From gene to function:
using new technologies for solving old problems

Daniel D'Andrea

Department of Physics
Sapienza University of Rome

A thesis submitted for the degree of

Doctor of Philosophy

Doctoral school in Pasteurian Science
Ph.D programme in Molecular Biology and Medicine
Programme coordinator: Prof. Marco Tripodi
XXVI Cicle - 2010/2013

Advisor
Prof. Anna Tramontano

Abstract

Recent advances in DNA sequencing have changed the field of genomics as well as that of proteomics making it possible to generate gigabases of genome and transcriptome sequence data at substantially lower cost than it was possible just ten years ago. In recent years, many high-throughput technologies have been developed to interrogate various aspects of cellular processes, including sequence and structural variation and the transcriptome, epigenome, proteome and interactome. These Next Generation Sequencing (NGS) experimental technologies are more mature and accessible than the computational tools available for individual researchers to move, store, analyse and present data in a user-friendly and reproducible fashion. My research work is placed in this scenario and focuses on the analysis of data produced by NGS technologies as well as on the development of new tools aimed at solving the different problems that arise during NGS data analysis. In order to achieve this aim, my group and I have dealt with several open biomedical problems in collaboration with different research groups of the Sapienza University. Some of these experiments have already given interesting results but mostly have represented the occasion and starting point for the development of new tools able to improve some crucial steps of the analyses, solve problems derived by the system complexity and make the results easier to understand for the researchers. Some examples are IsomirT, a tool for the small RNA-Seq analysis and isomiR identification, Phagotto, a tool for analysing deep sequencing data derived from phage-displayed libraries and FIDEA, a web server for the

functional interpretation of differential expression analysis. Recent reports have demonstrated that individual microRNAs can be heterogeneous in length and/or sequence producing multiple mature variants that have been dubbed isomiRs. IsomirT is a useful tool to improve and simplify the search for isomiRs starting directly from the results of a miRNA-sequencing experiment. By using it, we observed the behaviour of isomiRs in different cell types and in different biological replicates. Our results indicate that the distribution of the microRNA variants is similar among replicates and different among cells/tissues suggesting that the isomiRs have a functional role in the cell. The use of the NGS technologies for the analysis of antibody selected sequences both using phage display libraries and in vitro selection processes is becoming increasingly popular. By using these technologies, the experimental group headed by prof. Felici has introduced a new experimental pipeline, named PROFILER, aimed at significantly empowering the analysis of antigen-specific libraries. A key step to exploit this idea has been to develop a new tool, Phagotto, for processing and analysing the data derived by sequencing. PROFILER, in combination with Phagotto, seems ideally suited to streamline and guide rational antigen design, adjuvant selection, and quality control of newly produced vaccines. The publicly available web server FIDEA allows experimentalists to obtain a functional interpretation of the results derived from differential expression analysis and to test their hypothesis quickly and easily. The tool performs an enrichment analysis i.e. an analysis of specific properties that are distributed in a non random fashion in the up-regulated and down-regulated genes, taken both together and separately. It has been shown to be very useful and is being heavily used from scientists all over the world, more than 1500 requests for analysis have been submitted to the server in six months. Furthermore, during the course of the PhD I implemented pipelines for the speeding up and optimization of protocols for NGS data analysis and applied them to biomedical projects. Of course not all the proteins have a complete functional annotation and consequently the issue of predicting the function of proteins with a partial or no functional annotation arises.

This can be done both by exploiting the 3D structure of the protein or by inferring the function directly from the sequence. A real challenge, however, is the assessment of the accuracy of existing methods. In this context the help that critical assessment experiments can give is essential. We have had the possibility to be involved, as assessors, in the world wide experiment CASP (Critical Assessment of protein Structure Prediction). In particular, we are involved in the assessment of the residue-residue contacts in which the participant groups provide a list of predicted contacts between residues that hopefully can be used as constraints to fold the protein. We proposed and implemented new methodologies to understand which method works better and where future efforts should be focused.

To Gino and Rosaria.

Contents

List of Figures **v**

List of Tables **vii**

1 Introduction **1**
1.1 The next generation sequencing technologies 1
1.2 Functional interpretation of differentially expressed genes 2
1.3 Protein function prediction 4
1.4 Aim of the work 5

2 NGS technologies: applications and challenges **9**
2.1 RNA-Sequencing 10
2.1.1 Background 10
2.1.2 Methods 11
2.1.3 Application: long non coding RNAs and muscle differentiation 14
2.2 small RNA-Sequencing 21
2.2.1 Background 21
2.2.1.1 IsomiRs 22
2.2.2 IsomirT: a tool for the small RNA-Seq analysis and identification of miRNA variants 22
2.2.2.1 Methods 25

2.2.2.2 Application: tissue/cell specific isomiRs in brain development 30

2.3 ChIP-Sequencing 39

2.3.1 Background 39

2.3.2 Methods 39

2.3.3 Application: identification of target genes for the Hepatitis B Viruses X protein 45

2.4 Deep sequencing of phage display libraries to support antibody discovery 50

2.4.1 Background 50

2.4.2 Phagotto: a tool for analysing deep sequencing data of phage-displayed libraries of peptides 52

2.4.3 Results 54

3 FIDEA: server for the Functional Interpretation of Differential Expression Analysis 61

3.1 Background 61

3.2 Description 63

3.3 Implementation 69

4 Evaluation of residue-residue contact prediction in CASP10 71

4.1 Background 71

4.2 Methods 74

4.3 Results 75

5 Conclusions 85

6 Appendices 87

6.1 Table of all analysed experiments 88

6.2 Pubblications 91

References 113

List of Figures

2.1 Overview of the RNA-Seq analysis pipeline 13
2.2 Classification of the assembled transcripts 15
2.3 Expression of myogenic markers in the progression of prolifer- ating myoblasts into differentiation stages 16
2.4 Functional analysis on Gene Ontology Biological Process (heat map) 18
2.5 Functional analysis on Gene Ontology Biological Process (word cloud) 19
2.6 Functional analysis on Gene Ontology Biological Process (bubble heat map) 20
2.7 Schematic representation of miRNA variants 23
2.8 Example of alignment on pre-miRNA 24
2.9 IsomirT pipeline 27
2.10 New standard nomenclature for isomiRs 28
2.11 Example of isomirT output 29
2.12 IsomiR profiles in the different cell types 31
2.13 Non-negative matrix factorization 32
2.14 Shannon entropy of tissue specificity 33
2.15 Mutation profiles in different tissues 35
2.16 Percentage of mutations in seed 36
2.17 Occurrences of the main variants 38
2.18 Underlying idea of NED 42
2.19 Scheme of the database used to store ChIP-Seq data 44
2.20 Classification of the peaks considering genomic features of interest . 47

2.21 Classification of the peaks considering the chromatin states segmentation . . . 48
2.22 MicroRNAs possibly associated to the HBx protein . . . 49
2.23 Phagotto pipeline . . . 53
2.24 Abundance of natural frame NadA fragments . . . 56
2.25 Amino acid frequencies of the natural frame NadA fragments . . . 57
2.26 Enrichment factor of the NadA fragments . . . 58
2.27 Amino acid enrichment of the natural frame NadA fragments . . . 59

3.1 FIDEA: statistics of Differentially Expressed Genes . . . 64
3.2 Functional analysis by using GoSlim (heat map) . . . 66
3.3 Functional analysis by using GoSlim (word cloud) . . . 67
3.4 FIDEA: intersection among lists of genes . . . 68
3.5 Database schema . . . 70

4.1 Similarity among the groups . . . 76
4.2 Position of the first correct and incorrect contact . . . 79
4.3 Prediction of inter-domain contacts for target T0658 . . . 82
4.4 Inter-domain: position of the first correct and incorrect prediction . 84

List of Tables

4.1 Results of the analysis of the group performance for long-range contacts - Descriptive Statistics scores calculated for the predictions treated in the context of the complete contact maps for long-range contacts for FM domains. The results are sorted according to the MCC score. 77

4.2 Results of inter-domain predictions - Results of the Prediction of Long-Range Contacts in Which the Contacting Residues Belong to Two Different Domains 81

6.1 All the analysed experiments during my Ph.D.. The experiments were performed on Illumina GAIIx except exp. 8 that was performed on Illumina MiSeq. 89

1

Introduction

1.1 The next generation sequencing technologies

Since the Human Genome Project ended, biomedical research has witnessed a remarkable explosion in sequencing technologies that permit to ask a large number of questions about the genome. Some questions have been already been answered at an unprecedented speed and resolution (1) with the aid of multiple technologies which have contributed to the transition (2) from Sanger sequencing (first generation) to a much higher throughput "next generation" sequencing (NGS) (3; 4). These systems use different chemistry and offer unique benefits, which have been summarized elsewhere (5; 6). The common idea of these technologies is that following nucleic acid fragmentation and a series of ligation reactions and amplification steps, sequencing by synthesis is performed and millions of 35-400 bp (base pair) sequences, named reads, are created. Even though the reads are shorter than the sequences generated by Sanger system, these technologies have the advantages to generate gigabases of genomic data in a single experiment (5) and at substantially lower cost and time than was possible just ten years ago (7). These high-throughput technologies have been applied to investigate various aspects of cellular processes, including the transcriptome, epigenome, proteome and interactome (8), and therefore they have changed the field of genomics as well as that of proteomics (7). Whole transcriptome sequencing (RNA-Seq), used to decipher gene expression patterns, has become the most popular application and many articles (9; 10; 11; 12; 13) have asserted their advantages over the microarray platforms. Moreover, RNA-Seq can identify transcripts arising from gene fusion events (which are typical in cancer (14)) and can detect novel classes of non-coding RNAs (ncRNAs) such as large intergenic non-coding RNAs (lincRNAs) (15). Furthermore small RNA transcriptomes can be sequenced by using this system that can identify expression patterns as well as unveil novel microRNAs and other short RNAs (16). Chromatin immunoprecipitation coupled with deep sequencing (or ChIP-Seq) has also been used to isolate genomic regions that are bound by transcription factors or other DNA binding proteins (17) or to identify regions marked by histone modifications. In addition, high-throughput methods are being used to look at DNA looping or methylation patterns (i.e. MeDIP-Seq) (18), to provide a snapshot of long-range interactions among regions of DNA (chromosome confirmation capture) (19) or to define genetic and signalling pathways (i.e, through large-scale RNAi screens) (20; 21; 22; 23; 24).

Nowadays NGS technologies are more mature and accessible than the computational tools available to individual researchers to move, store, analyse and present data in a user-friendly and reproducible fashion (25).

My research work is placed in this scenario and focuses on the analysis of the data produced by NGS technologies, as well as on the development of tools able to improve some critical steps of the analyses, solve problems derived by system complexity and provide user friendly results.

