particular challenges, especially given that current methods are still very limited in formulating predictions based on gene sequence and structure, miRNA and transcription factor (TF) binding sites, and epigenetic modifications. For a review of our current knowledge of TFs and their activity, see [114]; the main data repositories are TRANSFAC, a database of TFs and their DNA binding sites [115], JASPAR, an open-access database of eukaryotic TF binding profiles [116], and ORegAnno, an open-access community-driven resource for regulatory annotation [117].

In principle, these information repositories make it possible to analyze any sequence for the presence of putative TF binding sites and to predict how binding would change following the introduction of mutations. In practice, however, the information relating to binding preferences is not very reliable as it is generally based on artificial *in vitro* systems. Furthermore, it is difficult to account for the effects of gene activation based on this information and it is also impossible to take into account any co-operation between individual binding sites. Although approaches based on NGS or ChIP-seq experiments would certainly improve the accuracy of the information available regarding true TF binding sites in different conditions, predicting the consequences of individual modifications in terms of the functional alterations produced is still difficult. The mapping of mutations in promoter regions and their correlation with TF binding sites thus provides us with only an indication of potentially interesting regions, but it does not yet represent an effective strategy to analyze mutations.

In the case of miRNAs and other non-coding RNAs, the 2012 *Nucleic Acids Research* database issue lists more than 50 databases providing information on miRNAs. As with the predictions of TF binding, it is possible to use these resources to explore the links between mutations and their corresponding sites. However, the methods currently available still cannot provide systematic predictions of the consequences of mutations in regions coding for miRNAs and other non-coding RNAs. Indeed, such approaches are becoming increasingly more difficult owing to the emergence of new forms of complex RNA, which pose further challenges to these prediction methods (reviewed in [118]).

Even if sequence analysis alone cannot provide a complete solution to the analysis of mutations in non-coding regions, combining such approaches with targeted gene expression experiments can shed further light on such events. In the context of personalized cancer treatment, combining genome and RNA sequencing of the same samples could enable the variation in coding capacity of different variants to be assessed directly. Hence, new methods and tools will be required to support the systematic analysis of such combined datasets.

In summary, predicting the functional consequences of point mutations in coding and non-coding regions still remains a challenge, requiring new and more powerful computational methods and tools. However, despite the inherent limitations, several useful methods and resources are now available, which, in combination with targeted experiments, should be explored further to analyze mutations more reliably in a context of personalized medicine.

Network analysis

Cancer and signaling pathways

Cancer has been repeatedly described as a systems disease. Indeed, the process of tumor evolution from primary to malignant forms, including metastasis to other tissues, involves competition between various cell lineages struggling to adapt to the changing conditions, both within and around the tumor. This complex process is closely associated with the occurrence of mutations and genetic alterations. In fact, it seems likely that rather

