Genomics 101 (2013) 75–85

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Database tools in genetic diseases research

Anna Monica Bianco [a,*], Annalisa Marcuzzi [a], Valentina Zanin [a], Martina Girardelli [a], Josef Vuch [a], Sergio Crovella [a,b]

[a] Institute for Maternal and Child Health – IRCCS "Burlo Garofolo," Trieste, Italy
[b] University of Trieste, Italy

ARTICLE INFO

Article history:
Received 19 March 2012
Accepted 1 November 2012
Available online 10 November 2012

Keywords:
Genetic diseases
Mutation
NCBI
Ensembl
SNPs
GWAS
IBD

ABSTRACT

The knowledge of the human genome is in continuous progression: a large number of databases have been developed to make meaningful connections among worldwide scientific discoveries. This paper reviews bioinformatics resources and database tools specialized in disseminating information regarding genetic disorders. The databases described are useful for managing sample sequences, gene expression and post-transcriptional regulation. In relation to data sets available from genome-wide association studies, we describe databases that could be the starting point for developing studies in the field of complex diseases, particularly those in which the causal genes are difficult to identify.

Contents

Abbreviations: NAR, Nucleic Acids Research; OMIM, Online Mendelian Inheritance in Man; GWAS, Genome Wide Association Studies; CDD, Conserved Domain Database; dbEST, Database of Expressed Sequence Tags; dbVar, Database of Genomic Structural Variation; dbGap, Database of Genotypes and Phenotypes; dbMHC, Database of Major Histocompatibility Complex; SBT, Sequencing Based Typing; dbSNP, Database of Short Genetic Variation; UCSC, Genome Browser of the University of California Santa Cruz; EMBL, European Molecular Biology Laboratory; EBI, European Bioinformatics Institute; TF, transcription factor; EST, Expression Sequence Tag; LD, linkage disequilibrium; GV, genetic variation; PPI, protein–protein interaction; IBD, inflammatory bowel disease; NOD2, nucleotide-binding oligomerization domain-containing protein 2; DECIPHER, Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources; HGNC, Hugo Gene Nomenclature Committee; DGV, Database of Genomic Variants; SNP, single nucleotide polymorphism.

* Corresponding author at: Institute for Maternal and Child Health – IRCCS "Burlo Garofolo," Via dell'Istria 65/1, 34137 Trieste, Italy. Tel./fax: +39 040 3785 422.
E-mail address: bianco@burlo.trieste.it (A.M. Bianco).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.11.001

1. Introduction

The speed of information progress and data exchange is a factor closely related to human progress. Databases are essential for storing while constantly updating and comparing data. Integrating new discoveries into existing databases makes them not only depositors and guardians of our science, but also key elements for the progress of scientific research.

Specific databases, including those related to "omics" (genomics, proteomics and metabolomics), collect experimental data and can be browsed with specifically designed software.

The Genome Online Database (GOLD) is an important resource for continuous centralized monitoring of genome and metagenome projects worldwide [1–3]. It plays a central role in bioinformatics by offering access to a wide variety of biological access. Millions of amino acids sequences, an increasing amount of genomic data, require a systematic approach to organize the information about their 3D structure, function and evolution. Numerous studies demonstrated the functional importance of the Protein Tandem Repeats and their involvement in human genetic diseases. These tandem repeats are considerably diverse, ranging from the repetition of a single amino acid to domains of 100 or more residues. They are frequently not perfect, containing a number of mutations (substitutions, insertions, deletions) accumulated during evolution, and some of them cannot be easily identified. To solve this problem, over the last few years, several improved software and particular Databases have been developed such as the Protein Tandem Repeat Database (PRDB) available at http://bioinfo.montp.cnrs.fr/?r=repeatDB that is a tool for large scale analysis of tandem repeats in proteins [4].

Biological databases offer access to a wide variety of biological data and play a central role in bioinformatics [5–7]. A biological database is an organized set of information and data from studies developed in research laboratories (both *in vitro* and *in vivo*), from bioinformatics (*in silico*) analysis and scientific publications. Databases consist of "entries" where data is structured and organized so it is both available and easy to use. First level Databases collect nucleotide (DNA, RNA) or amino acid (proteins) sequences and contain useful information for the identification of the species from which the sequences were obtained, as well as the related functions. The deoxyribonucleic acid (DNA) benefited from these technologies so that the human genetic map was represented by the earliest printed illustrations to the latest projections in 3D. Specialist databases gather more specific information on the taxonomy, the functions, the scientific publications and the diseases related to mutations of the nucleotide sequences.

Every year the Journal Nucleic Acids Research (NAR) publishes an article on all new biological databases and includes them in a list online. The 2012 19th edition includes 92 online databases covering a variety of molecular biological data, and 100 other papers concerning this topic. The NAR online Molecular Biology Database collection, available at http://www.oxfordjournals.org/nar/database/a/, after the update, includes 1380 databases sorted into 14 categories and 41 subcategories [8,9].

The aim of this manuscript is to describe databases useful for human genetics, and help the information search about genetic diseases, sample sequence, mutations, gene expression and protein expression data.

In monogenic diseases a mutation in a single gene is responsible for the disease. Single gene diseases occur in families and can be dominant or recessive, autosomal or sex-linked. When genome sequences started to be publicly available, researchers began to shift their focus from monogenic to polygenic and complex diseases, more frequent in the general population and involving several genes. The generation of widespread markers of genetic variation, the development of both new technologies and databases, allows investigators to associate the disease phenotype to genetic loci more easily. All data collected are stored and classified in a myriad of databases (Fig. 1). We describe a few bioinformatics tools useful in biomedical research of genetic diseases described in this issue.

2. Databases

2.1. Biomedical literature database

The first step in a scientific research project, particularly if it involves designing of an experiment, is the investigation of biomedical literature. The exponential growth in the number of published biomedical articles, together with the improved access to internet resources, has made data networking mandatory [10].

The oldest scientific article database is PubMed, inaugurated in 1997 (accessible online, developed by NCBI), and includes medical and biological abstracts and links to the papers [12]. PubMed is a part of NCBI's Entrez retrieval system, which provides access to a diverse set of 38 databases [13] and it provides a crucial bridge between the data of molecular biology, genetics and scientific literature. Literature search and analysis is the basis of any project and/or experiment because, by making available the information on a particular gene or specific disease, it allows us to formulate new hypotheses on which to work and further develop scientific reasoning [11]. Access to database is freely accessible for all users. Most citations in PubMed are about biologically relevant subjects (e.g. gene or disease), but since biomedical literature topics require much broader coverage it includes also a number of interdisciplinary subjects [12]. PubMed is the most powerful and updated site for bibliographic research in biomedicine; it also includes web-based systems such as PubMed Assistant [14], Alibaba [15] and PubMed-Ex [16]. PubMed Assistant provides useful functions such as keyword highlighting and easy export to citation managers, while both AliBaba and PubMed-EX are geared towards semantic enrichment by identifying gene, protein, disease and other biomedical entities from the text.

2.2. Common databases in genetics

Scientific research progress in several biomedical fields continuously contributes to supporting patients. The knowledge of human genome and all of its components is in constant evolution. Databases have been created to make meaningful connections among worldwide scientific discoveries, in a very clear way. The availability of information is necessary for answering several questions about the many systems and subsystems that are the basis of different mechanisms with the common base of the DNA.

The most common databases used in genetics and molecular biology are gene sequences and protein databases. Databases of gene sequences are divided in two types: primary databases which contain experimental results directly into the database without additional information and secondary databases, which combine findings from primary databases with other data such as sequence information on the species from which it derives, and gene variants. Concerning biomedical research and the analysis of rare genetic diseases, the most powerful are the General Databases where users can find comprehensive information about a specific illness and/or a disease gene [17].