1.2 Functional interpretation of differentially expressed genes

Analysis of high-throughput data typically provides scientists with a list of hundreds to thousands of differentially expressed genes or proteins. Although this list is extremely useful for the identification of genes that may be involved in a given phenomenon or phenotype, it often fails to provide a functional interpretation of the underlying biology of the condition being studied (26). The advent of high-throughput technologies permits to face a relevant issue, i.e. to extract the biological meaning and interconnections in a long list of differentially expressed genes. The standard approach consists in identifying groups of genes that operate in the same biological processes or pathways or have the same components or structures. This allows the identification, among the listed genes, of the functional categories to which they belong to. This approach is very useful for two reasons. First, it reduces the complexity of the analysis considering only a hundreds of functional categories for the experiment. Second, identifying active functional categories that differ between two conditions can have more explanatory power than a simple list of different genes or proteins (27). The first step of the functional analysis is often an enrichment analysis that is usually performed by statistically assessing whether a pathway or process is enriched in a given list of genes (26). In general an enrichment analysis on differentially expressed genes uses one or more variations of the following strategy (28; 29; 30; 31; 32; 33; 34). For each functional category, the genes that are involved in the specific category are counted. This process is repeated for an appropriate background (e.g., all expressed genes in a RNA-Seq experiment). Finally, all functional categories are tested for over- or under-representation by using a hyper-geometric, chi-square, or binomial distribution (35; 36). An enrichment analysis can be performed using different annotations of the genes such as KEGG pathways (37), Interpro (38), Gene Ontology Molecular Function, Biological Process and Cellular Component categories (39). Surely, the correct interpretation of the results is much more effective if associated to a deep knowledge of the biological system at hand and should, therefore, be preferentially done by the experimentalists. However, the experimentalists are not always sufficiently expert in exploiting different tools and databases and comparing results obtained from different experiments, although numerous public servers (such as DAVID (35; 40), g:Profiler (41; 42), Gorilla (43), High-Throughput GoMiner (44), Babelomics (45) and GeneCodis 3 (46)) allow functional enrichment analysis given a list of genes. For this reason I decided to provide a more user-friendly tool for the analysis of data from a functional view point. The system is called FIDEA and will be described in detail in Chapter 3.

Another issue regarding the functional characterization is that, despite the large number of annotations available, a surprisingly amount of genes is still not annotated and many of the existing annotations are incomplete or inaccurate. For instance, for Homo Sapiens, ∼50% of the GO annotations in the October 2013 release are not curated manually (since they are characterised by an evidence code inferred from electronic annotations) and have a lower quality than those annotations derived by direct experimental evidence (47).

Despite many efforts, a manual annotation of the entire genome is expected to take a very long time (∼13-25 years) (48). Therefore the issue of predicting the function of proteins with a partial or no functional annotation arises and consequently improving the coverage and accuracy of functional annotation has now become the real challenge. In general this can be done by inferring the function directly from the sequence or by exploiting the knowledge or a computational prediction of the 3D structure of the protein.

1.3 Protein function prediction

The definition of biological function is ambiguous and the exact meaning of the term varies on the bases of the context in which it is used (49; 50; 51). Generally, it describes biochemical, cellular and phenotypic aspects of the molecular events which involve the protein and how the protein interacts with the environment (52). However, due to the difficulties and the costs of the experimental characterization of protein function, it is not possible to perform a functional assay for every uncharacterised gene in every genome (53) and to keep up with the influx of data by manually curated annotation.

Given this state of affairs, the computational annotation of protein function has emerged as a problem at the forefront of computational and molecular biology and scientists have been turning to sophisticated computational methods for assistance in annotating the huge volume of sequence and structure data being produced (54). Although many solutions have been proposed in the last decades (55; 56; 57; 58; 59; 60), in many circumstances the computational functional inference still relies on traditional approaches such as those based on identifying basic local alignment hits among proteins with experimentally determined functions (52) (i.e. by using BLAST (61)). In this approach, also known as homology-based transfer, the researcher aims at finding a significant sequence similarity to another protein that is already in a public database and whose function has been experimentally characterized. The rationale is that if a sequence with high degree of similarity exists, then the sequences have evolved from a common ancestor and they might have similar functions if they can be identified as orthologous (i.e. derived from speciation events not involving duplications). As databases grow in the number of sequences they hold, homology-based transfer begins to break-down in several aspects (such as the observation that even with a high sequence similarity annotation transfer may be erroneous) as described by Friedberg (54). On the other hand the structure of a protein is more informative and is better preserved than the amino acid sequence allowing us to detect a structural similarity even when a high degree of sequence similarity is not present (62). If two proteins share similar 3D structures then a similarity in their functions is more likely (63; 64). Therefore the issue is whether, by using a good structure prediction method, we are able to identify structural features in a given protein and integrate that information into a functional prediction (65). Clearly it is crucial that the accuracy with which these methods predict the protein structure be high and in this context the help that critical assessment experiments can give is essential. In this thesis I am going to briefly discuss my involvement in the world wide experiments CASP, the Critical Assessment of protein Structure Prediction.

1.4 Aim of the work

My research work started with the study, application and analysis of the new techniques of sequencing. Therefore in the first part of my PhD I focused on the analysis of data produced by NGS technologies as well as on the development of new tools aimed at solving the several problems that arise from NGS data analysis. In order to carry out this work, I had the possibility to deal with very interesting problems regarding open biomedical issues. These problems were handled in collaboration with the groups of professor I. Bozzoni, and professor M. Levrero of the Sapienza University of Rome and in collaboration with the group of professor F. Felici of the University of Molise. The project in collaboration with Prof.

Bozzoni group aims at evaluating the contribution of lncRNAs in the molecular circuitries controlling myogenesis. In particular, cellular biology and advanced high-throughput RNA sequencing techniques are used to get a more comprehensive catalogue of muscle specific lncRNAs and to decipher how these molecules regulate gene expression/chromatin dynamics. For this purpose C_2C_{12} undifferentiated murine myoblasts were compared with C_2C_{12} murine myoblasts at three differentiation times: one day, three days and five days.
The project in collaboration with Prof. Levrero group focuses on HBV X protein (hepatitis B virus X protein (HBx)). HBx is an essential factor for viral replication, indeed it has been considered to be one of the most important causes of HBV-induced hepatocarcinogenesis. For this reason HBx might affect viral functions, as well as host cell functions, by modulating a wide variety of cellular processes, including transcription, cell cycle progression, DNA damage repair, and apoptosis (66). The specific aims of the project are to identify cellular target genes of HBx and to validate and refine existing hepatocellular carcinoma molecular signatures. High-throughput sequencing of anti-HBx ChIP-enriched DNA fragments (ChIP-Seq) was performed on wild type, mock and HBx-mt monomeric linear full length HBV DNA HepG2 transfected cells.
Finally, in collaboration with Prof. Felici group, sequencing of phage-displayed antigen-specific libraries using a high-throughput platform was performed. The principal aim is to identify and rank (according to their immunoreactivity) the epitopes of a model antigen, the meningococcal NadA virulence factor, using serum samples from vaccinated individuals.
For all these projects, in order to validate and to compare different approaches I implemented particular work-flows which have the possibility of modifying the parameters and to automatically compare the results with those of previous analysis performed with different parameters and/or different versions of the tools. Moreover I developed new tools in order to personalize or improve some crucial steps of the analyses, solve problems derived by the system complexity and make the results easier to understand for the researchers. Some examples are IsomirT, a tool for the small RNA-Seq analysis and isomiRs identification, FIDEA (67), a web server for the functional interpretation of differential expression analysis and Phagotto, a tool for the analysis of phage-displayed libraries of peptides built using deep sequencing. Moreover I have had the possibility to be involved, as

assessor, in the 10 th round of the Critical Assessment of protein Structure Prediction (CASP). In this world wide experiment I participated in the assessment of the residue-residue contacts in which the participant groups provide a list of contacts between residues that hopefully can be used as constraints to fold the protein. I proposed and implemented new methodologies to understand which method works better and which method works worse (68).

2

NGS technologies: applications and challenges

The recent development of next-generation sequencing technology has brought new opportunities and challenges to the field of bioinformatics. As tens of millions of short sequences, namely the reads, can be generated in a very short amount of time, efficient processing and analysis of sequencing data has become critical. Furthermore, combining sequencing with other experimental techniques, a number of approaches have been developed and several of these, although not yet completely satisfactory, have already proven their power in the vast amount of data they provided. One of the primary reasons for developing high-throughput technologies was the sequencing and resequencing of whole genomes with the aim of identifying variations and mutations that can be associated with a phenotype. Now, one of the most popular applications of these technologies is the study of the transcriptome through sequencing of RNAs (RNA-Seq). Compared to other RNA measuring technologies, such as qPCR and microarrays, RNA-Seq has higher throughput and lower noise and it can measure the expressions of tens of thousands of genes simultaneously in a single experiment in only a few days. Small RNA transcriptomes can also be sequenced in this way and can identify expression patterns as well as unveil novel microRNAs and other short RNAs (16). Genome-wide mapping of protein-DNA interactions and epigenetic marks is also essential for a full understanding of transcriptional regulation (69). Therefore the chromatin immunoprecipitation followed by sequencing (ChIP-Seq) used to isolate genomic regions that are bound by transcription factors or other proteins, or to identify the regions marked by histonic modifications are also exploded. They have broadened our understanding for deciphering the gene regulatory networks that underlie various biological processes (70). However, many problems associated with RNA-Seq or ChIP-Seq are still open and a variety of tools to serve for different purposes in NGS data analysis are yet to be developed. An example is given by RNA-Seq that is a complex multi-step process in which a nucleic acid is repeatedly transformed and modified: from RNA to cDNA followed by multiplexed PCR amplification. In this case, the available bioinformatics pipelines still introduce errors due to these transformations affecting the transcriptome assembly, normalization or differential expression analysis.

In this chapter I will describe some of the pipelines for analysing NGS data that I have implemented and automatized with the possibility of modifying the parameters and to automatically compare the results with those of previous analysis performed with different parameters and/or different versions of the tools. Moreover I will introduce some new tools and work-flows in order to personalize some crucial steps of the analyses, solve problems derived by the system complexity and make the results easier to understand for the researchers. Finally I will show some applications of these pipelines to open biological issues or to interesting biomedical problems in which my group and I have been involved.

2.1 RNA-Sequencing

2.1.1 Background

In a typical RNA-Seq experiment a sample of RNA is converted to a library of complementary DNA (cDNA) fragments and then sequenced on a high-throughput sequencing platform, such as Illumina Genome Analyzer, SOLiD or Roche 454 (71). Tens of millions of reads are obtained from this sequencing and then mapped to a reference genome or transcriptome. For each sample the mapped reads are assembled into gene-level, exon-level or transcript-level, depending on the aims of the experiment, whereas the unmapped reads are usually discarded. The quantity of mapped reads to a given gene/exon/transcript measures the expression level for this region of the genome or transcriptome. Compared with microarrays, which have been the dominant approach for studying gene expression in the last two decades, RNA-Seq technology has a wider measurable range of expression levels, less noise, higher throughput and more information to detect allele-specific expression, novel promoters, and isoforms (72). For these reasons, RNA-Seq is gradually replacing the array-based approach as the major method for the gene expression studies. Meanwhile, after around 5 years there is not yet a standard pipeline to analyse RNA-Seq data and it seems that the technology is faster then the computational methods used to analyse it.