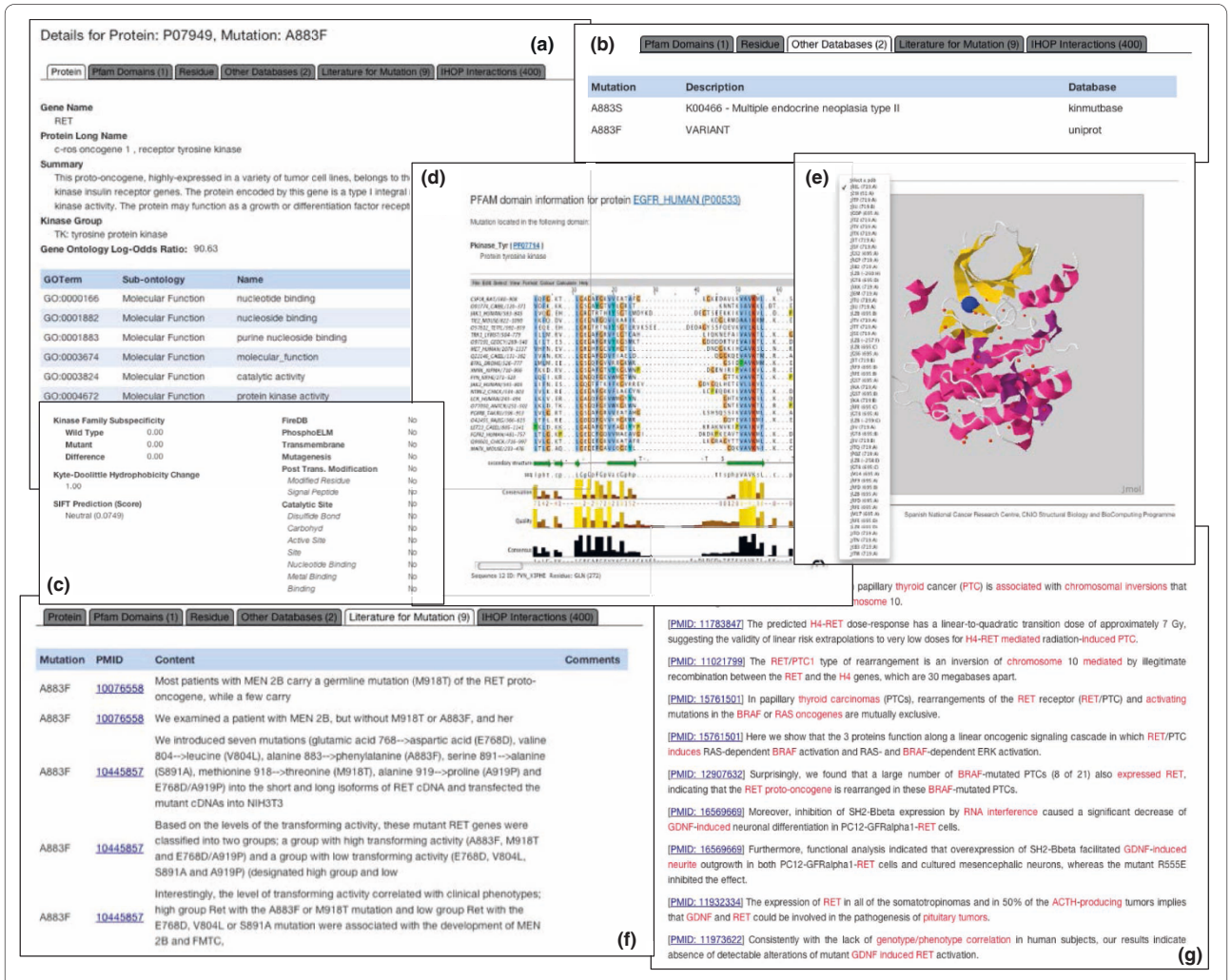

Figure 2. Screenshots representing the basic information provided by the wKinMut system for analyzing a set of point mutations in protein kinases [147,148]. The panels present: **(a)** general information about the protein kinase imported from various databases; **(b)** information about the possible consequences of the mutations extracted from annotated databases, each linked to the original source; **(c)** predictions of the consequences of the mutations in terms of the principal features of the corresponding protein kinase, including the results of the kinase-specific system KinMut [110] (Table 2); **(d)** an alignment of related sequences, including information about conserved and variable positions; **(e)** the position of the mutations in the corresponding protein structure (when available); **(f)** sentences related to the specific mutations from [77]; **(g)** information about the function and interactions of the protein kinase extracted from PubMed with the iHOP system [149,150]. A detailed description of the wKinMut system can be found in [147] and in the documentation of the web site [148].

than individual mutations themselves, combinations of mutations provide cell lineages with an advantage in terms of growth and their invasive capabilities. Given the complexity of this process, more elaborate biological models are needed to account for the role of networks of mutations in this competition between cell lineages [74].

Analyzing alterations in signaling pathways, as opposed to directly comparing mutated genes, has produced significant progress in interpreting cancer genome data [26]. In this study [119], a link between pancreatic cancer and certain specific signaling pathways was detected by carefully mapping the mutations detected in a set of cases. From this analysis, the general DNA damage pathway and several other pathways were broadly identified, highlighting the possibility of using drugs that target the proteins in these pathways to treat pancreatic cancer. Indeed, it was also relevant that the results from one patient in this study contradicted the relationship reported between pancreatic cancer and mutations in the DNA damage pathway. A manual analysis of the mutations in this patient revealed the crucial importance for treatment of a mutation in the *PALB2* gene, a gene not considered to be a component of the DNA damage pathway in the signaling database at the time of the initial analysis, even though it was clearly associated with the pathway in the scientific literature [27]. This observation

serves as an important reminder of the incomplete nature of the information organized in the current databases, the need for careful fact-checking and the difficulty in separating reactions that are naturally linked in cells into human annotated pathways.

From a systems biology viewpoint, it is clear that detecting common elements in cancer by analyzing mutations at the protein level is fraught with difficulty. Thus, shifting the analysis to the systems level by considering the pathways and cellular functions affected might offer a more general view of the relationship between mutations and phenotypes, helping to detect common biological alterations associated with specific types of cancer.

This situation was illustrated in our systematic analysis of cancer mutations and cancer types at the pathway and functional levels [120]. The associated system (Figure 3) allows the types of cancer and associated pathways to be explored, and it identifies common features in the input information (mutations obtained from small- and large-scale studies).

To overcome the limitations in defining the pathways and cell functions, as demonstrated in the study of pancreatic cancer [119], more flexible definitions of pathways and cell functions must be considered. Improvements to the main pathway information databases (that is, KEGG [121] and Reactome [122]), might be made possible by incorporating text mining systems to facilitate the task of annotation [123]. A further strategy to help detect proteins associated with specific pathways that might not have been detected by earlier biochemical approaches is to use information relating to the functional connections between proteins and genes, including gene control and protein interaction networks. For example, proteins that form complexes with other proteins in a given pathway can be considered as part of that pathway [124]. Candidates to be included in such analyses would be regulators, phosphatases and proteins with connector domains, in many cases corresponding to proteins that participate in more than one pathway and that provide a link between related cellular functions.

Even if the network- and pathway-based approaches are a clear step forward in analyzing the consequences of mutations, it is necessary to be realistic about their present limitations. Current approaches to network analysis represent static scenarios where spatial and temporal aspects are not taken into account: for example, the tissue and stage of tumor development are not considered. Furthermore, important quantitative aspects, such as the amount of proteins and the kinetic parameters of reactions, are generally not available. In other words, we still do not have at hand the comprehensive quantitative and dynamic models necessary to fully understand the consequences of mutations at the physiological level. Indeed, generating such models would require considerable experimental and computational effort, and as such it remains as one of the main challenges in systems biology today, if not the main challenge.

Linking drugs to genes/proteins and pathways

Even if comprehensive network-based approaches provide valuable information about the distribution of mutations and their possible functional consequences, they are still far from helping us reach the final objective of designing personalized cancer treatment. The final key preclinical stage is to associate the variation in proteins and pathways with drugs that directly or indirectly affect their function or activity. This is a direction that opens up a world of possibilities and may change the whole field of cancer research [125].

To go from possibilities to realities will require tools and methods that bring together the protein and pharmaceutical worlds (Table 3). The challenge is to identify proteins that when targeted by a known drug will interrupt the malfunctions in a given pathway or signaling system. This means that to identify potentially appropriate drugs, their effects must be described in different phases. First, adequate information must be compiled about the drugs and their targets in the light of our incomplete knowledge on the action *in vivo* of many drugs and the range of specificity in which many current drugs work. Second, the extent to which the effect of mutations that interrupt or overstimulate signaling pathways can be counteracted by the action of drugs must be assessed. This is a particularly difficult problem that requires an understanding of the consequences of the mutations at the network level, and the capacity to predict the appropriate levels of the network that can be used to counteract them (see above). Furthermore, the margin of operation is limited because most drugs tend to remove or diminish protein activity, as do most mutations. Hence, potential solutions will often depend on finding a node of the network that can be targeted by a drug and upregulated.