2.2.1. National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health was created in 1988 to develop information systems for many resources that can be accessed through the NCBI home page at www.ncbi.nlm.nih.gov. This complex database receives data from three sources: direct submissions from external investigators, internal collecting efforts and collaborations or agreements, with data providers and research consortia (both national and international). Within NCBI operates the Online Mendelian Inheritance in Man (OMIM) database [18,19], a catalog of human genes and genetics disorders; it contains information about linkage data, phenotypes and references on all inherited or heritable human known disorders. The OMIM comprises about 4613 diseases, 367 genes with an associated phenotype and 1317 genes (including 183 microRNAs) with known sequence. The

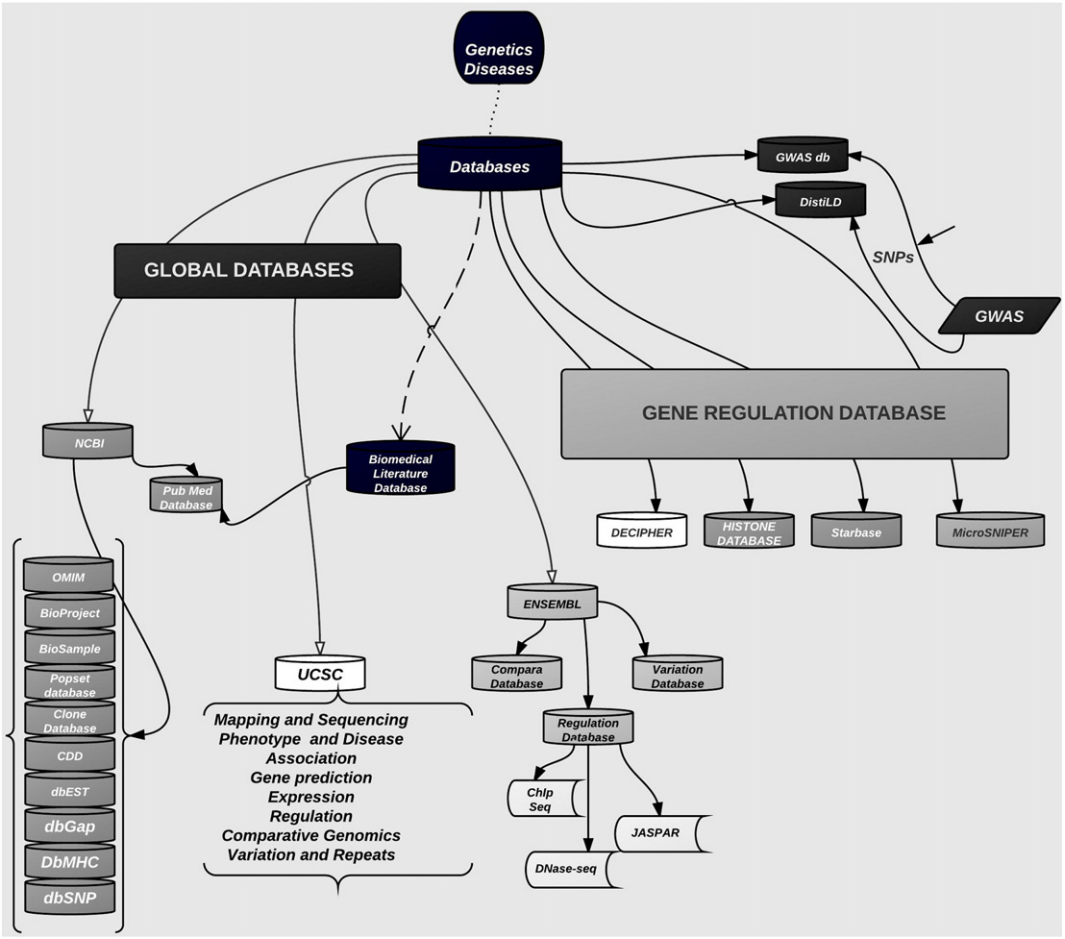

Fig. 1. Flow chart: databases useful in biomedical research of genetics diseases.

information provided covers bibliography, structure, function, association with disease and animal models.

A key feature of OMIM: it is completely interfaced to every other NCBI resources. By entering the name and/or the accession number of a gene, OMIM search results provide the complete description of the gene, of the diseases associated with it and also the description of other genes associated with these diseases. Being a very rich database, the interpretation of the results is not always simple: for each entry, results provide info about cloning, structure, chromosomal location, cytogenetic, animal models and bibliography.

These are some of the newest NCBI Databases:

- The BioProject database (www.ncbi.nlm.nih.gov/ bioproject/) enables users to submit comprehensive research studies ranging from focused genome sequencing projects to large international collaborations with multiple subprojects incorporating experiments resulting in nucleotide sequence sets, genotype/phenotype data, sequence variants and epigenetic information. This resource describes project scope, material and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases [20].
- The BioSample database (www.ncbi.nlm.nih.gov/biosample/) is a new resource that provides annotation for biological samples used in a variety of studies submitted to NCBI, including genome-wide association study (GWAS), epigenomics, genomics sequencing and microarrays. Currently BioSample contains over 600.000 samples and 90% of these coming from either short-read archive (SRA) or dbGaP [20].
- The PopSet database [21,22] (www.ncbi.nlm.nih.gov/popset/) is a collection of data that have been submitted to GenBank relating sequences and alignments derived from population, phylogenetic, mutation and ecosystem studies.
- The Clone database (CloneDB) (www.ncbi.nlm.nih.gov/clone/) integrates information about clones and libraries, including sequence data, map positions, distributor information, it also allows filtering by organism and vector type. The link with the Clone Finder helps locating of the clones by chromosomal position or by features such as genes, single nucleotide polymorphisms (SNPs), markers or transcript sequence accession number [22].
- A collection of sequence alignments and profiles, representing protein domains conserved in molecular evolution, and the alignments of the domains to known 3-dimensional protein structures in the MMDB (Molecular Modeling Database) (www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml) [22] constitute the Conserved Domain Database (CDD).
- Database of Expressed Sequence Tags (dbEST) contains short single-pass reads of cDNA (transcript) sequences. dbEST can be searched directly through the Nucleotide EST Database [23].
- Database of Genomic Structural Variation (dbVar) has been developed to archive information associated with large-scale genomic variation, including large insertions, deletions, translocations and inversions. This database also stores associations of defined variants with phenotype information [24].
- Entrez is an integrated database retrieval system that gives access to a diverse set of 35 databases that together contain over 570 million records [25]. It provides users with integrated access to sequence, mapping, taxonomy, and structural data. Furthermore, it also

supports graphical views of sequences and chromosome maps, very useful for genetic research.

- Databases of Genotypes and Phenotypes (dbGaP) contains results of different studies such as GWAS, medical resequencing, molecular diagnostic assays, which investigate the interaction between genotype and phenotype [26,27].
- Database of Major Histocompatibility Complex (dbMHC) consists of an interactive alignment viewer for HLA and related genes, an MHC microsatellite database, a sequence interpretation site for Sequencing Based Typing (SBT) and a primer/probe database [22,28].
- Database of Short Genetic Variations (dbSNP) [29] was originally created to support large-scale polymorphism discovery (e.g. HapMap), but it was rapidly adopted by the scientific community as the world archive for additional classes of variations such as insertions/deletions, microsatellites and non-polymorphic variants. It includes single nucleotide variations, microsatellites, small-scale insertions and deletions. dbSNP contains population-specific frequency and genotype data, experimental conditions, molecular context and mapping information for both neutral variations and clinical mutations [13].

2.2.2. Ensembl and Genome Browser of the University of California Santa Cruz

The work of genomes mapping has produced genomic maps with varying resolution, available for the human genome and for a large number of organisms of medical genetic interest frequently used as models (such as fruit fly, yeast, mouse and *C. elegans*). These data have often been integrated and are available in the form of databases accessible via the web. The web browser interfaces are linked to genomic databases with all the sequences produced by various genome-sequencing projects and the related notes. Through them it is possible to study the anatomy of the genomes at various degrees of detail, until the sequence, displaying at the same time all the structural and functional characteristics available for that section of the genome. In addition, it is possible to display mapping data for each stretch of DNA, where available. The most popular genome browsers are Ensembl and the UCSC Genome Browser. Both provide useful tools for data access from different sources.

2.2.2.1. Ensembl. Ensembl database (the name recalls the French word "ensemble" and "EMBL" European Molecular Biology Laboratory) database (http://www.ensembl.org) is a collaboration between the EMBL-European Bioinformatics Institute (EBI) and the Welcome Trust Sanger Institute (WTSI) to develop a software system that manages automated annotations on some eukaryotic genomes. The Ensembl project started in 1999, some years before the end of sequencing the human genome. It is based on an entirely 'open' philosophy: all data and program source code are available for the free and unrestricted use of both academic and commercial biomedical researchers worldwide. The main purpose of this web-database is to provide the reference of genome sequence data as a free resource for both scientists and patients' families associations and to integrate the genome with other biological data and ensure that everything could be accessible via the web [30,31]. Many other genomes have been added over the years including comparative genomics, variations and adjustment data. There is a continuous update of not only the genes belonging to the human, mouse, and zebrafish species [32] but also of other metazoa, plants, fungal invertebrates, bacteria, and unicellular eukaryotic and prokaryotic genomes. This update stems from the output of the Ensembl evidence-based automatic pipeline [33] and the manual annotation from the Havana project [34] in particular for human the Ensembl/Havana merged gene set continues to be equivalent to the Gencode gene set, the reference gene set for the Encode project [35]. It provides comprehensive, evidence-based gene annotations and comparative genomics resources including alignments and homology, ontology and paralogy relationships based on Ensembl [36].