2.1.2 Methods

In order to perform a RNA-Seq experiment, full length mRNA transcripts of different genes are extracted from the sample, retrotranscribed, fragmented into small pieces, filtered for suitable lengths, and finally sequenced on the machine. In this context I will not show the experimental details that are not relevant to our computational problems and that are explained in several works (6; 9; 73; 74; 75).
The RNA-Seq analysis is performed in three principal steps (as showed in Fig. 2.1): the alignment, the transcriptome assembly and the differential expression analysis. Depending on whether a reference genome assembly is available, current transcriptome assembly strategies generally fall into one of three categories: a reference-based strategy, a de novo strategy or a combined strategy that merges the two. In our cases, where a reference genome for the target transcriptome is available, the transcriptome assembly can be built upon it. Therefore, for each sample, the reads are first mapped to the reference genome using an aligner. After the alignment, overlapping reads from each locus are clustered to build a graph that then will be traversed to join compatible reads together into isoforms. Finally the differential expression analysis is performed comparing the assembled transcriptomes. Although there are several packages that may be used for the different phases of the RNA-Seq analysis (see (76; 77; 78) for the alternative read-alignment programs,

(79; 80) for transcriptome reconstruction and (81; 82; 83) for quantification and (84; 85; 86) for differential expression) I decided to use the pipeline proposed by Trapnell *et al.* (87; 88) in which the used tools are gaining wide acceptance and have been used in a number of recent high-resolution transcriptome studies such as the ENCODE project (89). In this protocol TopHat (90) aligns the reads to the genome and discovers sites of splicing. Cufflinks (91) uses this alignment against the genome to assemble the reads into transcripts and Cuffmerge merges all the assembled transcriptomes derived by different samples. The aligned reads and the assembled transcriptomes were finally used as input for Cuffdiff, a part of the Cufflinks package, to obtain the expression levels in FPKM (Fragments Per Kilobase per Million mapped reads) and to determine the genes and transcripts that are differentially expressed.

Replication, randomization and blocking are essential components of any well planned and properly analysed design (92), particularly in RNA-Seq experiments where the biological replicates are fundamental to estimate the mean and variance of the transcript expression. Consequently in one of the experiments that I have analyzed and that was designed without replicates (Experiment 1 in Appendix 6.1), the differentially expressed genes were only sorted according to their expression and Fold Change (the expression ratio between two samples). In particular, in each pairwise comparison between samples, I calculated the fold change for transcripts expressed in both samples and selected those with a fold change greater than 1. The resulting list was used to compute a threshold value corresponding to the third quartile of the data. Transcripts with fold change above the threshold were ordered according to the corresponding fold change. Transcripts not expressed (or with a value of FPKM lower than 0.1) in one of the samples and with a FPKM > 1 in the other were ordered according to the FPKM in the latter sample. The same procedure was used to select the genes used for the functional classification.

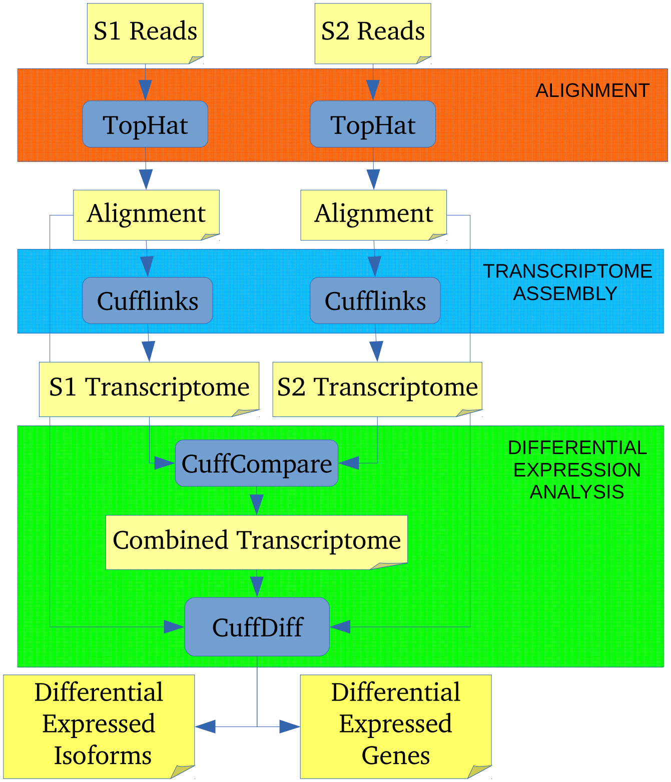


Figure 2.1: Overview of the RNA-Seq analysis pipeline - A typical pipeline for two theoretical biological samples is shown with the colours that indicate the principal steps of the analysis. The input/output files are shown in yellow rectangles; software tools or methods for each step are shown in blue rounded rectangles. First, the reads are mapped to the reference genome or transcriptome by using TopHat; mapped reads are assembled into the transcriptome by using Cufflinks. The two independent transcriptomes are then merged in order to create a combined transcriptome which is used as reference to perform a differential expression analysis. These last two steps are performed by using two methods of Cufflinks package, CuffCompare and CuffDiff. The pipeline produces a list of genes and isoforms with associated expressions, P-values and fold changes.

2.1.3 Application: long non coding RNAs and muscle differentiation

Although many studies have helped unveiling the function of many small non-coding RNAs, very little is known about the long non-coding (lncRNA) counterpart of the transcriptome. Thanks to the availability of sensitive detection techniques, in spite of their very low levels of expression in particular body compartments, specific patterns of lncRNA expression in different cell types, tissues and developmental conditions (93; 94) have been defined. Muscle differentiation is a powerful system for these investigations, because it can be both recapitulated in vitro and because the networks of transcription factors coordinating the expression of genes involved in muscle growth, morphogenesis, and differentiation are well known and evolutionarily conserved (95).

Myogenesis is an orderly and continuous process involving self-renewal, cell fate choice, proliferation and differentiation. The expression of the molecules that are responsible for coordinating this process must be tightly regulated in time and space to prevent myopathies.

In the last years several miRNAs with a specific role in both normal muscle differentiation and degenerative processes were identified (96; 97; 98). Recently, Cesana *et al.* discovered a long non-coding RNA, named linc-MD1, cross-regulating specific mRNAs by competing for miRNA binding *via* their miRNA recognition motifs. These findings opened the intriguing possibility that lncRNA-based mechanisms might influence different sets of transcripts during the execution of crucial metabolic pathways, as cellular differentiation programs.

The study, in which we are interested, aims at evaluating the contribution of lncRNAs in the molecular circuitries controlling myogenesis. In particular, cellular biology and advanced high-throughput RNA sequencing techniques are used to get a more comprehensive catalogue of muscle specific lncRNAs and to decipher how these molecules regulate gene expression/chromatin dynamics. For this purpose C_2C_{12} undifferentiated murine myoblasts (D0) were compared with C_2C_{12} murine myoblasts at three differentiation times: one day (D1), three days (3D) and five days (5D).

The assembled transcriptome of these cells consisted of 72,326 expressed transcripts (with a FPKM > 0.1) corresponding to 22,115 unique gene loci. About 69%

of these transcripts (known transcripts) were annotated in the Ensembl (99; 100) reference transcriptome (Release 74), 23% corresponded to possible new isoforms of known transcripts (novel isoforms) and 1% matched with antisense transcripts or exhibited a partial overlap to known genes (others). The remaining 7% of the analysed transcriptome did not correspond to any of the Ensembl reference transcripts (Fig. 2.2).

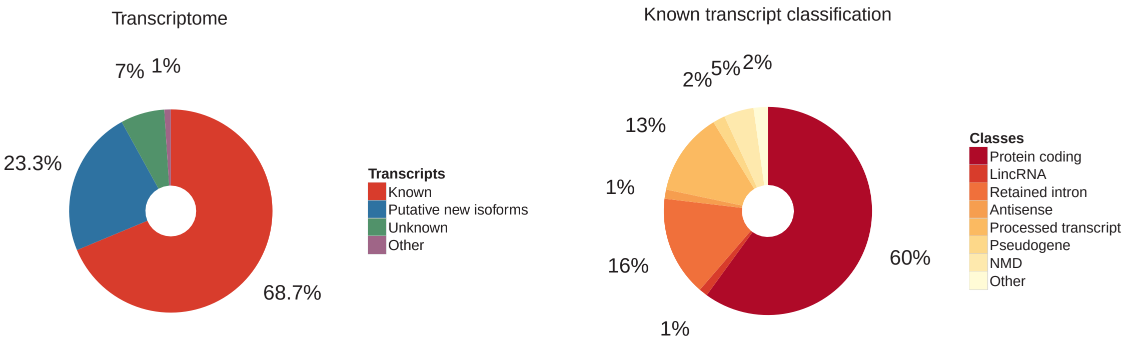


Figure 2.2: Classification of the assembled transcripts - The assembled transcriptome annotated using Release 74 of Ensembl (left) and classification of the known transcripts (right).

The ordered progression of proliferating myoblasts into differentiation stages was monitored by the expression of myogenic markers such as myogenin, MHC, Mef2c, MCK and Myod1. The observed profile reflects the expected modulation of myogenic markers, with the expression of Myod1 and myogenin at the early stages of miogenesis and the appearance of MHC, Mef2c, MCK occurring later during differentiation (Fig. 2.3).

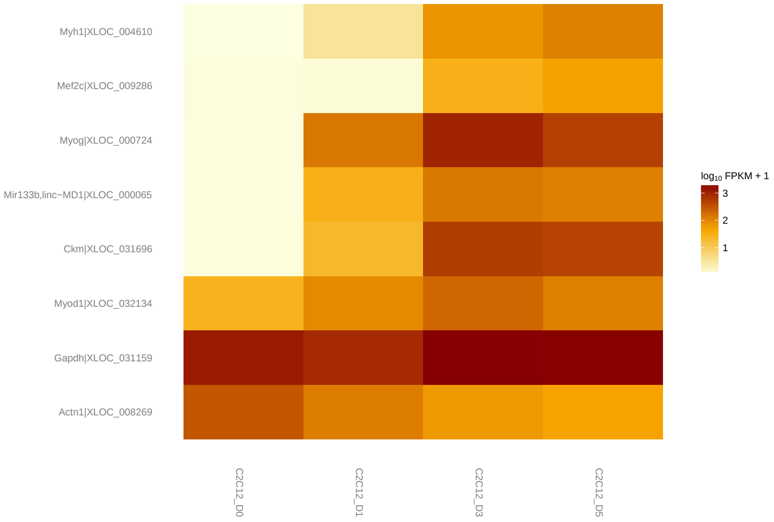


Figure 2.3: Expression of myogenic markers in the progression of proliferating myoblasts into differentiation stages - Heat-map shows the expression (in FPKM) of MHC, MEF2c, myogenin (MYOG), MCK and MyoD1 transcripts in C_2C_{12} myoblasts grown in proliferative (D0) or differentiative (D1, D3, D5) conditions.