Given the limited precision of current genome analysis strategies (as described above), the large number of potential mutations and possible targets related to cancer phenotypes are difficult to disentangle. Similarly, the limited precision of the drug-protein target relationships makes reducing the genome analysis to the identification of a single potential drug almost impossible. Fortunately, the use of complementary animal models (avatar mice, see above) consistently increases the number of possible combinations of drugs that can be tested for each specific case. Perhaps the best example of the possibilities of current systems is the PharmGKB resource [126] (Table 3), which was recently used to calculate the drug

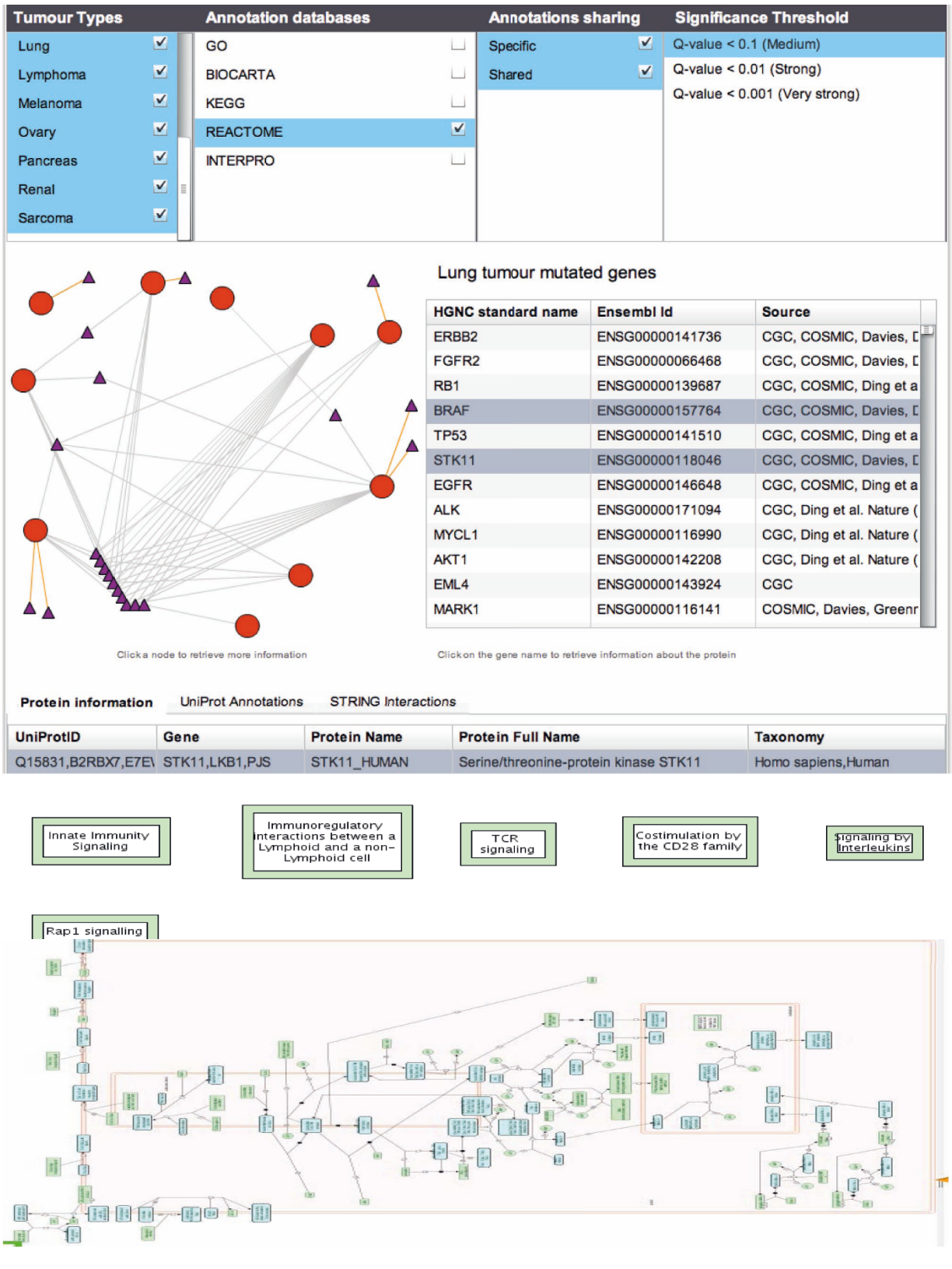

Figure 3. See overleaf for legend.

Figure 3. An interface (CONTEXTS) that we have developed for the analysis of cancer genome studies at the level of biological networks [122,151]. The upper panel shows the menus for selecting specific cancer studies, databases for pathway analysis (or set of annotations) and the level of confidence required for the relationships. From the user's requests, the system identifies the pathways or functional classes common to the different cancer studies, and the interface allows the corresponding information to be retrieved. The graph represent various cancer studies (those selected in the 'tumor types' panel are represented by red circles) using the pathways extracted from the Reactome database [152] as the background (the reference selected in the 'Annotation databases' panel and represented by small triangles). For the selected lung cancer study, the 'Lung tumor mutated genes' panel provides a link to the related genes indicating the database (source) from where the information was extracted. The lower panel represents the information on the pathways selected by the user ('innate immunity signaling') as directly provided by the Reactome database.

response probabilities after a careful analysis of the genome of a single individual [127]. Indeed, this approach provided an interesting example of the technical and organizational requirements of such an application (reviewed in [128]).

Toxicology is as an increasingly important field at the interface between genomics and disease, not least because of its influence on drug administration and its strategic importance for pharmaceutical companies. An important advance in this area will be to integrate information on mutations (and predictions of their consequences) within the context of a gene/protein, disease and drug network. In this area, the co-operation between pharmaceutical companies and research groups in the eTOX project [129] of the European 'Innovative Medicine Initiative' platform is particularly relevant (see also other IMI projects related to subjects discussed in this section [130]).

From our knowledge of disease-linked genes and protein-related drugs, the connection between toxicology and the secondary effects of drugs has been used to find associations between necrosis of breast and lung cancer [131]. Recent work has also achieved drug repositioning using analysis of expression profiles [132,133] and analyzed drug relationships using common secondary effects [134].

Conclusions and future directions

We have presented here a global vision of the issues associated with the computational analysis of personalized cancer data, describing the main limitations and possible developments of current approaches and the currently available computational systems.

The development of systems to analyze individual genome data is an ongoing activity in many groups and institutions, with diverse implementations tailored to their bioinformatics and clinical units. In the future, this type of pipeline will allow oncology units at hospitals to offer treatment for individual cancer patients based on the comparison of their normal and cancer genomic compositions with those of successfully treated patients. However, this will require the exhaustive analysis of genomic data within an analytical platform that covers the range of topics described here. Such genomic information has to be considered as an addition to the rest of the physiological and medical data that are essential for medical diagnosis.