The Gene Build team created the gene sets for the various species. The result of their work is stored in the core databases, by the Software team. This team also develops and maintains the BioMart data-mining tool [37]. Additional regulatory information was incorporated into Ensembl from high-throughput sequence assays of chromatin samples. All data are processed by an integrated mapping and processing pipeline using the eHive system [38].

Ensembl regulation database contains 369 ChIp-seq and DNase-seq data sets from 10 human and 5 mouse cell lines that include the genomic locations of binding regions for 74 different transcription factors (TF) as well as the locations of sites for 40 modified histones and 26 data sets that identify regions of open chromatin [39]. Moreover 25 of the TFs have binding matrices available through the JASPAR database and in particular it provides the positions of high probability TF-Binding sites within the binding regions [37].

The Variation database includes a large number of SNPs from human (dbSNP 132), cat (dbSNP 131), opossum (dbSNP 131) and pig (dbSNP 128). Each Ensembl release imports the structural variations from 1000 genomes of Pilot Project and structural variants from mouse, pig and dog [40].

The Compare, Variation and Regulation teams are responsible for handling the comparative the variation and regulatory data. They work along the Web team to make sure that all data are presented in a user-friendly way. The latest Ensembl's update has included variation data with the major human data release from dbSNP 132: Ensembl produces approximately five releases each year.

Releases are numbered sequentially (September 2011 was release 64) and include newly supported species, new assemblies and new or updated annotations of supported species. In recent times there have been significant improvements with the introduction of new softwares such as SIFT [41,42] and PolyPhen [43]. Both programs allow us to make predictions for the human proteome and the regulatory regions and to evaluate the consequences of nucleotide variations detected by direct sequencing. All changes are reported in terms of ontology standard sequence of definitions to ensure consistent across browsers [44].

So Ensembl was developed from a project aimed at producing a system capable of performing the automatic annotation of eukaryotic genomes and provide data visualization products. The software that runs the annotation is designed to follow the production of sequencing data step by step. The browser lets the user view and analyzes genes, transcripts as well as other collections of Expression Sequence Tag (EST) and genomic data to analyze the genome organization. It continues to provide extensive data resources for disease and phenotype annotations for germ and somatic variants. The data include somatic mutations from Catalogue of Somatic Mutations in Cancer (COSMIC) [45–47] and mutations from the public portion of the Human Gene Mutation Database (HGMD) [48]. Furthermore the data include many phenotype-variants associations from the NHGRI GWAS Catalog [49].

2.2.2.2. Genome Browser of the University of California Santa Cruz (Ucsc). The Genome Browser of the University of California Santa Cruz (UCSC) (http://genome.ucsc.edu/) is a popular Web-based tool for quickly displaying a requested portion of a genome at any scale [50,51]. It includes gene predictions, mRNA and expressed sequence tag alignments, simple nucleotide polymorphisms, expression and regulation data, phenotype and variation data and pair-wise or multiple species comparative genomics data. All information relevant to a region of interest is presented in one window, so the biological analysis is facilitated. Moreover tables underlying the Genome Browser tracks can be viewed and downloaded using the UCSC Table Browser the database [52,53].

Genome browser [52,54,55] facilitates the analysis of each gene through the following hyperlinks: Mapping and Sequencing Tracks, Phenotype and Disease Associations, Genes and Gene Prediction Tracks, mRNA and EST Tracks, Expression, Regulation, Comparative Genomics,

Neanderthal genome Assembly and Analysis, Variation and Repeats [55].

This database provides Softwares such as BLAT (to quickly find sequences of 95% and greater similarity of length 25 bases or more), Table Browser (to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track), and Gene Sorter (displays a sorted table of genes that are related to one another: the relationship can be one of several types, including protein-level homology, similarity of gene expression profiles or genomic proximity). In-Silico PCR searches a sequence database with a pair of PCR primers, using an indexing strategy for fast performance.

VisiGene Image Browser is a virtual microscope for in situ images. These images show where a gene is localized/expressed in an organism, sometimes reaching cellular resolution. With this browser users can retrieve images that meet specific search criteria, interactively zoom and scroll across the collection.

UCSC contains a large collection of genome sequences of vertebrates as well as those of insects and nematodes. It includes expression data, homology and map information. The browser lets the user select and scroll along the chromosome sequences by choosing the level of detail, to view information available in an integrated way. Moreover it is possible to correlate the information in different ways by highlighting similarities within subsets of genes.

The public availability of this comprehensive genetic information presents huge opportunities to develop new treatments for diseases based on understanding of the basic molecular processes of life.

2.2.3. Genome-Wide Association Studies (Gwas)

The thousands of SNPs associated with the risk of disorders, identified by GWAS [56], have substantially increased our knowledge in the field of genetics and molecular pathways underlying human traits and diseases.

GWASdb database freely available at http://jjwanglab.org/gwasdb provides an intuitive, multifunctional database for biologists and clinicians to explore genetic variation (GVs) and their functional inferences [57]. This database contains genetic variants with comprehensive functional annotations for each GV, genomic mapping information, regulatory effects (transcription factor binding sites, microRNA target sites and splicing sites), amino acid substitutions, evolution, gene expression and disease associations. Furthermore, GWASdb classifies these GVs according to diseases using Disease-Ontology Lite and Human Phenotype Ontology. It is able to perform pathway enrichment and PPI network association analysis for these diseases. GWASdb identifies marker SNPs, which are not necessarily the causal SNPs but are assumed to be in linkage disequilibrium (LD) with them.

With the aim of providing an example of how these bioinformatics tools can be used in human genetic research we describe how we used them to search the genetic component involved in inflammatory bowel disease (IBD): the GWAS allows us the evaluation of specific variants. Genetic variants associated with IBD can vary in frequency depending on the cohort ethnicity, raising the possibility that some variants could have emerged in the context of historical selective pressures. Current GWAS are typically powered to characterize variants with $>1\%$ frequency and do not include the contributions from rare variants. If pedigrees are available, rare variants discovery can be further targeted by fine mapping, as shown by the identification of interleukin-10 receptor subunit alpha (IL10RA) polymorphisms associated with the development of early-onset IBD [58]. By the analysis of the entire coding region of the IL10 receptor and two exons of NOD2 in four unrelated cases, in one early-onset Crohn patient, with variable clinical severity, we recently identified one potentially important interleukin-10 receptor subunit beta (IL10RB) gene variation in homozygosis, probably able to affect the signaling of interleukin (IL)-10, [59]. The fact that the number of identified loci do not account for the total heritability of Crohn disease (30% approximately) is a difficult issue. This is a problem referred to as missing heritability and specific genetic tests might be developed to improve counseling, while direct identification of modifier genes might assist in the recognition of new genetic, environmental and microbial causes for Crohn's disease [60].

To increase the usage of existing GWAS results, the DistiLD database [61] (http://distild.jensenlab.org) allows the visualization of disease-associated SNPs and genes in their chromosomal context. In this database the published GWAS are collected from several sources and linked to standardized international codes. HapMap data are analyzed to define LD blocks onto which SNPs and genes are mapped and finally a web interface makes it easy to query and to visualize disease-associated SNPs and genes within LD blocks. DistiLD database could be the starting point for many studies, especially for complex disease where the causal gene remain difficult to identify, by providing them the LD blocks associated with a given disease containing the set of genes in LD with the SNPs associated with the disease [61].

2.2.4. Genscan

Some nucleotide variations may lead to the formation of abnormal proteins. These variants, present in intron or in exon region, could induce an abnormal splicing mechanism that involves the synthesis of abnormal transcripts. A single point mutation may cause the production of a new protein isoform with or without the loss of the original protein. In presence of a particular variant, the corresponding exons and introns within the genome sequences could be predicted with the database GenScan (http://genes.mit.edu/GENSCAN.html). It can be used to identify the wild-type and the alternative exon-intron boundaries in genomic DNA [62].

2.2.5. Gene regulation databases

In the 46 human chromosomes (22 autosomal and 2 sexual), about 30,000–40,000 protein-coding genes are known. Most genetic disorders are the direct result of a gene mutation. Furthermore, for diseases with a complex pattern of inheritance, the single gene mutation doesn't cause the disease, but more variations are needed for the occurrence of the disease; in this case more genes contribute to increase the susceptibility of an individual to the disease, and the influence of the environmental factors must be taken into account. In this context we can also consider those patients suffering from a disease affecting the copy number of dosage-sensitive genes, or by the incorrect gene regulation and/or expression.