A more complete and comprehensive view of the biological processes can be obtained by moving the interest from the examined markers to all differentially expressed protein-coding genes and by investigating on their functional characterization. This can be done by using FIDEA (67), a web server for the functional interpretation of the results derived from differential expression analysis. The figures 2.4 and 2.5 show the results of the functional analysis on the biological processes annotated in Gene Ontology considering the differentially expressed genes between D0 and D5. The heatmap (Fig. 2.4) was made by considering up and down-regulated genes separately. It shows that biological processes related to muscle functionalities (i.e. *muscle system process* or *muscle contraction*) are enriched by the up regulated genes in differentiated cells. Also considering the up and down-regulated genes together we can see that biological processes related to muscle physiology are enriched. In the wordcloud (Fig. 2.5) it is possible to see that genes involved in these processes are mostly up-regulated in D5 cells (red colour). If we compare the results of functional analyses obtained considering the differentially expressed genes in D1, D3, D5 against D0 (Fig. 2.6) we can see that some processes such as *muscle system process* or *muscle contraction* are enriched by up-regulated genes in all the differentiation times; on other hand, some processes, such as *muscle cell differentiation* or *muscle cell development*, are more enriched by up-regulated genes in differentiation times D5 and D3 than D1.
According to the aim of the study we have focused our attention on differentially expressed lncRNAs in muscle differentiation. Among them, in collaboration with experimentalists, we have selected a subclass of 24 lncRNAs that are currently in the validation phase.

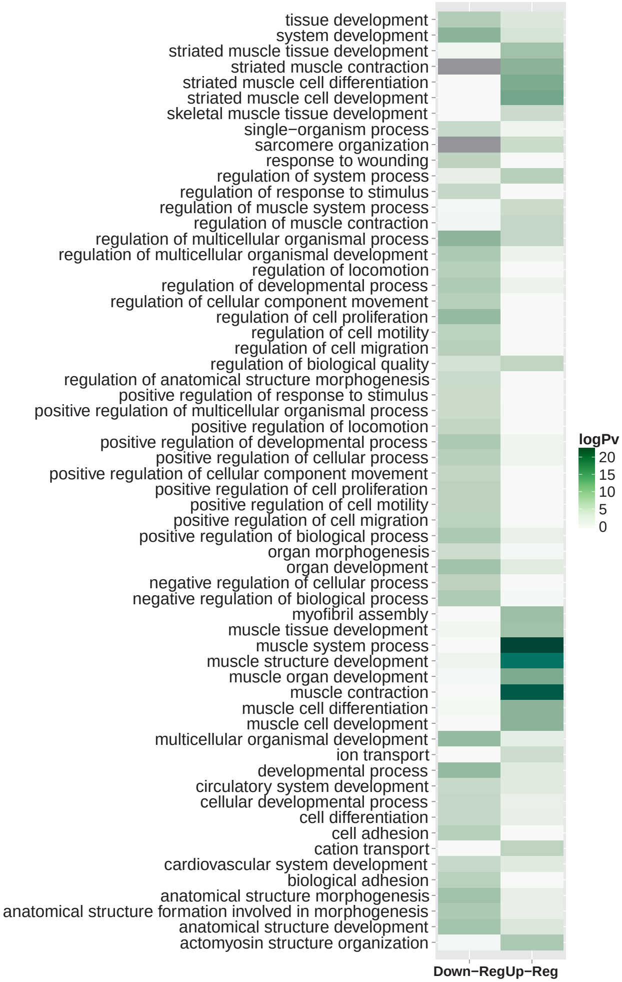


Figure 2.4: Functional analysis on Gene Ontology Biological Process (heat map) - The result of the functional analysis considering the differentially expressed genes between D0 and D5 is shown as a heat map showing the absolute log10 of the corrected P-value (colour of the cells). The functional analysis was made considering up and down-regulated genes separately.

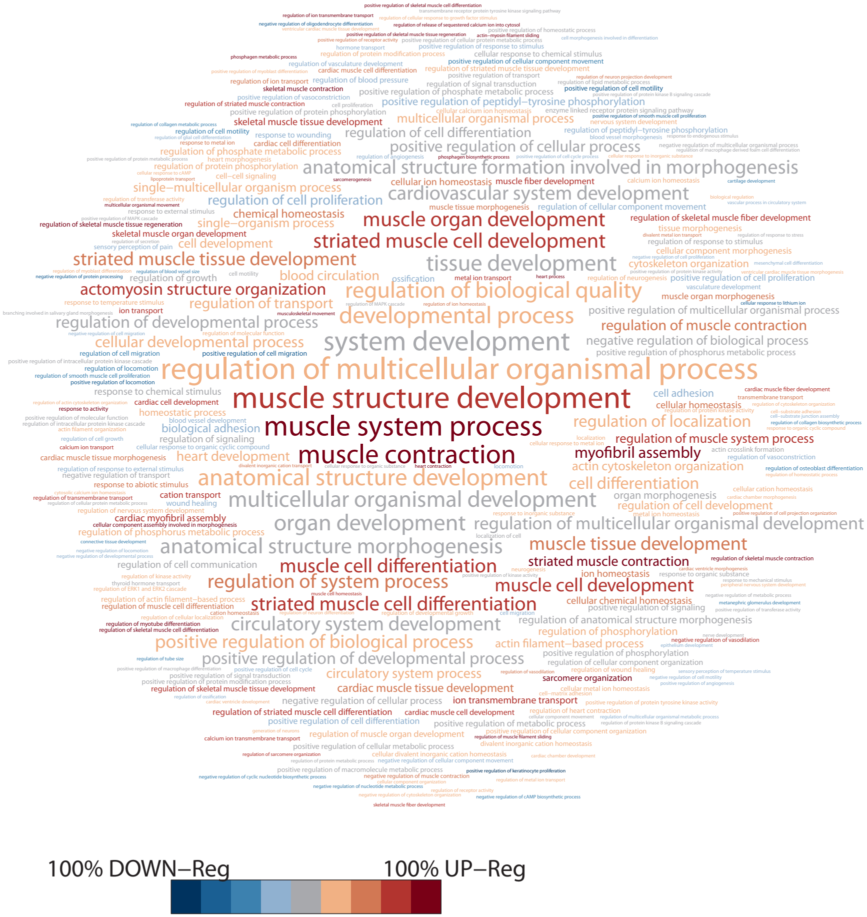


Figure 2.5: Functional analysis on Gene Ontology Biological Process (word cloud) - The result of the functional analysis considering the differentially expressed genes between D0 and D5 is shown as a word cloud where the functional categories are shown with a character size related to their enrichment (according to the corrected P-value) and in different colours according to the extent by which the pathways or categories are enriched by up- or down-regulated genes (red to blue, respectively). The functional analysis was made considering up and down-regulated genes together.

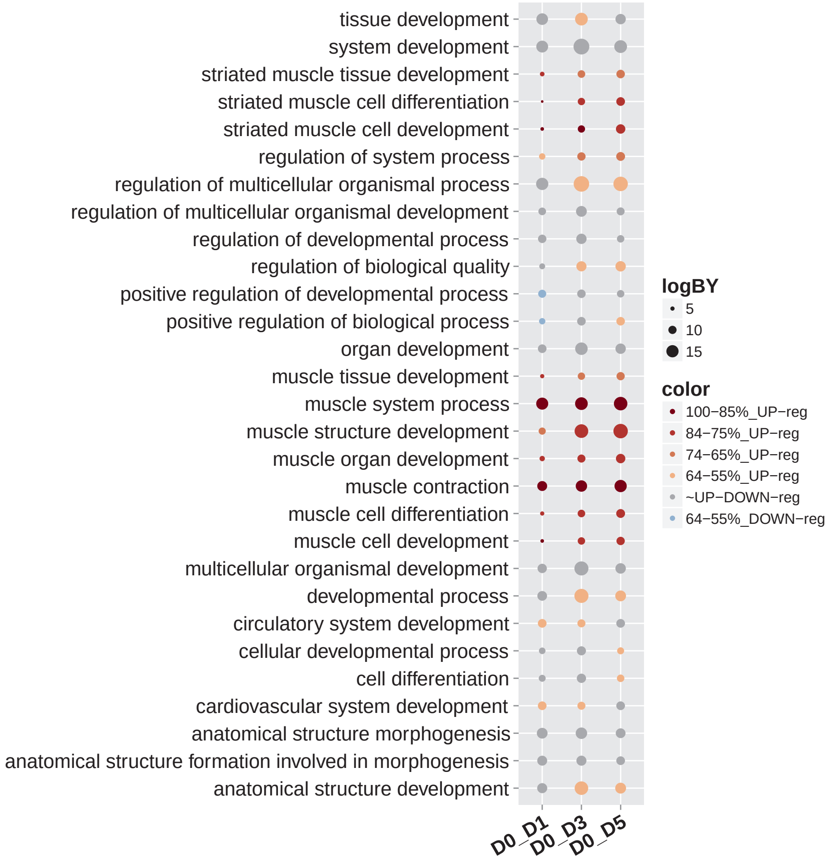


Figure 2.6: Functional analysis on Gene Ontology Biological Process (bubble heat map) - Results of the functional classification of the differentially expressed genes in D1, D3, D5 with respect to D0. Enriched functional categories are shown with a circle the size of which is proportional to the extent of enrichment (according to the corrected P-value). The different colours show the extent by which the pathways or categories are enriched by up- or down-regulated genes (red to blue, respectively).

2.2 small RNA-Sequencing

2.2.1 Background

In the last decade, one of the most significant advance in genomics has been the discovery of small (∼20-30 nucleotide [nt]) noncoding RNAs that regulate genes and genomes (101). These molecules, called MicroRNAs (miRNAs), seem to play an important gene-regulatory roles in animals and plants by guiding Argonaute (Ago) proteins to specific protein-coding sequences in transcripts (102) to direct their post transcriptional repression (103). This repression may occur at some of the most important levels of genome function, including chromatin structure, chromosome segregation, transcription, RNA processing, RNA stability, and translation (104; 105). Indeed functional studies indicate that miRNAs participate in the regulation of almost every investigated cellular process so far and that changes in their expression are associated with many human pathologies (104).