In practice, it seems likely that the initial systems will work in research environments to explore genomic information in cases of palliative treatment and most probably in cancer relapse. Specific regulations apply in these scenarios, and the time between the initial and secondary events provides a wider time window for the

Table 3. Resources with information connecting proteins and drugs

Name	Details	URL
ChEBI (Chemical Entities of Biological Interest)	Contains more than half a million chemical compounds classified according to their biological activity	http://www.ebi.ac.uk/chebi/
DrugBank	Contains detailed chemical, pharmacological and pharmaceutical data linked with information on the sequence, structure and pathways of potential targets. The database contains information on almost 500 drugs	http://www.drugbank.ca
Resources from Peer Bork's group, including STITCH, SuperDrug, SuperNatural and SuperTarget/Matador	Bork's group has developed a number of systems that help link drugs to their protein and genomic targets, including data on adverse drug effects and symptoms	http://www.embl.de/~bork
PharmGKB	Repository linking genomic information on 2,500 genetic variants with clinical data derived from pharmacogenomics studies, and the corresponding diseases and phenotypes	http://www.pharmgkb.org
TTD (Therapeutic Target Database)	Contains data for relations between 2,000 targets and more than 15,000 drugs, including information extracted from clinical trials	http://xin.cz3.nus.edu.sg/group/ttd/ttd.asp

analysis. These systems, such as the one we use in our institution, will combine methods and results in a more flexible and exploratory set-up than will need to be implemented in regulated clinical setups. The transition from such academic software platforms will require professional software development following industrial standards, and it will need to be developed in consortia between research and commercial partners. Initiatives such as the European flagship project proposal on Information Technology Future of Medicine (ITFoM) [135] could be an appropriate vehicle to promote such developments.

The incorporation of genomic information into clinical practice will require consultation with specialists in relevant areas, including genomics, bioinformatics, systems biology, pathology and oncology. Each of the professionals involved will have their own specific requirements, and thus the driving forces for users and developers of this system will naturally differ:

(1) Clinicians, the end users of the resulting data, will require an analytical platform that is sufficiently accurate and robust to work continuously in a clinical setting. This system must be easy to understand and capable of providing validated results at each stage of the analysis.

(2) Bioinformaticians developing the analytical pipeline will require a system with a modular structure that is based on current programming paradigms and that can be easily expanded by incorporating new methods. New technology should be easy to introduce, so that the methods used can be continuously evaluated, and they should be capable of analyzing large amounts of heterogeneous data. Finally, this system will have to fulfill stringent security and confidentiality requirements.

(3) Computational biologists developing these methods will naturally be interested in the scientific issues behind each stage of the analytical platform. They will be responsible for designing new methods, and they will have to collaborate with clinicians and biologists studying the underlying biological problems (the molecular mechanisms of cancer).

A significant part of the challenge in developing personalized cancer treatments will be to ensure effective collaboration between these heterogeneous groups (for a description of the technical, practical, professional and ethical issues see [127,136]), and indeed, better training and technical facilities will be essential to facilitate such co-operation [137]. In the context of the integration of bioinformatics into clinical practice, ethical issues emerge as an essential component. The pipelines and methods described here have the capacity to reveal unexpected relationships between genomic traces and disease risks. It is currently of particular interest to define how such findings that are not directly relevant for the medical condition at hand should be dealt with - for example, the possible need to disclose this additional information to the family (such as children of the patient), as they could be affected by the mutations. For a discussion on the possible limitations of release of genome results, see [138-141].

At the very basic technical level, there are at least two key areas that must be improved to make these developments possible. Firstly, the facilities used for the rapid exchange and storage of information must become more advanced and, in some cases, additional confidentiality constraints will need to be introduced on genomic information, scientific literature, toxicology and drug-related documentation, ongoing clinical trial information and personal medical records. Secondly, adequate interfaces must be tailored to the needs of the individual professionals, which will be crucial to integrate the relevant information. User accessibility is a key issue in the context of personalized cancer treatment, as well as in bioinformatics in general.

The organization of this complex scenario is an important aspect of personalized cancer medicine, which must also include detailed discussions with patients and the need to deal with the related ethical issues, although this is beyond the scope of this review. The involvement of the general public and of patient associations will be an important step towards improved cancer treatment, presenting new and interesting challenges for bioinformaticians and computational biologists working in this area.

Abbreviations

ChIP-seq, chromatin immunoprecipitation sequencing; GWAS, genome-wide association study; NGS, next-generation sequencing; RNA-seq, RNA sequencing; SNP, single nucleotide polymorphism; TF, transcription factor.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The work in the group of AV related to this review is supported by grants from the ENCODE Project (U54 HG0004555), eTOX (Grant 115002, Innovative Medicines Initiative) and Consolider E-Science (CSD2007-00050), the Instituto de Salud Carlos III COMBIOMED (RD07/0067/0014) and the Spanish Ministry of Science and Innovation (BIO2007-66855). MH's work is supported by grants from the Spanish FIS (PI10/01996) and USA National Cancer Institute (1R01-CA129963). The assistance provided by the Spanish Bioinformatics Institute (INB), a platform of the ISCII, is particularly appreciated. We are indebted to David G Pisano, Miguel Vazquez, Victor de la Torre, Gonzalo Gomez, Enrique Carrillo, Francisco Real, Jose MG Izarzugaza, Anäis Buadot, Enrico Glaab, Michael Tress, Daniel Rico, David de Juan for interesting discussions and insights.

Published: 30 July 2012

References

1. Personal Genomics [http://en.wikipedia.org/wiki/Personal_genomics]
2. **2020 visions.** *Nature* 2010, **463:**26-32.
3. Wiseman D, Tuff A, Peach J: **UK cancer genetics gets personal.** *Nature* 2011, **475:**37.