2.2.5.1. Decipher. Many genetic aberrations are new and extremely rare, often clinical interpretation is problematic and genotype-phenotype correlation quite uncertain. The interactive database called DECIPHER [63] (Database of Chromosomal Imbalance and phenotype in Humans Using Ensembl Resources) can be used, to facilitate the analysis of patients suffering from this kind of disorders (caused by chromosomal rearrangements). DECIPHER helps genetic counseling by retrieving relevant information from a variety of bioinformatics resources with a set of tools to help the designated interpretation of submicroscopic chromosomal imbalances, inversions and translocations. Known and predicted genes within an aberration are listed in the DECIPHER patient report, and genes of recognized clinical importance are highlighted and prioritized. This web-database integrates seamlessly with the Ensembl genome browser and interrogates the current version of the human genome assembly displayed in the Ensembl genome browser. Moreover it links to other genetic and medical databases, including Hugo Gene Nomenclature Committee (HGNC), (OMIM) On line Mendelian Inheritance in Man, PubMed, GeneReviews, Ensembl genes and Swiss-Prot. A frequently updated list of emerging bioinformatics databases comprises a feature graph tool that displays patient copy-number variations together with data from the Database of Genomic Variants (DGV) as well as variants identified in a selection of the major studies of copy-number variation in normal individuals. So DECIPHER enables clinical scientists worldwide to maintain records of

phenotype and chromosome rearrangement for their patients and, with informed consent, shares this information with the wider clinical research community through display in the genome browser Ensembl. By sharing cases worldwide, clusters of rare cases having phenotype and structural rearrangement in common can be identified, leading to the delineation of new syndromes and furthering understanding of gene function [64,65].

2.2.5.2. Histone Database. Both in Gene regulation and in chromatin organization the histones play central roles: they constitute the fundamental protein units of the nucleosome [66]. Core of histones are highly conserved across eukaryotes in terms of sequence and structure. Despite overall sequence conservation, extensive histone tail post-translational modifications, (in addition to histone variants present during development), contribute to epigenetic mechanism that signal transcriptional activation, repression and recombination events [67,68]. The Histone database is a comprehensive bioinformatics resource. It organizes, stores and groups the histones sequences into families. It maintains a collection of histone fold-containing sequences, and provides information on three-dimensional structures available in PDB. The last update contains 182 sequences also from 89 archaeal organisms, including members of all classified archaeal phyla. The Histone database contains entries that represent a total of 7356 unique NCBI taxonomic identifiers, which correspond approximately to the same number of organisms. The sequences of core, linker and archeal histones are available in FASTA format. They are also available as a series of multiple sequence alignments, one for each class of proteins [69].

2.2.5.3. Starbase and microsniper database. Functional studies attempting to determine causality have focused largely on coding variants, although non-coding SNPs can be associated with qualitative and quantitative changes. Alternative splicing exemplifies a qualitative change affected by non-coding modifications.

Non-coding RNAs, such as microRNAs [70], participate very actively in the non-encoded gene expression regulation [71]. miRNAs, small RNA molecules (about 22nucleotides), interact with their corresponding target mRNAs inhibiting the translation of the mRNA into proteins and cleaving the target mRNA [72]. This second effect diminishes the overall expression of the target mRNA. Researchers at Sun Yat-sen University, China have developed a novel database, starBase (sRNA target Base), to facilitate the comprehensive exploration of miRNA-target interaction maps from CLIP-Seq and Degradome-Seq data [73]. MicroRNAs are pivotal regulators of development and cellular homeostasis. They act as post-transcriptional regulators, controlling the stability and translation efficiency of their target mRNAs. The prediction of microRNA targets is a crucial component for understanding modulation of their mRNA expression. The current version includes high-throughput sequencing data generated from 21 CLIP-Seq and 10 Degradome-Seq experiments from six organisms. By analyzing millions of mapped CLIP-Seq and Degradome-Seq reads, the authors identified ~1 million Ago-binding clusters and ~2 million cleaved target clusters in animals and plants, respectively. Analyses of these clusters, and of target sites predicted by 6 miRNA target prediction programs, resulted in identification of approximately 400,000 and approximately 66,000 miRNA-target regulatory relationships from CLIP-Seq and Degradome-Seq data, respectively.

Two web servers are provided to discover novel miRNA target sites from CLIP-Seq and Degradome-Seq data. The web implementation supports diverse query types and exploration of common targets, gene ontology and pathways.

The StarBase is available at http://starbase.sysu.edu.cn/. It facilitates the integrative, interactive and versatile display of the discovery and comprehensive annotation of miRNA-target interaction maps from CLIP-Seq and Degradome-Seq data from six organisms: human, mouse, *Caenhorhabditis elegans*, *Arabidopsis thaliana*, *Oryza sativa* and *Vitis vinifera* [73]. Given the impact of microRNA on post-transcriptional regulation and its potential relation to complex diseases, there is substantial interest in methods designed to predict the microRNA targets and effect of SNPs on microRNA binding.

A web-based application, MicroSNiPer [74], has been developed to predict the impact of a SNP on putative microRNA targets. It is freely accessible at http://cbdb.nimh.nih.gov/microsniper and allows researchers to estimate the impact of a SNP on a putative miRNA target, providing information about the creation or disruption of putative miRNA binding sites in a gene due to the presence of alternative SNP alleles. This application interrogates the 3'-untranslated region and predicts if a SNP within the target site will disrupt/eliminate or enhance/create a microRNA binding site. MicroSNiPer computes these sites and examines the effects of SNPs in real time. MicroSNiPer is a user-friendly web-based tool. Its advantages include ease of use, flexibility and straightforward graphical representation of the results. This web-tool provides a high degree of flexibility at the input stage. It is useful for labs carrying out in-house discovery and characterization of novel transcripts and SNPs. This package makes it possible to combine catalogued and novel SNPs, as well as to manually enter or edit 3′UTRs based on newly available experimental data. Computational prediction of SNP impact on miRTS (followed by the experimental validation of the miRNA binding) can determine whether a functional SNP is affecting gene expression or not.

To provide an example derived from of our group research activity, studying the genetic factors involved in IBD [75,76], we know that, in the context of immune responses regulation, NOD2 [77,78] can encode truncated variants that inhibit their signaling pathways [79]. Genetic changes may affect transcription-factor-binding sequences, locus accessibility, translational efficiency and trans-regulators such as non-coding RNAs and microRNAs (miRNAs). In this regard, a Crohn's-disease-associated synonymous variant in IRGM (c.313C>T) perturbs regulation by miR-196A and miR-196B, and is associated with altered IRGM expression in patients with Crohn's disease who bear this SNP [80].

3. Discussion

The rapid growth of genome sequencing projects increased the number of useful information: nowasdays the genome browser is not only a visualization system but also an interactive platform to support open data access and collaborative work. Thus a customizable genome browser framework with rich functions and flexible configuration is needed to facilitate the variety of genome research projects. Currently the database is an entity in which users can store data in a structured manner with minimum possible repetitions.

The projects studying the genomes of different organisms need to generate and categorize different types of data including, for example, the information necessary to determine first the physical maps, the DNA sequences, the genes and proteins with their functional descriptions (in a simple and orderly way). The collection of this information, which begun in the early eighties, led to a huge growth of the data contained in databases, and with current sequencing techniques the content of these banks are doubling every two years. It is easy to predict that this growth rate will further increase, as will problems arising from the management of this large amount of data. Several methods have been developed to facilitate the access to data and allow studying methods of analysis in response to the growing number and complexity of data generated during the project sequencing of the genomes of different organisms, and in particular the Human Genome Project, Some typical objectives of bioinformatics are the development of tools to maintaining the information coming from various sources: physical and genetic maps, chromosome mapping, cytogenetic map, polymorphisms, and information about the genomic and protein sequences. All these tools should be not only data containers but also offer a precise guidance specifying the steps for searching a gene mutation in monogenic disorder and how you can join data by searching for multiple genetic factors thus influencing

the onset of complex diseases. Moreover, the following sources should also be cited:

a) the collection and the organization of the genetic information associated with human diseases;
b) the development of computing programs for the analysis, and the interpretation of various types of data such as nucleotide and amino acid sequences, domains and structures of proteins;
c) the development of graphical user interfaces that can effectively display the information;
d) the design and construction of a data network for the collection, distribution, and continuous updating of the whole information produced.

All data collected will be stored and classified in the myriad of databases available. We must consider that all data entered in these databases (for several reasons) are not always able to shed light in an appropriate manner. Therefore, despite the enormous potential of bioinformatics, supplementary tools will be needed for simplified and easy mining in complex databases. Bioinformatics, on the one hand, has the important objective to organize and validate the databases of various origins and, on the other hand, to develop analytical software and paths for the integration of different information. Moreover, an information database for about all phenotypes related to a possible Mendelian form of hereditary disease has not been fully established (in spite of the very useful and powerful OMIM), and it may overlap with other characterized phenotypes. As we can see, many questions remain to be answered regarding the identity of single genes and their role in human diseases.