Since miRNAs only need as few as 7 nucleotides of complementarity to bind to their target, computational and experimental approaches indicate that more than 60% of human protein coding genes are predicted to contain miRNA binding sites (106; 107). Consequently, thousands of different genes may be subject to regulation by a single miRNA or miRNA family (106; 108). The most common motif is a perfect pairing between nucleotides 2 and 7 at the 5' end of the miRNA, which is called the “seed” region, and the 3' untranslated regions (UTRs) of the target site (103). Despite this, imperfect pairing of the 5' end to a target or other type of pairing have been seen (109). For example centred sites have been described, at which the middle region of the miRNA makes contiguous base pairs with a target sequence (110) or in *C. elegans* the imperfect pairing of the 5' end of the miRNA to a target is compensated by extensive 3' end interactions (111; 112; 113). The small size of mature miRNAs makes them ideal for characterization using RNA-seq technologies (73; 114). Unlike hybridization approaches such as microarray, massive-scale sequencing provides a way to profile miRNAs without a priori knowledge of expression (115; 116). Therefore deep sequencing of the small RNA fraction within cells yields an incredibly rich amount of data from which we can determine not only the expression levels of known miRNAs but also detect novel miRNAs or other small RNA species, such as piRNAs (Piwi interacting RNA) or snoRNAs (Small nucleolar RNA) (117).

2.2.1.1 IsomiRs

Typically, miRNAs are annotated as a single defined sequence, despite the fact that, for many of them, several variants with different length and/or sequence for the same miRNA have been observed (118). These variants, named isomiRs (119), were originally discarded as sequencing (120; 121) or alignment artefacts (122), poor quality or degraded RNA (123), sloppy Drosha/Dicer excision (124; 125), or simply as "trivial variants" (126), although some studies argued that measurement noise cannot account for the high frequency of these variants (127; 128). Indeed, in recent years, massively parallel sequencing and sophisticated computational algorithms have confidently confirmed the existence of isomiRs and identified a vast array of them in various species (116). Moreover, comparison of isomiR profiles across different cell lines or tissue types revealed that isomiRs are non-randomly distributed (129), suggesting that their biogenesis can be cell type specific (130; 131) and therefore they can be true physiological miRNA variants. However, despite their consistent appearance in several datasets, the biological relevance of isomiRs remains controversial and, above all, it is unclear to what extent isomiRs are functionally significant. A small but growing number of reports suggest that, in certain cases, alternative isomeric forms may have different properties, but because there are limited tools available for modulating cellular levels of specific isomeric forms and measuring their effects, many reports at this stage are suggestive rather than conclusive (132).

2.2.2 IsomirT: a tool for the small RNA-Seq analysis and identification of miRNA variants

By using deep sequencing it seems clear that the sequences of many miRNAs vary from the standard sequences (also named canonical) that are reported in public databases like miRBase (133; 134). These variants can encompass substitutions, insertions or deletions, 5' and/or 3' end templated or non-templated additions, and 5' and/or 3' cleavage variations (Fig. 2.7).

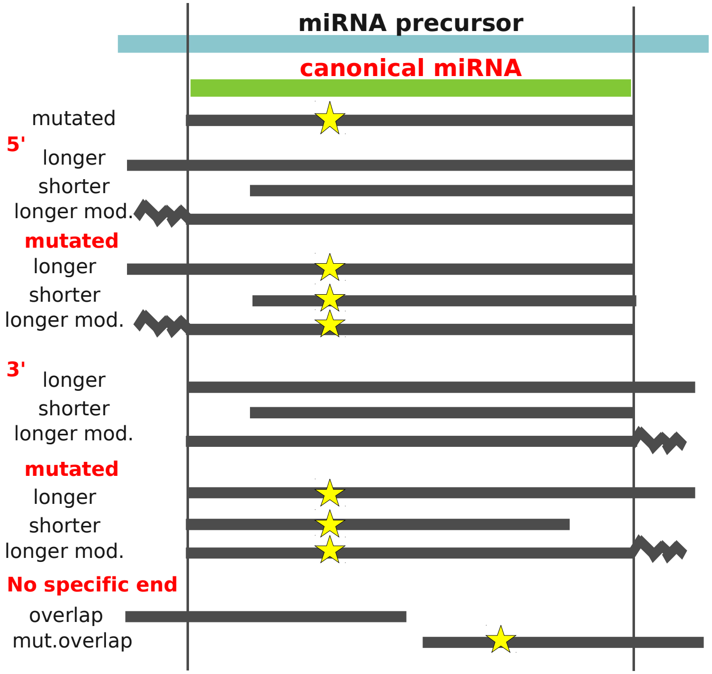


Figure 2.7: Schematic representation of miRNA variants - Example of a canonical miRNA (green) and each mutually exclusive isomiR category (gray). The non-templated forms are shown as a zig-zag line and the mutations as stars.

To understand if these variants are functional and how they are biologically relevant, my colleagues and I decided to survey multiple biological states from multiple samples. In order to improve and automate the search for isomiRs in small RNA sequenced libraries we developed a simplified tool, named isomirT, although there are some public tools that do this (in particular (135; 136; 137; 138)). The reason is that, since these tools identify the variants by aligning the reads to annotated miRNA precursors (pre-miRNA), their results refer to the pre-miRNAs. We instead need that all occurrences refer to the mature miRNAs which are the real biological players present into the cell. Furthermore an identification of mature miRNAs from pre-miRNAs is not always straightforward. In fact it may affect biases in the expression estimation, especially in the case of mature miRNAs derived by different pre-miRNAs. Considering as practical example miR-92a-3p that can be derived by two precursors, the mir-92a-1 located on chromosome 13 and mir-92a-2 on chromosome X. A read, occurring 1000 times, has a base more respect the miR-92a-3p as in Fig. 2.8. Aligning on pre-miRNAs we will have as result that the read aligns on both pre-miRNAs with 1000 occurrences. On the other hand by directly referring to mature miRNA the result will be that for miR-92a-3p a longer variant with 1000 occurrences is present.

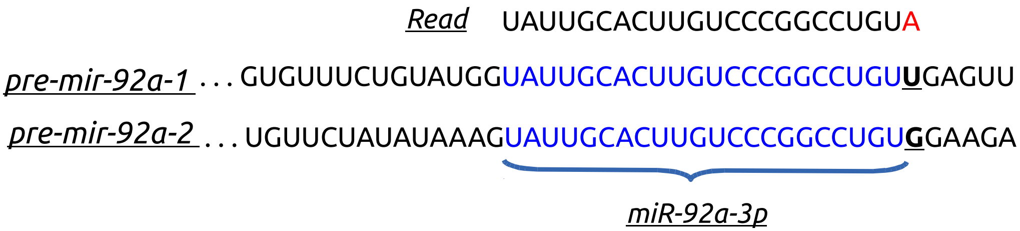

Read
UAUUGCACUUGUCCCGGCCUGUA
pre-mir-92a-1
. . . GUGUUUCUGUAUGGUAUUGCACUUGUCCCGGCCUGUUGAGUU
pre-mir-92a-2
. . . UGUUCUAUAUAAAGUAUUGCACUUGUCCCGGCCUGUGGAAGA
miR-92a-3p

Figure 2.8: Example of alignment on pre-miRNA - The blue bases represent the mature miRNA. The underlined bases represent the differences between the pre-mir-92a-1 and pre-mir-92a-2. In this example the read has an A base (red) more respect the mature miRNA and this base does not match any pre-miRNAs.

IsomirT allows the identification of templated or nontemplated 5' and/or 3' end variations and polymorphic isomiRs considering the canonical miRNA sequence as well as other regions on the same miRNA precursor.

This framework is very useful to identify and compare isomiR abundance in different sequencing libraries, providing an initial view for biologically relevant isomiRs. The tool is a stand-alone software that rapidly processes the reads (in fastq or fasta format) derived from sequencing; it gives as result all the variants for each miRNA, reporting the occurrences and the sequences with the relative variations. Furthermore graphs about general statistics are produced, such as the distributions of variants, the mutations in seeds or the type of mutations. In order to understand whether the isomiRs are functional or random modifications we investigated how they behave during differentiation or in different cell type by using several public datasets derived by recent articles (139; 140; 141).

2.2.2.1 Methods

After sequencing, the raw reads are filtered according the quality. As the average length of the reads in a small RNA-Seq experiment (∼36 nt) is greater than the average size of a miRNA (∼17-25 nt), we expect to find part of the 3'-adaptor at the 3'-end of the sequence. Therefore the reads are passed through an adaptor filter that searches for reads whose 3'-ends align to at least 6 nt of the 3'-adaptor. Finally potential contaminations are detected and filtered out. The remaining reads are formatted into a non-redundant FASTA file containing for each unique sequence the copy number and the amino acid sequence.

In order to associate each read to a specific miRNA variant we designed and implemented the work-flow showed in Fig. 2.9. For each read we verify if a *canonical* or subsequently a “canonical with mutation” (named *mutated*) form exists. For mutated forms, we allow up to one mismatch which permits us to identify miRNAs that carry mutations or that may have undergone RNA editing. Progressively we look for mutually exclusive *shorter, longer* and *overlapping* variants that exactly align on miRNA precursor. The variants that align on miRNA precursor and don't fall in previous classes are labelled as *new putative* variants. The reads do not align on any miRNA precursor but include a canonical form or overlap with one of these are identified and labelled as *longer modified* and *overlap modified* respectively (in other words they are variants with additional nucleotides that do not match any precursor miRNA sequence). The second part of the pipeline looks for the mutated variants. We consider only the mutated variants for which at least a "canonical with mutation" form was detected. We verify if the read is contained in one or more mutated canonical miRNA (named *shorter mutated* variant) or if the read contains one or more mutated canonical miRNA. In the latter case we distinguish the *mutated longer* and the *mutated longer modified* variants according to the miRNA precursor contains or not the read. Finally we detect the *mutated overlap* variants if the reads overlap with one or more mutated canonical miRNA.

In this disseration I will consider the "main variant" the variant with the most frequent occurrences in a specific miRNA. For the reference miRNAs and precursors I refer to the latest version of miRBase (release 20 at now). In order to give the relative abundance of each isomiR, the counts of the reads that have multiple alignment are equally apportioned to each variant to which they align and then they are normalized to the number of variants. We decided to discard, as potential sequence errors or artefacts, all the variants with an occurrence lower than the 5% of the occurrence of the main variant.
All the variants are reported in an output file with their sequence and a full description of the alignment on canonical miRNA, as well as possible addition or cleavage variations or mismatches with related position on canonical miRNA.
We introduced also a new standard nomenclature for isomiRs. This nomenclature is based on canonical miRNA sequence and describes the variant type, the side, the added or absent sequence, the position and the substitution for a mutated variant (Fig. 2.10). As an example: the notation *miR-124-3p_Lg-5p-A* refers to the variant longer (*Lg*) at 5' (*5p*) respect to the canonical sequence of *miR-124-3p*. In this case the addiction is the only *A* base. The *miR-124-3p_Mt-C-20-A* refers to an isomiR of *miR-124-3p* that contains a mutation (*Mt*) from *C* to *A* at position 20 in the canonical miRNA.
In Fig. 2.11 a typical output of isomirT is also shown.

