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
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.
Bioinformatics steps for personalized genome analysis

| Principal developments | Key problems |
|------------------------|--------------|
| (1) Genome analysis: DNA and RNA sequences from different experimental sources and using various technologies and platforms | Methods for the prediction of consequences of mutations at protein family level |
|                       | Methods for the prediction of binding sites for transcription factor and miRNAs |
|                       | Analysis of alterations in splice sites and splice factors |
|                       | Prediction of the consequences of alterations in epigenetic markers |
| (2) Consequences of mutations and genomic alterations: Mutations in coding and non-coding regions, changes in gene expression, genome structure alterations | Databases and analysis platforms for pathways and networks |
|                       | Systems linking disease and symptoms to molecular entities |
|                       | Emerging simulations of biological networks and pathways |
| (3) Network level analysis: metabolic and signaling pathways, gene control networks and functional classes | Public repositories of drugs and small molecules |
|                       | Systems linking protein networks with drug targets |
| (4) Drug related to proteins and pathways | Initial systems able to display genomic information and complex analysis in biomedical/clinical environments |
| (5) Collaborative interfaces | Data provenance and traceability |
|                       | Robustness and performance |
|                       | Adequate information for users and different levels of access |

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 in the middle row, and the top 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.
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 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                                             | [142]                                                    |
| Ensembl   | Maps known mutations and SNPs in the human genome from other databases      | [143]                                                    |
| OMIM      | Online Mendelian Inheritance in Man; a large collection of disease annotations, often for monogenetic diseases | [144]                                                    |
| COSMIC    | Catalog of somatic mutations in cancer                                       | [145]                                                    |
| CGC       | Cancer Gene Census                                                           | [146]                                                    |
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).

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.pcbb.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.cno.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 |

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 Sören 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
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 scientific literature [27].

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].
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.
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 cooperation 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].

**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/tdt/tdt.asp |

**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...
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 (US4 HG004555), 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). MR’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 ISCIII, 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, Anaís Buadot, Enrico Glaia, 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, 475:57.
4. Callaway E. Cancer-genome testing ramps up. Nature 2010, 467:766-767.

5. Callaway E. Norway to bring cancer-genome 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.

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(9):619-624.

9. Corless CL. Medicine. Personalized cancer diagnostics. Science 2011, 334:1117-1219.

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/sites/jmgolden/2012/06/21/cancer-data-and-the-fallacy-of-the-1000-genome/]

12. Lupski JR, Reid RG, Gonzaga-Jauregui C, Rio Derosi C, Chen DC, Nazareth L, Barrière M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irkat RK, Tom W, Fanta N, Muzny DM. GB: Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. Neuropathol. 2011, 406:1181-1191.

13. Gui Y, Guo G, Huang Y, Xu T, 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 T, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Li 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 Med 2011, 17:875-878.

14. Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harvick CL, Burnett KN, Aalipour GJ, Algazi A, Dooling DJ, Fulton RS, McCarroll SA, McCombie WR, Mardis ER, Van Veldhuisen S, Wendl MC, Wilson RK, et al.: Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of multiple myeloma. Nature 2011, 471:473.

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, Stomer A, Esquela R, Pfleger D, Sougnez C, Onofrio R, Carter SL, Park K, Hruban RH, Dooling DJ, Fulton RS, McCombie WR, Mardis ER, et al.: Initial genome sequencing and functional genomics: The changes that count. Science 2012, 333:1130-1134.

17. Link DC, Schuettelpfel LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M, Kozakiewicz K, McCarroll SA, Varambally R, Chen C, Li X, Zhou L, He M, 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 T, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Li 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 Med 2011, 17:875-878.

18. Link DC, Schuettelpfel LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M, Kozakiewicz K, McCarroll SA, Varambally R, Chen C, Li X, Zhou L, He M, 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 T, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Li 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 Med 2011, 17:875-878.

19. Link DC, Schuettelpfel LG, Shen D, Wang J, Walter MJ, Kulkarni S, Payton JE, Ivanovich J, Goodfellow PJ, Le Beau M, Kozakiewicz K, McCarroll SA, Varambally R, Chen C, Li X, Zhou L, He M, 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 T, Zhao X, Wan S, Ye R, Liang C, Liu Z, Huang P, Li 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 Med 2011, 17:875-878.

20. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

21. Quesada V, Conde L, Villamar N, Órdoñez GR, Jares I, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Hruban RH, Dooling DJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

22. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

23. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

24. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

25. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

26. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

27. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

28. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

29. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

30. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

31. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

32. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.

33. Snijders AM, Cibulskis K, Marra MA, Zeng Z, Dooling DJ, Packard RJ, Fulton RS, Mardis ER, et al.: Whole-genome sequencing of the DNA methyltransferase gene Dnmt3a in acute monocytic leukemia. Nat Genet 2011, 43:309-315.
Wong M, Bhattacharjee A, Eichler EE, Barnshaw 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, Labs EW, Nickerson DA, Shendure J, Barnshaw MJ. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010, 42:30-35.

39. Summerer D, Schracker 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, Gluskin G, Smits AF, Hubley R, Shannon PT, Rowe L, Pant KP, Goodman N, Barnshaw M, Shendure J, Drmanac R, Jordne 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, Simonis JF, Kim PM, Palejev D, Carnero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Ch I, Yang F, Carter NP, Hurles ME, Weissman SM, Hinks TJ, 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, Klingenhofer A, Scherf M, Siefert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt M, O’Keefe S, Haas S, Vingron M, Lehach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2009, 323:944-948.

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, Wu D, 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, 29:725-730.

44. Chen K, Walls J, McAllan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendt MC, Zhang Q, Locke DP, Shi X, Fulton RS, Levy TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009, 6:771-781.

45. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 2009, 6:513-520.

46. Hormozdiani 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, Lu L, Dereski H, Lander ES, Hehir K, Shete S, Wang X, Zhou B, Zhao D, Emberson JL, Neale BM, Nicol D, Obenaauer JC, Naive C, Mardis ER, Wilson RK, Downing JR, Zhang J. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 2011, 8:552-654.

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, 453:1219-1243.

50. Cirulli ET, Singh A, Shanaa KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert MJ, Hirschhorn JN. [http://www.genome.gov/gwastudies] Genome-wide association loci for human diseases and traits. Nat Rev Cancer 2010, 10:936-937.
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: PharmoGKB: 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, Capritti E, Daneshouj R, Karczewski KJ, Altman RB: Bioinformatics challenges for personalized medicine. Bioinformatics 2011, 27(17):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, Kysiai-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(1):1000788.
132. Iorio F, Bosotti R, Scaheri E, Belcastro V, Mithbaokar P, Ferrero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Iaccoci 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, Heinen 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 MJ, Joly Y: Incidental findings in data-intensive postgenomics studies and legal liability of clinician-researchers: ready for vaccinomics? OMICS 2011, 15:615-624.
140. Bookman EB, Langeherne AA, Eckfeldt JH, Glass KC, Jarvik GP, Klag M, Koski G, Motulsky A, Wilford B, Manolio TA, Fabritz 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, Cuthts 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, Kiarath A, Keefe D, Kokocinski F, Kuleshova E, Lawson D, Longden I, Metzopp 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. Izarzegua JM, Kralinger 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.cnsio.eu/]
152. Reactome (http://www.reactome.org/)