4. Conclusions and application of databases example

In conclusion, we would like to quote our own experience about the usefulness of Databases in the study of genetic variants associated with inflammatory bowel disease. From an analysis carried out on patients with very early onset of the disease, we have found some SNPs that we have then studied deeper with the use of the databases cited here. By inserting a specific code in the dbSNP database, it responded with the specific nucleotide sequence, with links to OMIM, MapView, PubMed, GeneView, SeqView, Protein 3D, and other information as Allele Origin, MAF/MinorAlleleCount and the Clinical Significance. These allowed us to discover: the chromosomal location, relevant data in literature, the complete vision of the gene in question and its sequence, the connection with changes mapped on the protein and so on. By entering the specific "rs" in the dbUCSC and/or Ensembl we can know, not only the chromosomic position and the probable association with a particular disease, but also as much as it is possible to know about the change, because we can read information from different databases linked with it, such as dbSNP135, dbSNP 131, HapMap SNPs from the CEU Population, and the nucleotide conservation in the species and so on.

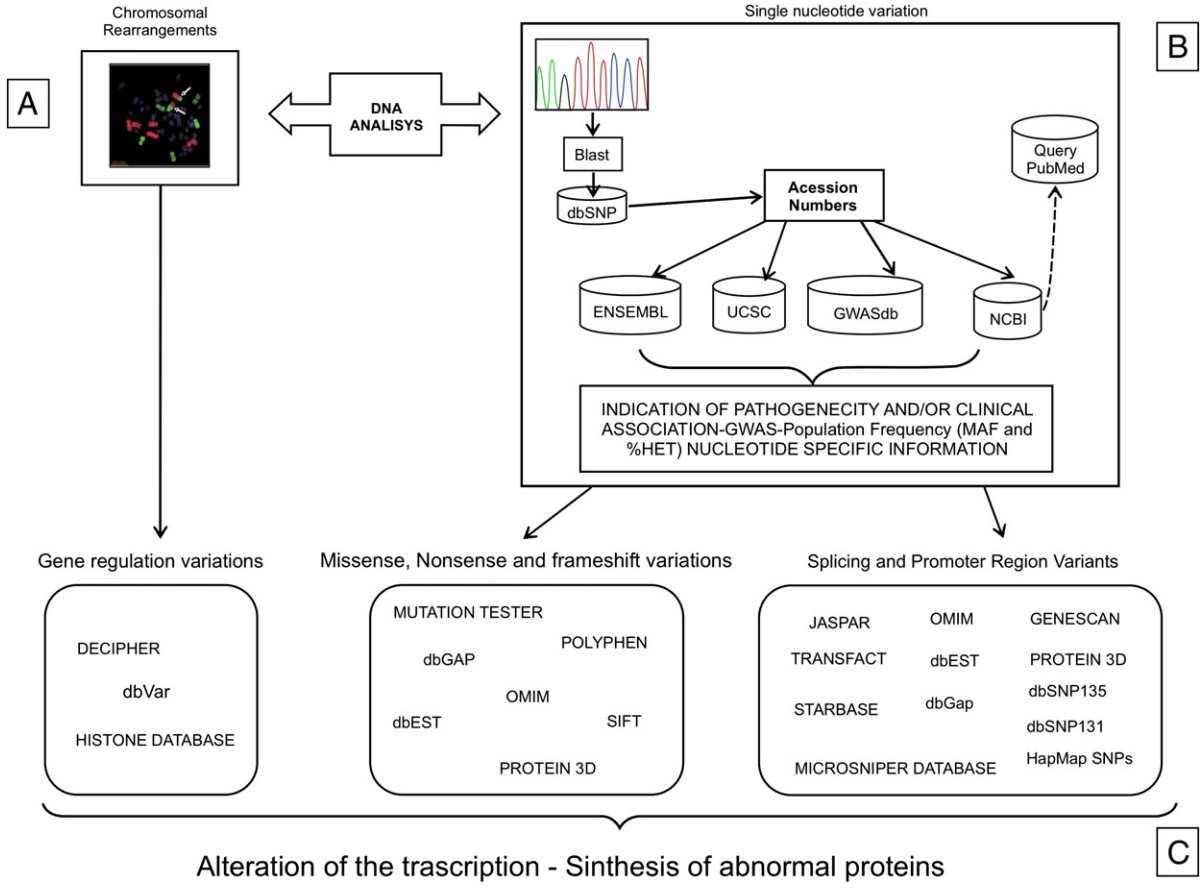

Fig. 2. DNA analysis. A. Decipher is a database for the designated interpretation of submicroscopic chromosomal imbalances, inversions and translocations; histone database, a comprehensive bioinformatic resource that organizes and provides information on three-dimensional structures and; dbVar that has been developed to archive information associated with large-scale genomic variation and stores associations of defined variants with phenotype information. B. Single nucleotide variation. From a Blast in dbSNP comes out (only if the nucleotide change is known) the refseq identifier. Through the refseq is possible access in the general Databases (NCBI, Ensembl, UCSC, GWASdb) for known general and specific information such as Population frequency, chromosomal position, gene information and if it is known the clinical association. C. With the accession number or with the same sequence is possible to access to many other databases (OMIM, Polyphen, Sift, Mutation Tester, Starbase, MicroSniper, Protein 3D, dbSNP131, dbSNP135, HapMap SNPs, GENESCAN, dbEST, that interact each other in many ways, and can be up to predict a possible alteration of the transcription (mutations in the promoter region), a synthesis of abnormal proteins (changes in the process of splicing and/or mutations that involve a stop codon).

Furthermore when a nucleotide variation that implies the resulting amino acid change is present within Ensembl, there are also several links that indicate the prediction of pathogenicity.

Beyond listing of databases for genetic studies, there are interesting considerations to be done, finally the message we would like to convey is that the approach to the study of Mendelian and/or Complex genetic disorders can be completed and, why not, much simplified by the use of databases. A scheme that can be adopted to understand the consequences of different types of genomic anomalies is shown in Fig. 2. In the case of large chromosomal rearrangements, it is possible to use databases such as Decipher (which is an interactive database with a set of tools to help the designated interpretation of submicroscopic chromosomal imbalances, inversions and translocations), Histone database (a comprehensive bioinformatics resource that organizes and provides information on three-dimensional structures) and dbVar (developed to archive information associated with large-scale genomic variation and stores associations of defined variants with phenotype information).

In the case of a variation of a single nucleotide it could be a good idea to identify this variant in the data bank. A very explanative database is the dbSNP of NCBI. By performing a Blast using as query 30–40 nucleotides of the sequence containing the nucleotide variation the SNP identifier comes out (only if the nucleotide change is known). Through the refseq (rs) it is possible to search access the general databases (NCBI, Ensembl, UCSC, GWASdb) for known general and specific information, such as Population frequency, chromosomal position, gene information, and if it is known, the clinical association as well. With the accession number or with the sequence it is possible to access also many other databases such as Polyphen, Sift, Starbase, MicroSniper.

Most of them in fact interact with each other in many ways, allowing the answers to appear interconnected in such a way that starting from a simple variation in a sequence, the use of these tools can lead up to predict a possible alteration of the transcription (mutations in the promoter region), a synthesis of abnormal proteins (changes in the process of splicing and/or mutations that involve a stop codon).

Indeed we believe that establishing connections between different databases is an important way to improve the databases themselves, helping the user with additional research. Certainly, we have restricted our description to the most conspicuous and important databases, but also to those we believe contribute to the advancement of genomics and bioinformatic in scientific world. Despite the preliminary findings presented in our study due to limited number of databases described, we would like to open a discussion aimed at improving the knowledge and proper use of genomic databases and analysis programs in the search for genetic defects associated with complex human diseases, being them an essential source of information for the variegated world of scientific communities.