4. Callaway E: **Cancer-gene testing ramps up.** *Nature* 2010, **467:**766-767.
5. Callaway E: **Norway to bring cancer-gene tests to the clinic.** *Nature* 2012, doi:10.1038/nature.2012.9949.
6. Baker M: **Functional genomics: The changes that count.** *Nature* 2012, **482:**257, 259-262.
7. Brunham LR, Hayden MR: **Medicine. Whole-genome sequencing: the new standard of care?** *Science* 2012, **336:**1112-1113.
8. Parkinson DR, Johnson BE, Sledge GW: **Making personalized cancer medicine a reality: challenges and opportunities in the development of biomarkers and companion diagnostics.** *Clin Cancer Res* 2012, **18:**619-624.
9. Corless CL: **Medicine. Personalized cancer diagnostics.** *Science* 2011, **334:**1217-1218.
10. Hayden EC: **Sequencing set to alter clinical landscape.** *Nature* 2012, **482:**288.
11. Golden F: *Cancer Data and the Fallacy of the $1000 Genome.* Forbes 2012 [http://www.forbes.com/sites/jimgolden/2012/06/21/cancer-data-and-the-fallacy-of-the-1000-genome/]
12. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362:**1181-1191.
13. Gui Y, Guo G, Huang Y, Hu X, Tang A, Gao S, Wu R, Chen C, Li X, Zhou L, He M, Li Z, Sun X, Jia W, Chen J, Yang S, Zhou F, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Liu C, Jiang H, Wang Y, Zheng H, Sun L, Liu X, Jiang Z, *et al.*: **Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder.** *Nat Genet* 2011, **43:**875-878.
14. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, Anderson KC, Ardlie KG, Auclair D, Baker A, Bergsagel PL, Bernstein BE, Drier Y, Fonseca R, Gabriel SB, Hofmeister CC, Jagannath S, Jakubowiak AJ, Krishnan A, Levy J, Liefeld T, Lonial S, Mahan S, Mfuko B, Monti S, Perkins LM, *et al.*: **Initial genome sequencing and analysis of multiple myeloma.** *Nature* 2011, **471:**467-472.
15. **Integrated genomic analyses of ovarian carcinoma.** *Nature* 2011, **474:**609-615.
16. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, *et al.*: **The genomic complexity of primary human prostate cancer.** *Nature* 2011, **470:**214-220.
17. Link DC, Schuettpelz LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M, Koboldt DC, Dooling DJ, Fulton RS, Bender RH, Fulton LL, Delehaunty KD, Fronick CC, Appelbaum EL, Schmidt H, Abbott R, O'Laughlin M, Chen K, McLellan MD, Varghese N, Nagarajan R, Heath S, Graubert TA, Ding L, Ley TJ, Zambetti GP, *et al.*: **Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML.** *JAMA* 2011, **305:**1568-1576.
18. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, *et al.*: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148:**1293-1307.
19. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, Escaramis G, Jares P, Beà S, González-Díaz M, Bassaganyas L, Baumann T, Juan M, López-Guerra M, Colomer D, Tubío JM, López C, Navarro A, Tornador C, Aymerich M, Rozman M, Hernández JM, Puente DA, Freije JM, Velasco G, Gutiérrez-Fernández A, Costa D, Carrió A, Guijarro S, Enjuanes A, *et al.*: **Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia.** *Nature* 2011, **475:**101-105.
20. Varela I, Tarpey P, Raine K, Huang D, Ong CK, Stephens P, Davies H, Jones D, Lin ML, Teague J, Bignell G, Butler A, Cho J, Dalgliesh GL, Galappaththige D, Greenman C, Hardy C, Jia M, Latimer C, Lau KW, Marshall J, McLaren S, Menzies A, Mudie L, Stebbings L, Largaespada DA, Wessels LF, Richard S, Kahnoski RJ, Anema J, *et al.*: **Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma.** *Nature* 2011, **469:**539-542.
21. Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J, Zhang N, El-Naggar AK, Jasser SA, Weinstein JN, Treviño L, Drummond JA, Muzny DM, Wu Y, Wood LD, Hruban RH, Westra WH, Koch WM, Califano JA, Gibbs RA, Sidransky D, Vogelstein B, Velculescu VE, Papadopoulos N, Wheeler DA, Kinzler KW, *et al.*: **Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1.** *Science* 2011, **333:**1154-1157.
22. Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T, Sakamoto H, Wang L, Ojima H, Shimada K, Kosuge T, Okusaka T, Kato K, Kusuda J, Yoshida T, Aburatani H, Shibata T: **High-resolution characterization of a hepatocellular carcinoma genome.** *Nat Genet* 2011, **43:**464-469.
23. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, Shi JY, Zhu YM, Tang L, Zhang XW, Liang WX, Mi JQ, Song HD, Li KQ, Chen Z, Chen SJ: **Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia.** *Nat Genet* 2011, **43:**309-315.
24. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, Ramsay AJ, Beà S, Pinyol M, Martínez-Trillos A, López-Guerra M, Colomer D, Navarro A, Baumann T, Aymerich M, Rozman M, Delgado J, Giné E, Hernández JM, González-Díaz M, Puente DA, Velasco G, Freije JM, Tubío JM, Royo R, Gelpí JL, Orozco M, Pisano DG, Zamora J, Vázquez M, *et al.*: **Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia.** *Nat Genet* 2012, **44:**47-52.
25. Hidalgo M, Bruckheimer E, Rajeshkumar NV, Garrido-Laguna I, De Oliveira E, Rubio-Viqueira B, Strawn S, Wick MJ, Martell J, Sidransky D: **A pilot clinical study of treatment guided by personalized tumorgrafts in patients with advanced cancer.** *Mol Cancer Ther* 2011, **10:**1311-1316.
26. 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, *et al.*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321:**1801-1806.
27. Zhang F, Ma J, Wu J, Ye L, Cai H, Xia B, Yu X: **PALB2 links BRCA1 and BRCA2 in the DNA-damage response.** *Curr Biol* 2009, **19:**524-529.
28. Lifton RP: **Individual genomes on the horizon.** *N Engl J Med* 2010, **362:**1235-1236.
29. Navin N, Hicks J: **Future medical applications of single-cell sequencing in cancer.** *Genome Med* 2011, **3:**31.
30. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, Muthuswamy L, Krasnitz A, McCombie WR, Hicks J, Wigler M: **Tumour evolution inferred by single-cell sequencing.** *Nature* 2011, **472:**9094
31. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang LE, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, *et al.*: **Detectable clonal mosaicism from birth to old age and its relationship to cancer.** *Nat Genet* 2012, **44:**642-650.
32. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, Cullen M, Epstein CG, Burdett L, Dean MC, Chatterjee N, Sampson J, Chung CC, Kovaks J, Gapstur SM, Stevens VL, Teras LT, Gaudet MM, Albanes D, Weinstein SJ, Virtamo J, Taylor PR, Freedman ND, Abnet CC, Goldstein AM, Hu N, *et al.*: **Detectable clonal mosaicism and its relationship to aging and cancer.** *Nat Genet* 2012, **44:**651-658.
33. Groesser L, Herschberger E, Ruetten A, Ruivenkamp C, Lopriore E, Zutt M, Langmann T, Singer S, Klingseisen L, Schneider-Brachert W, Toll A, Real FX, Landthaler M, Hafner C: **Postzygotic HRAS and KRAS mutations cause nevus sebaceous and Schimmelpenning syndrome.** *Nat Genet* 2012, **44:**783-787.
34. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, Ruzzo EK, Singh A, Campbell CR, Hong LK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB: **The characterization of twenty sequenced human genomes.** *PLoS Genet* 2010, **6:**e1001111.
35. 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.
36. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.
37. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T,

Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461:**272-276.

38. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ: **Exome sequencing identifies the cause of a mendelian disorder.** *Nat Genet* 2010, **42:**30-35
39. Summerer D, Schracke N, Wu H, Cheng Y, Bau S, Stähler CF, Stähler PF, Beier M: **Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform.** *Genomics* 2010, **95:**241-246.
40. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328:**636-639.
41. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318:**420-426.
42. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.
43. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H, Ma H, Zhang F, Feng S, Zhang W, Du H, Tian G, Li J, Zhang X, Li S, Bolund L, Kristiansen K, de Smith AJ, Blakemore AI, Coin LJ, Yang H, Wang J, Wang J: **Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly.** *Nat Biotechnol* 2011, **29:**723-730.
44. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation.** *Nat Methods* 2009, **6:**677-681.
45. Medvedev P, Stanciu M, Brudno M: **Computational methods for discovering structural variation with next-generation sequencing.** *Nat Methods* 2009, **6:**S13-S20.
46. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC: **Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.** *Genome Res* 2009, **19:**1270-1278.
47. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J: **CREST maps somatic structural variation in cancer genomes with base-pair resolution.** *Nat Methods* 2011, **8:**652-654.
48. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.
49. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.** *Nature* 2008, **453:**1239-1243.
50. Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert JJ, Goldstein DB: **Screening the human exome: a comparison of whole genome and whole transcriptome sequencing.** *Genome Biol* 2010, **11:**R57.
51. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA: **Alternative expression analysis by RNA sequencing.** *Nat Methods* 2010, **7:**843-847.
52. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.
53. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464:**773-777.
54. Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, Howald C, Foissac S, Ucla C, Chrast J, Ribeca P, Martin D, Murray RR, Yang X, Ghamsari L, Lin C, Bell I, Dumais E, Drenkow J, Tress ML, Gelpí JL, Orozco M, Valencia A, van Berkum NL, Lajoie BR, Vidal M, Stamatoyannopoulos J, Batut P, Dobin A, Harrow J, *et al.*: **Evidence for transcript networks composed of chimeric RNAs in human cells.** *PLoS One* 2012, **7:**e28213.
55. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, Del Pozo A, Tress M, Johnson R, Guigo R, Valencia A: **Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts.** *Genome Res* 2012, **22:**1231-1242.
56. Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L: **Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing.** *Proc Natl Acad Sci U S A* 2011, **108:**9172-9177.
57. David CJ, Manley JL: **Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged.** *Genes Dev* 2010, **24:**2343-2364.
58. Mori J, Takahashi Y, Tanimoto T: **SF3B1 in chronic lymphocytic leukemia.** *N Engl J Med* 2012, **366:**1057; author reply 1057-1058.
59. He L, He X, Lowe SW, Hannon GJ: **microRNAs join the p53 network--another piece in the tumour-suppression puzzle.** *Nat Rev Cancer* 2007, **7:**819-822.
60. Uziel T, Karginov FV, Xie S, Parker JS, Wang YD, Gajjar A, He L, Ellison D, Gilbertson RJ, Hannon G, Roussel MF: **The miR-17~92 cluster collaborates with the Sonic Hedgehog pathway in medulloblastoma.** *Proc Natl Acad Sci U S A* 2009, **106:**2812-2817.
61. Wyman SK, Parkin RK, Mitchell PS, Fritz BR, O'Briant K, Godwin AK, Urban N, Drescher CW, Knudsen BS, Tewari M: **Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries.** *PLoS One* 2009, **4:**e5311.
62. Nygaard S, Jacobsen A, Lindow M, Eriksen J, Balslev E, Flyger H, Tolstrup N, Møller S, Krogh A, Litman T: **Identification and analysis of miRNAs in human breast cancer and teratoma samples using deep sequencing.** *BMC Med Genomics* 2009, **2:**35.
63. Dalgliesh GL, Furge K, Greenman C, Chen L, Bignell G, Butler A, Davies H, Edkins S, Hardy C, Latimer C, Teague J, Andrews J, Barthorpe S, Beare D, Buck G, Campbell PJ, Forbes S, Jia M, Jones D, Knott H, Kok CY, Lau KW, Leroy C, Lin ML, McBride DJ, Maddison M, Maguire S, McLay K, Menzies A, Mironenko T, *et al.*: **Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes.** *Nature* 2010, **463:**360-363.
64. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP: **Increased methylation variation in epigenetic domains across cancer types.** *Nat Genet* 2011, **43:**768-775.
65. Martín-Subero JI, Esteller M: **Profiling epigenetic alterations in disease.** *Adv Exp Med Biol* 2011, **711:**162-177.
66. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, McLaren S, Lin ML, McBride DJ, Varela I, Nik-Zainal S, Leroy C, Jia M, Menzies A, Butler AP, Teague JW, Quail MA, Burton J, Swerdlow H, Carter NP, Morsberger LA, Iacobuzio-Donahue C, Follows GA, Green AR, Flanagan AM, Stratton MR, *et al.*: **Massive genomic rearrangement acquired in a single catastrophic event during cancer development.** *Cell* 2011, **144:**27-40.
67. Rausch T, Jones DT, Zapatka M, Stütz AM, Zichner T, Weischenfeldt J, Jäger N, Remke M, Shih D, Northcott PA, Pfaff E, Tica J, Wang Q, Massimi L, Witt H, Bender S, Pleier S, Cin H, Hawkins C, Beck C, von Deimling A, Hans V, Brors B, Eils R, Scheurlen W, Blake J, Benes V, Kulozik AE, Witt O, Martin D, *et al.*: **Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations.** *Cell* 2012, **148:**59-71.
68. Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, Nezi L, Protopopov A, Chowdhury D, Pellman D: **DNA breaks and chromosome pulverization from errors in mitosis.** *Nature* 2012, **482:**5358.
69. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.
70. **A Catalog of Published Genome-Wide Association Studies** [http://www.genome.gov/gwastudies]
71. Goldstein DB: **Common genetic variation and human traits.** *N Engl J Med* 2009, **360:**1696-1698.
72. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** *Nat Rev Genet* 2008, **9:**356-369.
73. Andrewes CH: **The complex epidemiology of respiratory virus infections.** *Science* 1964, **146:**1274-1277.
74. Baudot A, Real FX, Izarzugaza JM, Valencia A: **From cancer genomes to cancer models: bridging the gaps.** *EMBO Rep* 2009, **10:**359-366.
75. Greaves M, Maley CC: **Clonal evolution in cancer.** *Nature* 2012, **481:**306-313.