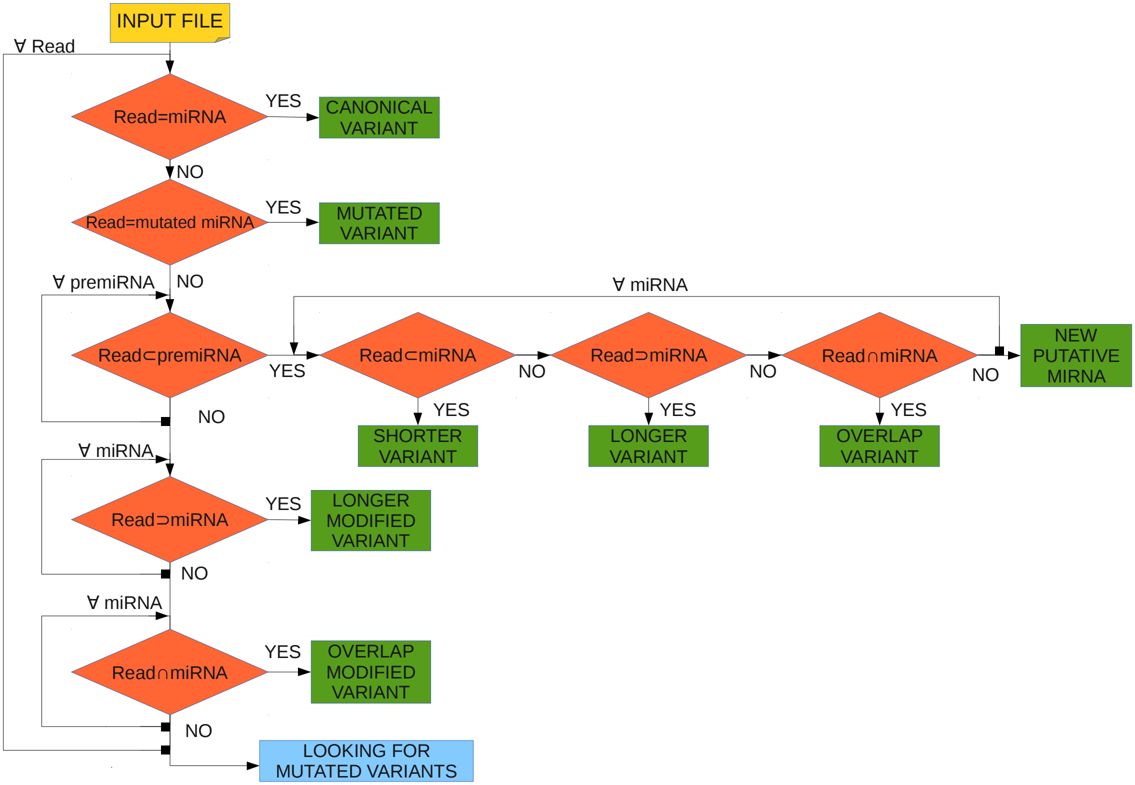


Figure 2.9: IsomirT pipeline - Schematic representation of isomirT work-flow.

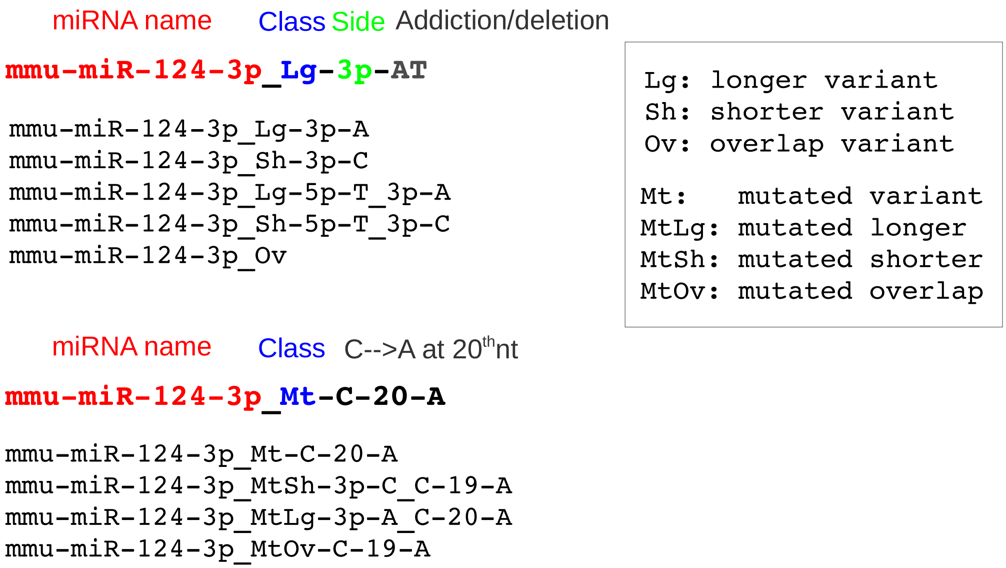
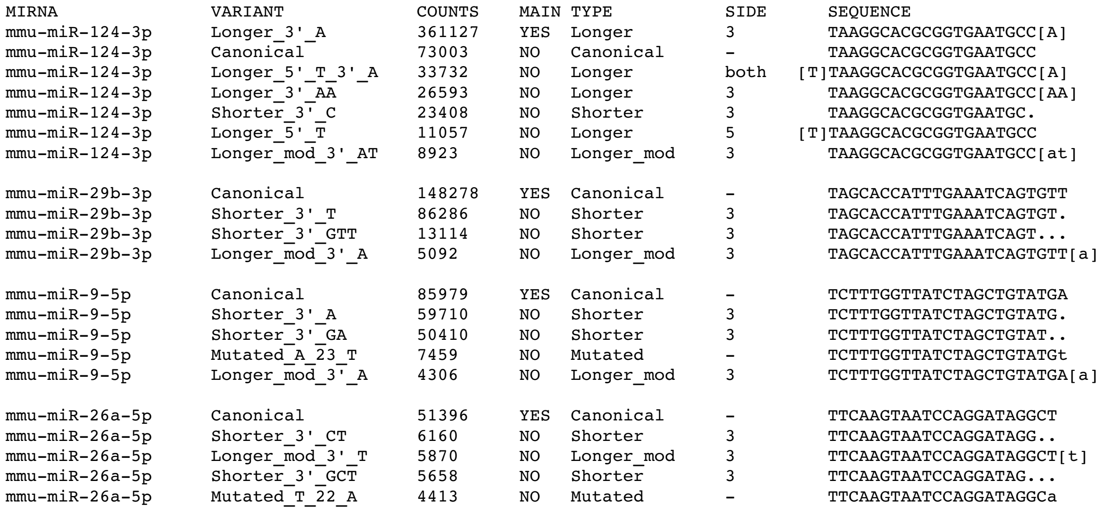


Figure 2.10: New standard nomenclature for isomiRs - This nomenclature is based on canonical miRNA sequence and describes the variant type (first part after the miRNA name), the side (first part after the variant type), the added or absent sequence (second part after the variant type), the position and the substitution for a mutated variant (first, second and third parts after variant type Mt).



```
MIRNA              VARIANT            COUNTS    MAIN TYPE           SIDE      SEQUENCE
mmu-miR-124-3p     Longer_3'_A        361127    YES  Longer         3            TAAGGCACGCGGTGAATGCC[A]
mmu-miR-124-3p     Canonical          73003     NO   Canonical      -            TAAGGCACGCGGTGAATGCC
mmu-miR-124-3p     Longer_5'_T_3'_A   33732     NO   Longer         both      [T]TAAGGCACGCGGTGAATGCC[A]
mmu-miR-124-3p     Longer_3'_AA       26593     NO   Longer         3            TAAGGCACGCGGTGAATGCC[AA]
mmu-miR-124-3p     Shorter_3'_C       23408     NO   Shorter        3            TAAGGCACGCGGTGAATGC.
mmu-miR-124-3p     Longer_5'_T        11057     NO   Longer         5         [T]TAAGGCACGCGGTGAATGCC
mmu-miR-124-3p     Longer_mod_3'_AT   8923      NO   Longer_mod     3            TAAGGCACGCGGTGAATGCC[at]

mmu-miR-29b-3p     Canonical          148278    YES  Canonical      -            TAGCACCATTTGAAATCAGTGTT
mmu-miR-29b-3p     Shorter_3'_T       86286     NO   Shorter        3            TAGCACCATTTGAAATCAGTGT.
mmu-miR-29b-3p     Shorter_3'_GTT     13114     NO   Shorter        3            TAGCACCATTTGAAATCAGT...
mmu-miR-29b-3p     Longer_mod_3'_A    5092      NO   Longer_mod     3            TAGCACCATTTGAAATCAGTGTT[a]

mmu-miR-9-5p       Canonical          85979     YES  Canonical      -            TCTTTGGTTATCTAGCTGTATGA
mmu-miR-9-5p       Shorter_3'_A       59710     NO   Shorter        3            TCTTTGGTTATCTAGCTGTATG.
mmu-miR-9-5p       Shorter_3'_GA      50410     NO   Shorter        3            TCTTTGGTTATCTAGCTGTAT..
mmu-miR-9-5p       Mutated_A_23_T     7459      NO   Mutated        -            TCTTTGGTTATCTAGCTGTATGt
mmu-miR-9-5p       Longer_mod_3'_A    4306      NO   Longer_mod     3            TCTTTGGTTATCTAGCTGTATGA[a]

mmu-miR-26a-5p     Canonical          51396     YES  Canonical      -            TTCAAGTAATCCAGGATAGGCT
mmu-miR-26a-5p     Shorter_3'_CT      6160      NO   Shorter        3            TTCAAGTAATCCAGGATAGG..
mmu-miR-26a-5p     Longer_mod_3'_T    5870      NO   Longer_mod     3            TTCAAGTAATCCAGGATAGGCT[t]
mmu-miR-26a-5p     Shorter_3'_GCT     5658      NO   Shorter        3            TTCAAGTAATCCAGGATAG...
mmu-miR-26a-5p     Mutated_T_22_A     4413      NO   Mutated        -            TTCAAGTAATCCAGGATAGGCa
```

Figure 2.11: Example of isomirT output - A typical output of isomirT.

2.2.2.2 Application: tissue/cell specific isomiRs in brain development

In order to understand whether the isomiRs are functional we investigated how they behave during differentiation or in different cell type by using several public datasets derived by recent articles (139; 140; 141) (Experiments 9, 10, 11 in Appendix 6.1). All the datasets give similar results therefore I chose to show only these relating to the study of He *et al.* in which there is a cell-type based analysis of miRNA profiles in the mouse neocortex and cerebellum (in particular there are 5 different embryonic stem (ES) cell lines in triplicate). The researhers introduced a novel miRNA tagging and affinity-purification method (miRAP). This method is based on the fact that mature miRNAs are incorporated into RNA-induced silencing complex (RISC), in which the Argonaute protein AGO2 directly binds miRNAs and their mRNA targets. Therefore by using AGO2 antibody, miRNAs in all cells of the tissue are coprecipitated with AGO2 or tAGO2 (a GFP-MYC-AGO2 fusion protein). Finally the RNAs prepared from immunoprecipitation product is subjected to deep sequencing (139). The study has the principal aim to systematically analyse miRNA expression in neurons. Furthermore it reveals the expression of a large fraction of known miRNAs with distinct profiles in glutamatergic and GABAergic neurons and subtypes of GABAergic neurons.