5. Application of databases information in IBD study

Inflammatory bowel diseases (IBD) are idiopathic, chronic and relapsing inflammatory conditions of the gastrointestinal tract and comprise the chronic relapsing inflammatory disorders Crohn's disease and Ulcerative Colitis. While the Crohn disease may involve any part of the

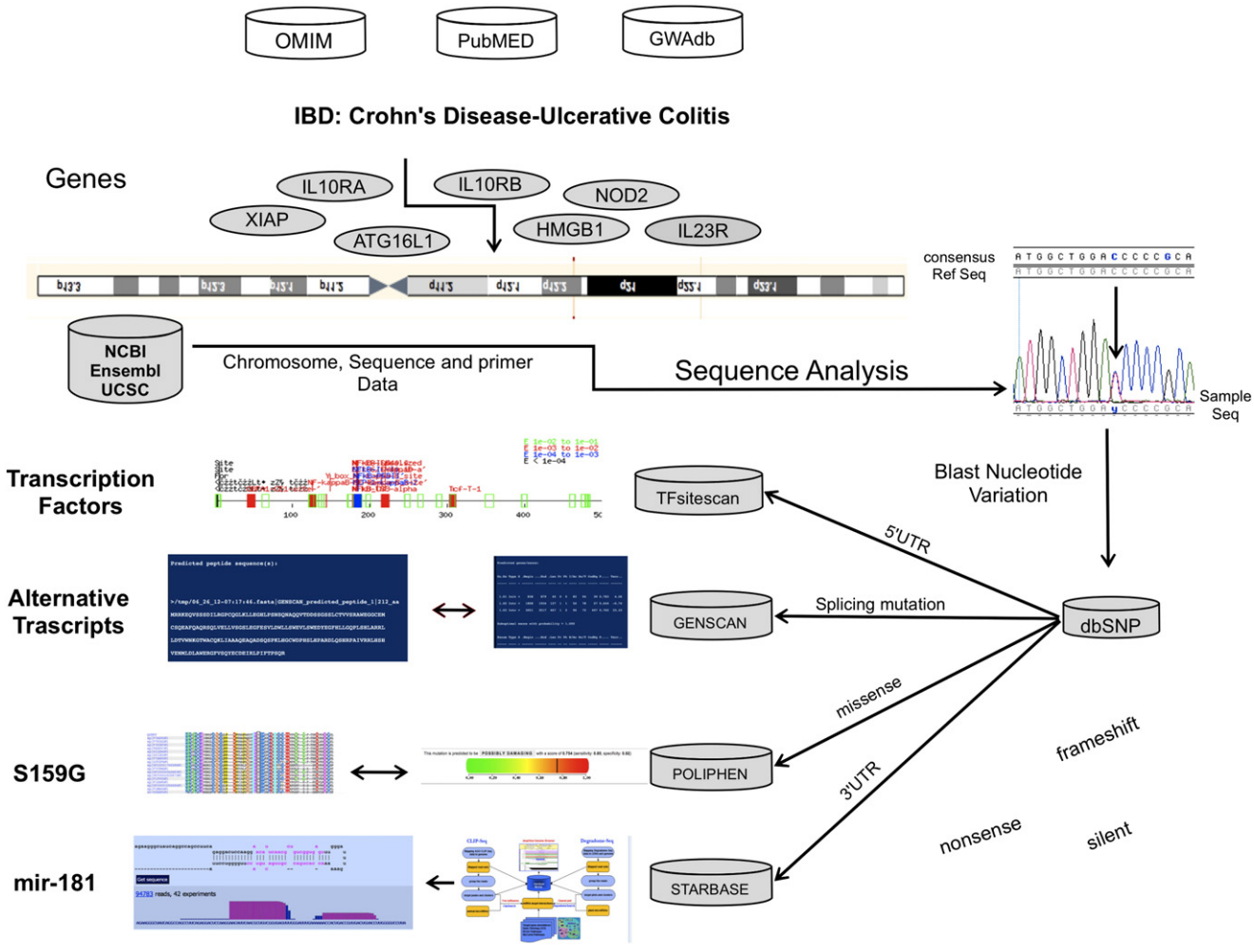

Fig. 3. Variant analysis.

gastrointestinal tract, but most frequently the terminal ileum and colon with transmural discontinuous lesions, in the ulcerative colitis the inflammation is continuous and limited to rectal and colonic mucosal layers. Both diseases are complex genetic traits as inheritance does not follow any simple Mendelian models and both genetic and environmental factors seem to be important in disease aetiology. Family history is a risk factor for developing IBD, with a peak incidence in early life, although individuals of any age can be affected. Progress in IBD knowledge has been notable, with many GWAS publications increasing the number of confirmed associated loci. Our research group is currently studying some of the IBD susceptibility genes that came out from the literature, especially by GWAS. Through a search in PubMed, according to the different results, we decided to analyze IBD patients by gene amplification and sequencing for some genes known to be associated to Crohn's disease and/or to Ulcerative Colitis. All patients were characterized for some gene of interests and, in case of identification of new variants we used bioinformatics tools to predict the pathogenic effect.

Fig. 3 shows schematically the use of databases useful for studying a genetic disease, and especially how their use can be helpful in promoting a clear understanding of a nucleotide variant. The use of databases such as PubMed, OMIM, and GWAdb (in case of non-syndromic and multifactorial diseases) is useful not only for providing information about the selected disease, but let us know about the previous studies, allowing a better and appropriate choice of genes.

Crossing NCBI, Ensembl, and UCSC information, a detailed evaluation was performed for each gene by the investigation of the chromosomal location, the genomic structure, the different transcripts and all polymorphic variants notes. Then was performed the canonical amplification and the sequencing of the entire coding region, including the intron–exon boundaries and the 5′ and the 3′ un-translated regions of the gene.

Variant analysis:

The first step is to investigate each variant in the database of SNPs (SNPdb) by performing a simple Blast that give us general information about the variant.

The second step consists to investigate the potential damage of the variant by the use of specific database. With the Polyphen database it is possible to determine if a nucleotide variant in first or second base of reading frame triplet could originate a change in the amino acid sequence that could be or not harmful for the protein. When we encountered variants in the intron–exon junctions and/or nearby in sequences important for the splicing process (Enhancer and Silencer), Genscan helps us to predict whether it creates a cryptic splice site and therefore an alternative transcript. A way to predict if variants in regulatory regions may determine or not the binding of specific miRnas is to consult both Starbase and MiRbase.

For example, in our IBD patients we identified both in homozygosis and in heterozygosis a nucleotide variant in the exon 2 of the IL10RB gene (This variant presents as polymorphism (rs2834167) in the dbSNP is a A to G variation (c.A139G) that determines the amino acid change from Lysine (K) to glutammic acid (E), (K47E). By Polyphen investigation this mutation is predicted to be possibly damaging with a score of 0.583.

So the databases can allow us to predict first and to direct the operator to the functional studies targeted to validate the prediction made by the bioinformatic tool. All described bioinformatics tools are also very useful in deepening the results obtained with the next generation sequencing.

Acknowledgments

This work was supported by a grant from the Institute for Maternal and Child Health IRCCS “Burlo Garofolo,” Italy (RC 40/2011).