76. Baker CJ, Rebholz-Schuhmann D: **Between proteins and phenotypes: annotation and interpretation of mutations.** *BMC Bioinformatics* 2009, **10 Suppl 8:**I1.
77. Krallinger M, Izarzugaza JM, Rodriguez-Penagos C, Valencia A: **Extraction of human kinase mutations from literature, databases and genotyping studies.** *BMC Bioinformatics* 2009, **10 Suppl 8:**S1.
78. Nagel K, Jimeno-Yepes A, Rebholz-Schuhmann D: **Annotation of protein residues based on a literature analysis: cross-validation against UniProtKb.** *BMC Bioinformatics* 2009, **10 Suppl 8:**S4.
79. Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H: **Automatic extraction of mutations from Medline and cross-validation with OMIM.** *Nucleic Acids Res* 2004, **32:**135-142.
80. Doughty E, Kertesz-Farkas A, Bodenreider O, Thompson G, Adadey A, Peterson T, Kann MG: **Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature.** *Bioinformatics* 2011, **27:**408-415.
81. Riazanov A, Laurila JB, Baker CJ: **Deploying mutation impact text-mining software with the SADI Semantic Web Services framework.** *BMC Bioinformatics* 2011, **12 Suppl 4:**S6.
82. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics* 2007, **23:**1862-1865.
83. Garten Y, Altman RB: **Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text.** *BMC Bioinformatics* 2009, **10 Suppl 2:**S6.
84. Rance B, Doughty E, Demner-Fushman D, Kann MG, Bodenreider O: **A mutation-centric approach to identifying pharmacogenomic relations in text.** *J Biomed Inform* 2012, doi: 10.1016/j.jbi.2012.05.003.
85. Valencia A: **Automatic annotation of protein function.** *Curr Opin Struct Biol* 2005, **15:**267-274.
86. Godzik A, Jambon M, Friedberg I: **Computational protein function prediction: are we making progress?** *Cell Mol Life Sci* 2007, **64:**2505-2511.
87. Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, Sander C: **GeneQuiz: a workbench for sequence analysis.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**348-353.
88. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6 Suppl 1:**S1.
89. Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A: **The FEBS Letters/BioCreative II.5 experiment: making biological information accessible.** *Nat Biotechnol* 2010, **28:**897-899.
90. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A: **Critical assessment of methods of protein structure prediction - Round VIII.** *Proteins* 2009, **77 Suppl 9:**1-4.
91. **Automated Function Prediction SIG 2012: Critical Assessment of Function Annotations** [http://biofunctionprediction.org]
92. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A: **UniProtKB/Swiss-Prot.** *Methods Mol Biol* 2007, **406:**89-112.
93. Lopez G, Valencia A, Tress M: **FireDB--a database of functionally important residues from proteins of known structure.** *Nucleic Acids Res* 2007, **35:**D219-D223.
94. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32:**D129-D133.
95. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state.** *J Mol Biol* 2007, **372:**774-797.
96. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N: **ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.** *Nucleic Acids Res* 2010, **38:**W529-W533.
97. Ward RM, Venner E, Daines B, Murray S, Erdin S, Kristensen DM, Lichtarge O: **Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates.** *Bioinformatics* 2009, **25:**1426-1427.
98. Reva B, Antipin Y, Sander C: **Determinants of protein function revealed by combinatorial entropy optimization.** *Genome Biol* 2007, **8:**R232.
99. Rausell A, Juan D, Pazos F, Valencia A: **Protein interactions and ligand binding: from protein subfamilies to functional specificity.** *Proc Natl Acad Sci U S A* 2010, **107:**1995-2000.
100. Muth T, García-Martín JA, Rausell A, Juan D, Valencia A, Pazos F: **JDet: interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and structures.** *Bioinformatics* 2012, **28:**584-586.
101. Baldi P, Brunak S: *Bioinformatics, The Machine Learning Approach*. Cambridge, MA: MIT Press; 2011.
102. 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* 2011, **88:**440-449.
103. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M: **PMUT: a web-based tool for the annotation of pathological mutations on proteins.** *Bioinformatics* 2005, **21:**3176-3178.
104. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.
105. Bromberg Y, Capriotti E: **SNP-SIG Meeting 2011: Identification and annotation of SNPs in the context of structure, function, and disease.** *BMC Genomics* 2012, **13 Suppl 4:**S1. [http://snps.uib.es/snp-sig/]
106. Callaway E: **Mutation-prediction software rewarded.** *Nature* 2010 doi:10.1038/news.2010.679.
107. Knudsen M, Wiuf C: **The CATH database.** *Hum Genomics* 2010, **4:**207-212.
108. Izarzugaza JM, Baresic A, McMillan LE, Yeats C, Clegg AB, Orengo CA, Martin AC, Valencia A: **An integrated approach to the interpretation of single amino acid polymorphisms within the framework of CATH and Gene3D.** *BMC Bioinformatics* 2009, **10 Suppl 8:**S5.
109. Izarzugaza JM, Redfern OC, Orengo CA, Valencia A: **Cancer-associated mutations are preferentially distributed in protein kinase functional sites.** *Proteins* 2009, **77:**892-903.
110. Izarzugaza JMG, del Pozo A, Vazquez M, Valencia A: **Prioritization of pathogenic mutations in the protein kinase superfamily.** *BMC Genomics* 2012, **13 Suppl 4:**S3.
111. Torkamani A, Schork NJ: **Distribution analysis of nonsynonymous polymorphisms within the human kinase gene family.** *Genomics* 2007, **90:**49-58.
112. Reva B, Antipin Y, Sander C: **Predicting the functional impact of protein mutations: application to cancer genomics.** *Nucleic Acids Res* 2011, **39:**e118.
113. **CBS Prediction Server** [http://www.cbs.dtu.dk/services/]
114. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet* 2009, **10:**252-263.
115. Wingender E, Dietze P, Karas H, Knüppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24:**238-241.
116. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-D94.
117. Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJ: **ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation.** *Bioinformatics* 2006, **22:**637-640.
118. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10:**155-159.
119. Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E, Brune K, Jaffee EM, Iacobuzio-Donahue CA, Maitra A, Parmigiani G, Kern SE, Velculescu VE, Kinzler KW, Vogelstein B, Eshleman JR, Goggins M, Klein AP: **Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene.** *Science* 2009, **324:**217.
120. Baudot A, de la Torre V, Valencia A: **Mutated genes, pathways and processes in tumours.** *EMBO Rep* 2010, **11:**805-810.
121. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38:**D355-D360.
122. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L: **Reactome: a database of reactions, pathways and biological processes.** *Nucleic Acids Res* 2011, **39:**D691-D697.
123. Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, Hirschman L, Cesareni G, Valencia A: **The FEBS Letters/BioCreative II.5 experiment: making biological information accessible.** *Nat Biotechnol* 2010, **28:**897-899.
124. Glaab E, Baudot A, Krasnogor N, Valencia A: **Extending pathways and processes using molecular interaction networks to analyse cancer genome data.** *BMC Bioinformatics* 2010, **11:**597.