Observing the isomiR profiles (Fig. 2.12), the first interesting result is that the canonical miRNAs do not necessarily constitute the most expressed class of miRNA variants. Their occurrence varies among different cells/tissues and reaches at most the half of the expressed miRNAs in a given sample. Moreover the comparison of isomiR profiles among the different cell types reveals a great consistency among the biological replicates of a given sample and, at the same time, clear differences among different samples (result confirmed also by NMF method (142) (Fig. 2.13)). For example in Fig. 2.12 in Camk2a cells the canonical variants occur the 50% of the times and the 3' modifications the most of the other times. But this is different in Cerebellum cells where the most frequent variants are in 3' modification. These observations suggest that the phenomenon of isomiRNAs is cell/tissue specific and, at the same time, it exhibits strong repeatability. We calculated the Shannon Entropy and the tissue specificity (Fig. 2.14) of each miRNA variant in order to compare canonical miRNAs to other variants. From this comparison, it emerges that canonical miRNAs and other variants have similar tissue specificity.

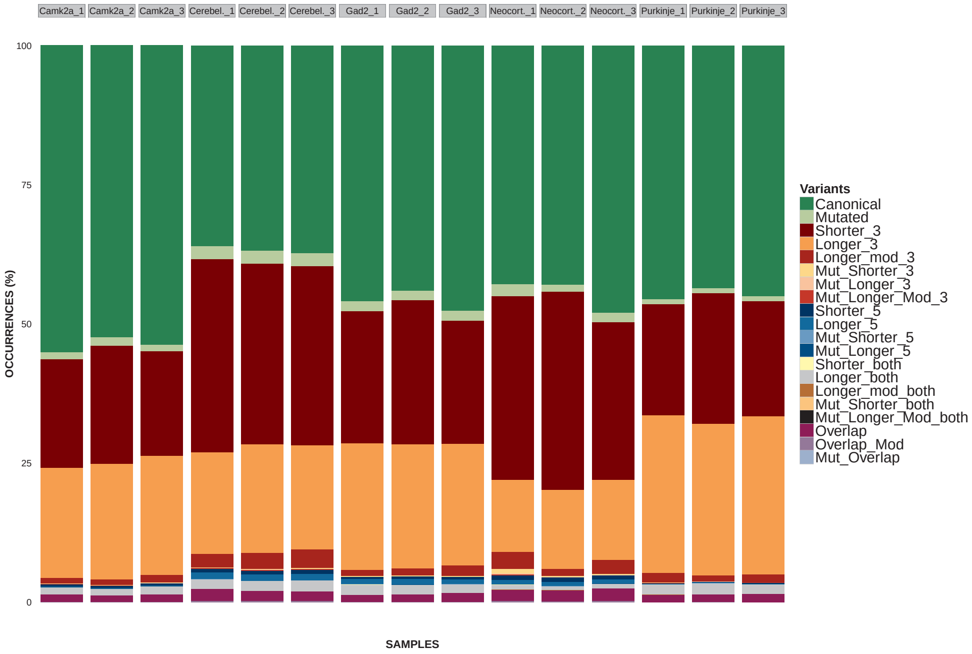


Figure 2.12: IsomiR profiles in the different cell types - Comparison of isomiR profiles among the different cell types.

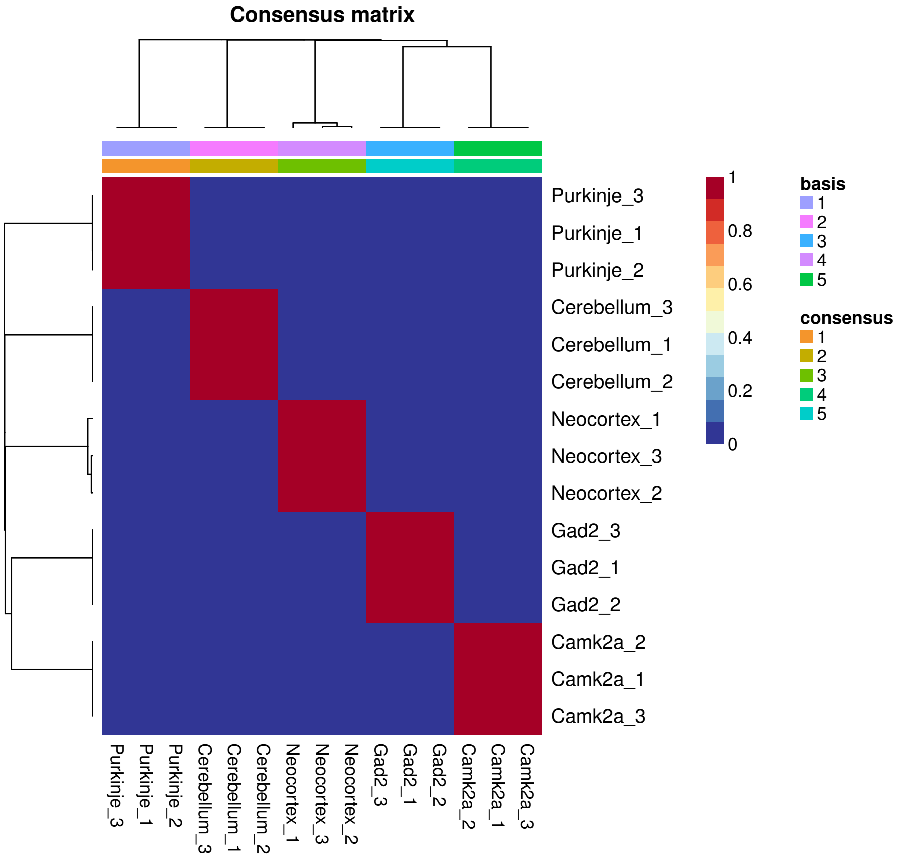


Figure 2.13: Non-negative matrix factorization - NMF computation and model selection were performed according to Brunet *et al.* (142). According to cophenetic correlation coefficients for NMF-clustered matrices, the NMF class assignment for K = 5 was the most robust. Samples are hierarchically clustered by using distances derived from consensus clustering matrix entries, colored from 0 (deep blue, samples are never in the same cluster) to 1 (dark red, samples are always in the same cluster).

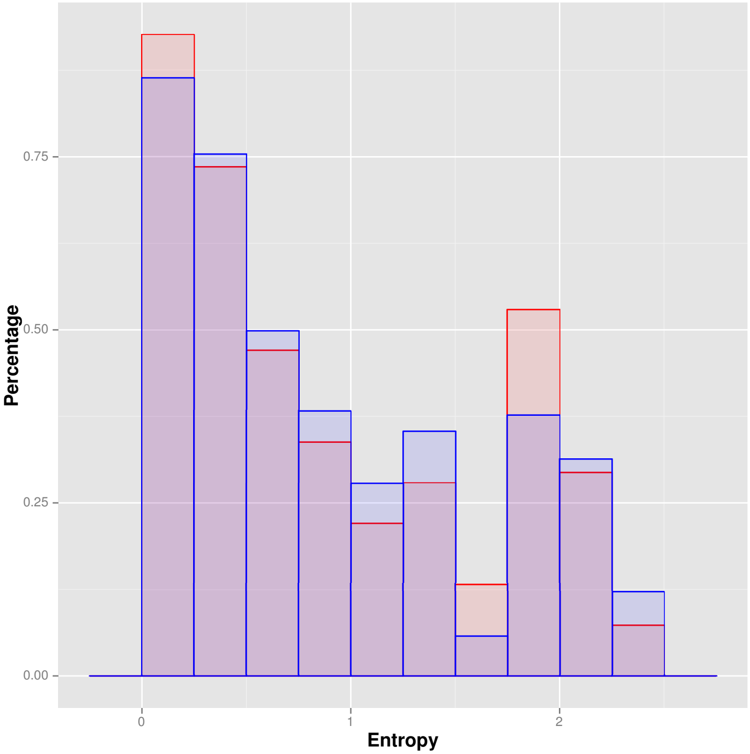


Figure 2.14: Shannon entropy of tissue specificity - Histogram of the percentage of the Shannon entropy of the tissue specificity for the canonical miRNAs (red) and all the other variants (blue).

Most categories of isomiRs (such as the 3' *shorter* or *longer*) were detected with a number of occurrences comparable to those of canonical miRNA. Moreover their presence in specific miRNAs and in several tissues argues that these molecules should not be artifacts of sequencing. As we expected, the most frequently observed variants, in terms of both the number of miRNAs displaying them and their overall abundance, are at the 3' end. These variants may play a role in miRNA degradation and be associated in miRNA stability and efficiency (132; 135).
Although generally rare, 5' and polymorphic variants represent a significant proportion of the population of some miRNAs. These findings indicate and confirm that most isomiRs do not possess different targeting specificities but few isomiRs (with an addition or deletion of 5' nucleotides) may instead have dramatic effects on target selection (78; 129; 140; 143). Some categories of isomiRs seem to be less relevant than others and it is possible to argue that may more likely be artefacts of the library preparation and/or sequencing. Even though this consideration cannot be ruled out, the step of filtering introduced to remove the less abundant variants strengthens our thought that these variants can be real isomiRs.

Most miRNAs do not exhibit a high frequency of mutations and therefore the relevance of these variants is likely to be limited to a small miRNA subset. Comparing the distributions of the mutations across the different cell types we also observe a consistency among the biological replicates of a given sample and an evident differences across the samples (Fig. 2.15). I also observed the mutations that fall within the seed regions and therefore may affect the target selection (Fig. 2.16). Further investigations will be needed to identify which targets are affected by these mutations.