References

[1] N.C. Kyrpides, Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide, Bioinformatics 15 (1999) 773–774.
[2] A. Bernal, U. Ear, N. Kyrpides, Genomes OnLine Database (GOLD): a monitor of genome projects world-wide, Nucleic Acids Res. 29 (2001) 126–127.
[3] I. Pagani, K. Liolios, J. Jansson, I.M. Chen, T. Smirnova, B. Nosrat, V.M. Markowitz, N.C. Kyrpides, The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata, Nucleic Acids Res. 40 (2012) D571–D579.
[4] J. Jorda, T. Baudrand, A.V. Kajava, PRDB: Protein Repeat DataBase, Proteomics 12 (2012) 1333–1336.
[5] A.D. Baxevanis, The Molecular Biology Database Collection: an updated compilation of biological database resources, Nucleic Acids Res. 29 (2001) 1–10.
[6] A.D. Baxevanis, The importance of biological databases in biological discovery, Curr. Protoc. Bioinformatics 1 (Mar 2006) 1.1. (Review).
[7] A.D. Baxevanis, The importance of biological databases in biological discovery, Curr. Protoc. Bioinformatics 1 (Sep 2009) 1.1.
[8] M.Y. Galperin, X.M. Fernandez-Suarez, The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection, Nucleic Acids Res. 40 (2012) D1–D8.
[9] M.Y. Galperin, G.R. Cochrane, Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009, Nucleic Acids Res. 37 (2009) D1–D4.
[10] L. Hunter, K.B. Cohen, Biomedical language processing: what's beyond PubMed? Mol. Cell 21 (2006) 589–594.
[11] R. Islamaj Dogan, G.C. Murray, A. Neveol, Z. Lu, Understanding PubMed user search behavior through log analysis, 2009 Database (Oxford) (Nov 27 2009) bap018 (Epub).
[12] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical literature, 2011 Database (Oxford) (2011 Jan 18) baq036.
[13] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, J. Ye, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 40 (2012) D13–D25.
[14] J. Ding, L.M. Hughes, D. Berleant, A.W. Fulmer, E.S. Wurtele, PubMed Assistant: a biologist-friendly interface for enhanced PubMed search, Bioinformatics 22 (2006) 378–380.
[15] C. Plake, T. Schiemann, M. Pankalla, J. Hakenberg, U. Leser, AliBaba: PubMed as a graph, Bioinformatics 22 (2006) 2444–2445.
[16] R.T. Tsai, H.J. Dai, P.T. Lai, C.H. Huang, PubMed-EX: a web browser extension to enhance PubMed search with text mining features, Bioinformatics 25 (2009) 3031–3032.
[17] P.G. Baker, A. Brass, Recent developments in biological sequence databases, Curr. Opin. Biotechnol. 9 (1998) 54–58.
[18] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick, Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, Nucleic Acids Res. 33 (2005) D514–D517.
[19] V.A. McKusick, Mendelian inheritance in man and its online version, Am. J. Hum. Genet. 80 (2007) 588–604 (OMIM).
[20] T. Barrett, K. Clark, R. Gevorgyan, V. Gorelenkov, E. Gribov, I. Karsch-Mizrachi, M. Kimelman, K.D. Pruitt, S. Resenchuk, T. Tatusova, E. Yaschenko, J. Ostell, BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata, Nucleic Acids Res. 40 (2012) D57–D63.
[21] M. Arenas, M. Patricio, D. Posada, G. Valiente, Characterization of phylogenetic networks with NetTest, BMC Bioinformatics 11 (2010) 268 (20).
[22] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, Ye J, Database resources of the National Center for Biotechnology Information Nucleic Acids Res. 40 (2012) D13–D25.
[23] M.S. Boguski, T.M. Lowe, C.M. Tolstoshev, dbEST–database for “expressed sequence tags”, Nat. Genet. 4 (1993) 332–333.
[24] D.M. Church, I. Lappalainen, T.P. Sneddon, J. Hinton, M. Maguire, J. Lopez, J. Garner, J. Paschall, M. DiCuccio, E. Yaschenko, S.W. Scherer, L. Feuk, P. Flicek, Public data archives for genomic structural variation, Nat. Genet. 42 (2010) 813–814.
[25] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res. 35 (2007) D26–D31.
[26] M.D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, N. Popova, S. Pretel, L. Ziyabari, M. Lee, Y. Shao, Z.Y. Wang, K. Sirotkin, M. Ward, M. Kholodov, K. Zbicz, J. Beck, M. Kimelman, S. Shevelev, D. Preuss, E. Yaschenko, A. Graeff, J. Ostell, S.T. Sherry, The NCBI dbGaP database of genotypes and phenotypes, Nat. Genet. 39 (2007) 1181–1186.
[27] GAIN Collaborative Research Group, T.A. Manolio, L.L. Rodriguez, L. Brooks, G. Abecasis; Collaborative Association Study of Psoriasis, D. Ballinger, M. Daly, P. Donnelly, S.V. Faraone; International Multi-Center ADHD Genetics Project, K. Frazer, S. Gabriel, P. Gejman; Molecular Genetics of Schizophrenia Collaboration, A. Guttmacher, E.L. Harris, T. Insel, J.R. Kelsoe; Bipolar Genome Study, E. Lander, N. McCowin, M.D. Mailman, E. Nabel, J. Ostell, E. Pugh, S. Sherry, P.F. Sullivan; Major Depression Stage 1 Genomewide Association in Population-Based Samples Study, J.F. Thompson, J. Warram; Genetics of Kidneys in Diabetes (GoKinD) Study, D. Wholley, P.M. Milos, F.S. Collins, New Models of Collaboration in Genome-Wide Association Studies: The Genetic Association Information Network, Nat. Genet. 39 (2007) 1045–1051.