125. Goldstein I, Madar S, Rotter V: **Cancer research, a field on the verge of a paradigm shift?** *Trends Mol Med* 2012, **18:**299-303.
126. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE: **PharmGKB: the Pharmacogenetics Knowledge Base.** *Nucleic Acids Res* 2002, **30:**163-165.
127. Ormond KE, Wheeler MT, Hudgins L, Klein TE, Butte AJ, Altman RB, Ashley EA, Greely HT: **Challenges in the clinical application of whole-genome sequencing.** *Lancet* 2010, **375:**1749-1751.
128. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB: **Bioinformatics challenges for personalized medicine.** *Bioinformatics* 2011, **27:**1741-1748.
129. **eTOX** [http://www.etoxproject.eu]
130. **The Innovative Medicines Initiative** [http://www.imi.europa.eu]
131. Audouze K, Juncker AS, Roque FJ, Krysiak-Baltyn K, Weinhold N, Taboureau O, Jensen TS, Brunak S: **Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks.** *PLoS Comput Biol* 2010, **6:**e1000788.
132. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A, di Bernardo D: **Discovery of drug mode of action and drug repositioning from transcriptional responses.** *Proc Natl Acad Sci U S A* 2010, **107:**14621-14626.
133. Dudley JT, Deshpande T, Butte AJ: **Exploiting drug-disease relationships for computational drug repositioning.** *Brief Bioinform* 2011, **12:**303-311.
134. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P: **A side effect resource to capture phenotypic effects of drugs.** *Mol Syst Biol* 2010, **6:**343.
135. **Information Technology Future of Medicine** [http://www.itfom.eu]
136. Beskow LM, Linney KN, Radtke RA, Heinzen EL, Goldstein DB: **Ethical challenges in genotype-driven research recruitment.** *Genome Res* 2010, **20:**705-709.
137. Tramontano A, Valencia A: *Education and Research Infrastructures* in *Cancer Systems Biology: Bioinformatics and Medicine: Research and Clinical Applications.* Edited by Cesario A, Marcus FB. Springer: 2011.
138. Hayden EC: **Secrets of the human genome disclosed.** *Nature* 2011, **478:**17.
139. Zawati MH, Hendy M, Joly Y: **Incidental findings in data-intensive postgenomics science and legal liability of clinician-researchers: ready for vaccinomics?** *OMICS* 2011, **15:**615-624.
140. Bookman EB, Langehorne AA, Eckfeldt JH, Glass KC, Jarvik GP, Klag M, Koski G, Motulsky A, Wilfond B, Manolio TA, Fabsitz RR, Luepker RV: **Comment on "Multidimensional results reporting to participants in genomic studies: getting it right".** *Sci Transl Med* 2011, **3:**70le1.
141. Kohane IS, Taylor PL: **Multidimensional results reporting to participants in genomic studies: getting it right.** *Sci Transl Med* 2010, **2:**37cm19.
142. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
143. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, *et al.*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35:**D610-D617.
144. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33:**D514-D517.
145. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC).** *Curr Protoc Hum Genet* 2008, **Chapter 10:**Unit 10.11.
146. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4:**177-183.
147. Izarzugaza JM, Krallinger M, Valencia A: **Interpretation of the consequences of mutations in protein kinases: combined use of bioinformatics and text mining.** *Frontiers Systems Physiol* 2012 31.
148. **wKinMut** [http://wkinmut.bioinfo.cnio.es]
149. Fernández JM, Hoffmann R, Valencia A: **iHOP web services.** *Nucleic Acids Res* 2007, **35:**W21-W26.
150. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet* 2004 **36:** 664.
151. **CONTEXTS** [http://contexts.bioinfo.cnio.es/]
152. **Reactome** [http://www.reactome.org]

doi:10.1186/gm362
Cite this article as: Valencia A, Hidalgo M: **Getting personalized cancer genome analysis into the clinic: the challenges in bioinformatics.** *Genome Medicine* 2012, **4:**61.