[28] W. Helmberg, R. Dunivin, M. Feolo, The sequencing-based typing tool of dbMHC: typing highly polymorphic gene sequences, Nucleic Acids Res. 32 (2004) W173–W175.
[29] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311.
[30] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X.M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp, T. Hubbard, Ensembl 2004, Nucleic Acids Res. 32 (2004) D468–D470.
[31] E. Birney, D. Andrews, M. Caccamo, Y. Chen, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, S. Graf, M. Hammond, J. Herrero, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, F. Kokocinski, E. Kulesha, D. London, I. Longden, C. Melsopp, P. Meidl, B. Overduin, A. Parker, G. Proctor, A. Prlic, M. Rae, D. Rios, S. Redmond, M. Schuster, I. Sealy, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, A. Stabenau, J. Stalker, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark, T.J. Hubbard, Ensembl 2006, Nucleic Acids Res. 34 (2006) D556–D561.
[32] P.J. Kersey, D. Lawson, E. Birney, P.S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kahari, R.J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A.J. Vilella, A. Yates, Ensembl Genomes: extending Ensembl across the taxonomic space, Nucleic Acids Res. 38 (2010) D563–D569.
[33] S.C. Potter, L. Clarke, V. Curwen, S. Keenan, E. Mongin, S.M. Searle, A. Stabenau, R. Storey, M. Clamp, The Ensembl analysis pipeline, Genome Res. 14 (2004) 934–941.
[34] I. Ezkurdia, A. Del Pozo, A. Frankish, J.M. Rodriguez, J. Harrow, K. Ashman, A. Valencia, M.L. Tress, Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function, Mol. Biol. Evol. 29 (9) (Sep 2012) 2265–2283.
[35] V. Curwen, E. Eyras, T.D. Andrews, L. Clarke, E. Mongin, S.M. Searle, M. Clamp, The Ensembl automatic gene annotation system, Genome Res. 14 (2004) 942–950.
[36] A.J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, E. Birney, EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates, Genome Res. 19 (2009) 327–335.
[37] E. Portales-Casamar, S. Thongjuea, A.T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W.W. Wasserman, A. Sandelin, JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, Nucleic Acids Res. 38 (2010) D105–D110.
[38] J. Severin, K. Beal, A.J. Vilella, S. Fitzgerald, M. Schuster, L. Gordon, A. Ureta-Vidal, P. Flicek, J. Herrero, eHive:An Artificial Intelligence workflow system for genomic analysis 11 (May 11 2010) 240.
[39] J. Severin, K. Beal, A.J. Vilella, S. Fitzgerald, M. Schuster, L. Gordon, A. Ureta-Vidal, P. Flicek, J. Herrero, eHive: an artificial intelligence workflow system for genomic analysis, BMC Bioinformatics 11 (2010) 240.
[40] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, N. Johnson, A.K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H.S. Riat, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernández-Suarez, J. Harrow, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, S.M. Searle, Ensemble 2012 Nucleic Acids Res 40 (D1) (Jan 2012) D84–D90.
[41] N.L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins, Nucleic Acids Res. 40 (Jul 2012) W452–W457 (Web Server issue).
[42] P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4 (2009) 1073–1081.
[43] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[44] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The Sequence Ontology: a tool for the unification of genome annotations, Genome Biol. 6 (2005) R44.
[45] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, R. Wooster, The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website, Br. J. Cancer 91 (2004) 355–358.
[46] S.A. Forbes, G. Tang, N. Bindal, S. Bamford, E. Dawson, C. Cole, C.Y. Kok, M. Jia, R. Ewing, A. Menzies, J.W. Teague, M.R. Stratton, P.A. Futreal, COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer, Nucleic Acids Res. 38 (2010) D652–D657.
[47] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J.W. Teague, P.J. Campbell, M.R. Stratton, P.A. Futreal, COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer, Nucleic Acids Res. 39 (2011) D945–D950.
[48] P.D. Stenson, E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S. Thomas, S. Abeysinghe, M. Krawczak, D.N. Cooper, Human Gene Mutation Database (HGMD): 2003 update, Hum. Mutat. 21 (2003) 577–581.
[49] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 9362–9367.
[50] P.A. Fujita, B. Rhead, A.S. Zweig, A.S. Hinrichs, D. Karolchik, M.S. Cline, M. Goldman, G.P. Barber, H. Clawson, A. Coelho, M. Diekhans, T.R. Dreszer, B.M. Giardine, R.A. Harte, J. Hillman-Jackson, F. Hsu, V. Kirkup, R.M. Kuhn, K. Learned, C.H. Li, L.R. Meyer, A. Pohl, B.J. Raney, K.R. Rosenbloom, K.E. Smith, D. Haussler, W.J. Kent, The UCSC Genome Browser database: update 2011, Nucleic Acids Res. 39 (2011) D876–D882.
[51] T.R. Dreszer, D. Karolchik, A.S. Zweig, A.S. Hinrichs, B.J. Raney, R.M. Kuhn, L.R. Meyer, M. Wong, C.A. Sloan, K.R. Rosenbloom, G. Roe, B. Rhead, A. Pohl, V.S. Malladi, C.H. Li, K. Learned, V. Kirkup, F. Hsu, R.A. Harte, L. Guruvadoo, M. Goldman, B.M. Giardine, P.A. Fujita, M. Diekhans, M.S. Cline, H. Clawson, G.P. Barber, D. Haussler, W. James Kent, The UCSC Genome Browser database: extensions and updates 2011, Nucleic Acids Res. 40 (2012) D918–D923.
[52] K.R. Rosenbloom, T.R. Dreszer, J.C. Long, V.S. Malladi, C.A. Sloan, B.J. Raney, M.S. Cline, D. Karolchik, G.P. Barber, H. Clawson, M. Diekhans, P.A. Fujita, M. Goldman, R.C. Gravell, R.A. Harte, A.S. Hinrichs, V.M. Kirkup, R.M. Kuhn, K. Learned, M. Maddren, L.R. Meyer, A. Pohl, B. Rhead, M.C. Wong, A.S. Zweig, D. Haussler, W.J. Kent, ENCODE whole-genome data in the UCSC Genome Browser: update 2012, Nucleic Acids Res. 40 (2012) D912–D917.
[53] A.S. Zweig, D. Karolchik, R.M. Kuhn, D. Haussler, W.J. Kent, UCSC genome browser tutorial, Genomics 92 (2008) 75–84.
[54] R.M. Kuhn, D. Karolchik, A.S. Zweig, T. Wang, K.E. Smith, K.R. Rosenbloom, B. Rhead, B.J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A.S. Hinrichs, R.A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G.P. Barber, D. Haussler, W.J. Kent, The UCSC Genome Browser Database: update 2009, Nucleic Acids Res. 37 (2009) D755–D761.
[55] B. Rhead, D. Karolchik, R.M. Kuhn, A.S. Hinrichs, A.S. Zweig, P.A. Fujita, M. Diekhans, K.E. Smith, K.R. Rosenbloom, B.J. Raney, A. Pohl, M. Pheasant, L.R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R.A. Harte, B. Giardine, T.R. Dreszer, H. Clawson, G.P. Barber, D. Haussler, W.J. Kent, The UCSC Genome Browser database: update 2010, Nucleic Acids Res. 38 (2010) D613–D619.
[56] F. Begum, D. Ghosh, G.C. Tseng, E. Feingold, Comprehensive literature review and statistical considerations for GWAS meta-analysis, Nucleic Acids Res. 40 (9) (May 2012) 3777–3784.
[57] M.J. Li, P. Wang, X. Liu, E.L. Lim, Z. Wang, M. Yeager, M.P. Wong, P.C. Sham, S.J. Chanock, J. Wang, GWASdb: a database for human genetic variants identified by genome-wide association studies, Nucleic Acids Res. 40 (2012) D1047–D1054.
[58] E.O. Glocker, D. Kotlarz, K. Boztug, E.M. Gertz, A.A. Schaffer, F. Noyan, M. Perro, J. Diestelhorst, A. Allroth, D. Murugan, N. Hatscher, D. Pfeifer, K.W. Sykora, M. Sauer, H. Kreipe, M. Lacher, R. Nustede, C. Woellner, U. Baumann, U. Salzer, S. Koletzko, N. Shah, A.W. Segal, A. Sauerbrey, S. Buderus, S.B. Snapper, B. Grimbacher, C. Klein, Inflammatory bowel disease and mutations affecting the interleukin-10 receptor, N. Engl. J. Med. 361 (2009) 2033–2045.
[59] A. Marcuzzi, M. Girardelli, A.M. Bianco, S. Martelossi, A. Magnolato, A. Tommasini, S. Crovella, Inflammation profile of four early onset Crohn patients, Gene 493 (2012) 282–285.
[60] A.M. Bianco, V. Zanin, M. Girardelli, A. Magnolato, S. Martellossi, A. Tommasini, A. Marcuzzi, S. Crovella, A common genetic background could explain early-onset Crohn's disease, Med. Hypotheses 78 (2012) 520–522.
[61] A. Palleja, H. Horn, S. Eliasson, L.J. Jensen, DistiLD Database: diseases and traits in linkage disequilibrium blocks, Nucleic Acids Res. 40 (2012) D1036–D1040.
[62] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, J. Mol. Biol. 268 (1997) 78–94.
[63] H.V. Firth, S.M. Richards, A.P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. Van Vooren, Y. Moreau, R.M. Pettett, N.P. Carter, DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources, Am. J. Hum. Genet. 84 (2009) 524–533.
[64] A.J. Sharp, R.R. Selzer, J.A. Veltman, S. Gimelli, G. Gimelli, P. Striano, A. Coppola, R. Regan, S.M. Price, N.V. Knoers, P.S. Eis, H.G. Brunner, R.C. Hennekam, S.J. Knight, B.B. de Vries, O. Zuffardi, E.E. Eichler, Characterization of a recurrent 15q24 microdeletion syndrome, Hum. Mol. Genet. 16 (2007) 567–572.
[65] E. Klopocki, H. Schulze, G. Strauss, C.E. Ott, J. Hall, F. Trotier, S. Fleischhauer, L. Greenhalgh, R.A. Newbury-Ecob, L.M. Neumann, R. Habenicht, R. Konig, E. Seemanova, A. Megarbane, H.H. Ropers, R. Ullmann, D. Horn, S. Mundlos, Complex inheritance pattern resembling autosomal recessive inheritance involving a microdeletion in thrombocytopenia-absent radius syndrome, Am. J. Hum. Genet. 80 (2007) 232–240.
[66] S.A. Sullivan, L. Aravind, I. Makalowska, A.D. Baxevanis, D. Landsman, The histone database: a comprehensive WWW resource for histones and histone fold-containing proteins, Nucleic Acids Res. 28 (2000) 320–322.
[67] J. Ausio, Histone variants—the structure behind the function, Brief. Funct. Genomic. Proteomic. 5 (2006) 228–243.
[68] P.B. Talbert, S. Henikoff, Histone variants—ancient wrap artists of the epigenome, Nat. Rev. Mol. Cell Biol. 11 (2010) 264–275.
[69] L. Marino-Ramirez, K.M. Levine, M. Morales, S. Zhang, R.T. Moreland, A.D. Baxevanis, D. Landsman, The Histone Database: an integrated resource for histones and histone fold-containing proteins, 2011 Database (Oxford) (2011) bar048.
[70] D.P. Bartel, MicroRNAs: genomics, biogenesis, mechanism, and function, Cell 116 (2004) 281–297.
[71] D.P. Bartel, MicroRNAs: target recognition and regulatory functions, Cell 136 (2009) 215–233.
[72] V.N. Kim, J.W. Nam, Genomics of microRNA, Trends Genet. 22 (2006) 165–173.
[73] J.H. Yang, J.H. Li, P. Shao, H. Zhou, Y.Q. Chen, L.H. Qu, StarBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data, Nucleic Acids Res. 39 (2011) D202–D209.

[74] M. Barenboim, B.J. Zoltick, Y. Guo, D.R. Weinberger, MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets, Hum. Mutat. 31 (2010) 1223–1232.
[75] C.W. Lees, J.C. Barrett, M. Parkes, J. Satsangi, New IBD genetics: common pathways with other diseases, Gut 60 (2011) 1739–1753.
[76] M.G. Neuman, R.M. Nanau, Single-nucleotide polymorphisms in inflammatory bowel disease, Transl. Res. 160 (2012) 45–64.
[77] N. Azzam, H. Nounou, O. Alharbi, A. Aljebreen, M. Shalaby, CARD15/NOD2, CD14 and toll-like 4 receptor gene polymorphisms in Saudi patients with Crohn's disease, Int. J. Mol. Sci. 13 (2012) 4268–4280.
[78] S. Lesage, H. Zouali, J.P. Cezard, J.F. Colombel, J. Belaiche, S. Almer, C. Tysk, C. O'Morain, M. Gassull, V. Binder, Y. Finkel, R. Modigliani, C. Gower-Rousseau, J. Macry, F. Merlin, M. Chamaillard, A.S. Jannot, G. Thomas, J.P. Hugot, E.-I. Group, E. Group, G. Group, CARD15/NOD2 mutational analysis and genotype–phenotype correlation in 612 patients with inflammatory bowel disease, Am. J. Hum. Genet. 70 (2002) 845–857.
[79] P. Rosenstiel, K. Huse, A. Till, J. Hampe, S. Hellmig, C. Sina, S. Billmann, O. von Kampen, G.H. Waetzig, M. Platzer, D. Seegert, S. Schreiber, A short isoform of NOD2/CARD15, NOD2-S, is an endogenous inhibitor of NOD2/receptor-interacting protein kinase 2-induced signaling pathways, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 3280–3285.
[80] P. Brest, P. Lapaquette, M. Souidi, K. Lebrigand, A. Cesaro, V. Vouret-Craviari, B. Mari, P. Barbry, J.F. Mosnier, X. Hebuterne, A. Harel-Bellan, B. Mograbi, A. Darfeuille-Michaud, P. Hofman, A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease, Nat. Genet. 43 (2011) 242–245.