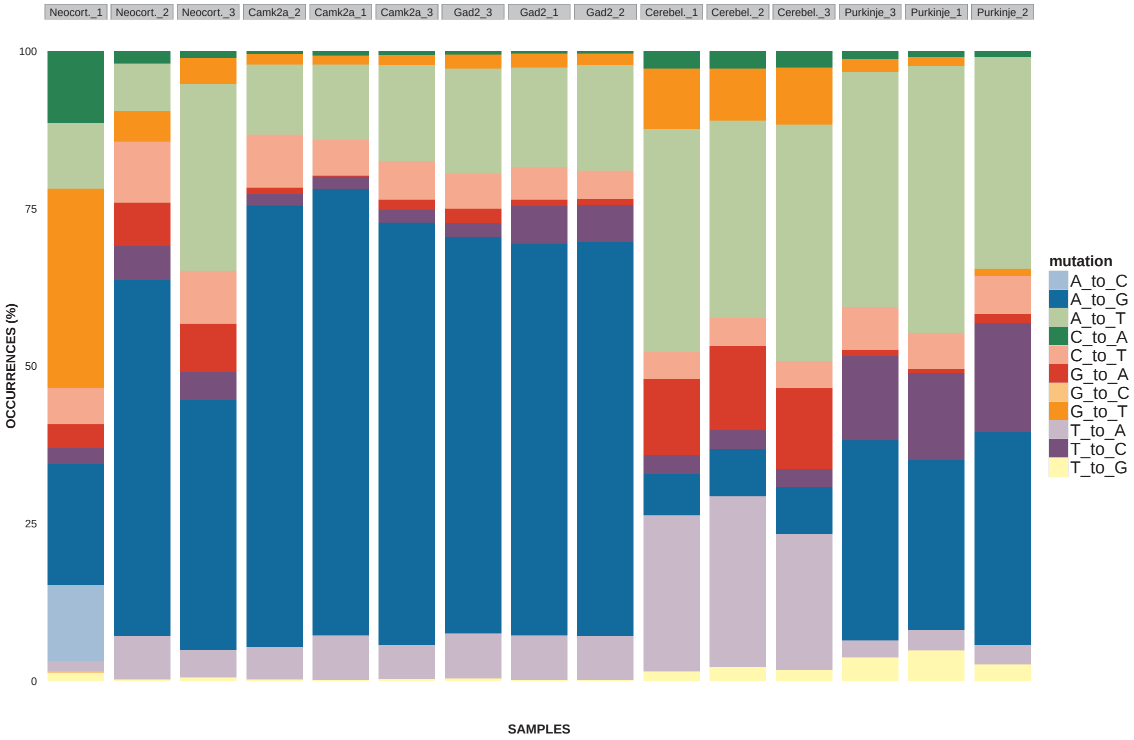


Figure 2.15: Mutation profiles in different tissues - Comparison of the mutation profiles among the different cell types .

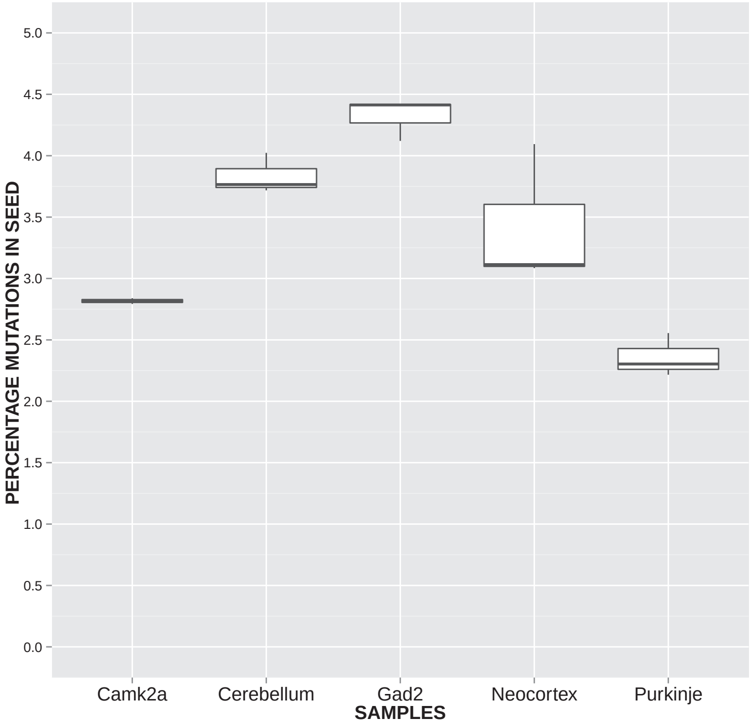


Figure 2.16: Percentage of mutations in seed - Percentage of mutations in the seed region for each cell type.

I also studied in detail the behaviour of the individual miRNAs and their relative main variants (Fig. 2.17). Many main variants are expressed according to the cell type. For example, in the Purkinje cells (class of neurons located in the cerebellar cortex) the main variant for mmu-miR124-5p is canonical while for the other cells the main variant is shorter than the canonical ones. A similar situation is observed for mmu-miR-9-5p. Again, these findings show that in some cases the canonical sequence does not represent the most expressed variant for a given miRNA. It is noteworthy that most isomiRs are not subject to dramatic tissue-dependent variations, suggesting that, although the existence of isomiRs is a widespread feature of miRNA biology, the differential regulation of activity through selective isomiR production is likely to be limited.

Further experimental work is required to demonstrate these supposed functionalities, because most work to date has largely been limited to the bioinformatics identification of these variants. Although further studies are required, we believe that this approach may be a starting point to indicate relations between the length and/or sequence variation of miRNAs and their stability and functionality, which may result in biologically significant outcomes.

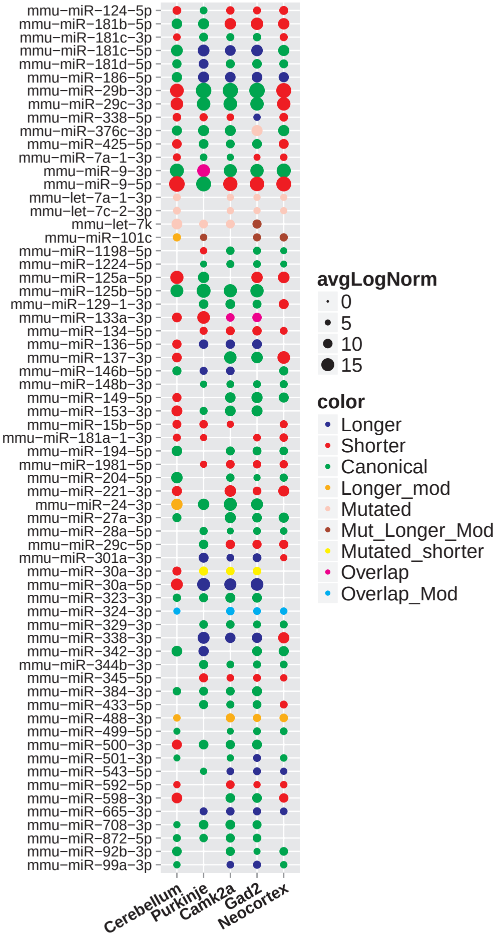


Figure 2.17: Occurrences of the main variants - In this plot the miRNAs with the related main variants (colour) and their related occurrences (size of the circles) are shown. Only the more expressed (up the median) miRNAs are plotted.

2.3 ChIP-Sequencing

2.3.1 Background

The article of Venter *et al.* published in 2001 (144) starts with the assertion that *"the decoding of the DNA that constitutes the human genome has been widely anticipated for the contribution it will make toward understanding human evolution, the causation of disease, and the interplay between the environment and heredity in defining the human condition"*. However, it has become increasingly clear that the control of the genome itself is managed through much more complex interactions, involving modifications to the DNA and its associated proteins, and that protein-DNA interactions have huge impacts on the phenotype of an individual (145; 146; 147; 148). These modifications, joined with the activity of the transcription factors (TF) that selectively interact with specific promoters, repressors and enhancers, have the ability to regulate the gene expression turning genes on and off in ways that would be impossible to predict from the genome sequence alone (149). Measurements of protein-DNA interactions by chromatin immunoprecipitation (ChIP) that first used microarrays, are now being studied by deep DNA sequencing versions (ChIP-Seq) that offer distinct advantages in increased specificity, sensitivity and genome-wide comprehensiveness (150). Indeed, the ability to directly map genome-wide protein-DNA interactions *in vivo* offers an essential complement for a deeper understanding of the cell dynamics (151). Such genome-wide measurements have, for example, allowed to generate new models of nucleosome positioning (152; 153), unveil potential functions for histone modifications (154; 155; 156), quantify the evolution and variability of transcription-factor binding sites (157; 158) and reveal unexpected regulatory relations (159; 160).

2.3.2 Methods

Chromatin immunoprecipitation is a process with the ability to collect fragments of DNA known to be bound to a specific protein of interest from living cells (161). For ChIP-Seq, these fragments are selected according to length for sequencing and for specific problem under study, usually between 100 and 600 nt. After an amplification step (either before or after size selection) the fragments are then directly sequenced without the need for cloning steps. Experimental details about the chromatin immunoprecipitation followed by deep sequencing can be found in several works (17; 150; 154; 162).

Once the reads have been obtained, they are aligned to the reference genome by using the aligner Bowtie (163). Zero or one mismatch are allowed due to possible single nucleotide polymorphisms or sequencing errors and the uniquely mapping reads are retained for further analysis. The aligned reads are then used to find regions that are enriched in the ChIP sample respect the input DNA to identify 'peaks' of the specific protein-DNA binding. There are many different, open source algorithms that can be used for peak calling (69; 148) and due to the specific problem under study I decided to use MACS (164; 165; 166) and PeakSeq (167). The general idea for a peak calling program (see an overview in (69) or (168)) is to use the location and directionality of the reads to identify the enriched regions. By using the reads, the entire fragments of DNA can be computationally inferred, possibly considering the average size of the DNA fragments that were isolated during the library construction. At each nucleotide in the genome, the number of overlapping fragments is counted to obtain a peak height (169). To capture local biases in the genome MACS and PeakSeq use a dynamic Poisson distribution to model a local background. Furthermore PeakSeq takes into account also the variability in genomic mappability. Candidate peaks with p-values below a user-defined threshold are then identified and also FDR is estimated dividing the number of control peaks over the ChIP sample by the number of ChIP peaks over the control sample (170). For our purposes, I used the strategies and pipelines that subsequently were used, among others, as guidelines in ENCODE project (171).
As for the RNA-Seq, I have implemented and automated these pipelines as workflows in order to analyse different data with the possibility of modifying the parameters and to automatically compare the results with those of previous analysis performed with different parameters and/or different versions of the tools.
A first impression about the quality of a ChIP-Seq experiment can be obtained by local inspection of the mapped reads using a genome browser such as UCSC (172). Although not quantitative, this approach is very useful both for bioinformatics and experimentalists because the shape of the peak and the strength of the signal relative to the control sample can provide a sense of ChIP quality (171). A true signal

is expected to show a clear asymmetrical distribution of reads mapping to the forward and reverse strands around the midpoint (peak) of accumulated reads (168). In our analysis, the peak finding was integrated with several downstream analyses that, associating the peaks with functionally relevant genomic regions (such as gene promoters, transcription start sites or intergenic regions), aimed at annotating and characterizing the extracted enriched regions. Usually the correlation of the peaks to the genes in their neighbourhoods includes the identification of the closest gene to each peak and its relative position: upstream of transcription start site (TSS), in the intron, exon, 5'/3'-UTR, or downstream of transcription termination site (TTS). Consequently when there are multiple nearby genes for a peak, the peak is associated with the gene whose TSS is the closest (173; 174; 175). By using this solution it is possible to lose some genes therefore I implemented a new method, named NED, with which a researcher, choosing the size of the regions before and after the TSS of the genes, identifies all the peaks that fall in the specified regions (Fig. 2.18). By using one or more annotations, the researcher can also detect all the peaks that fall in a particular genomic regions such as of opened chromatin, or on which other ChIP experiments were performed. In this way the peaks were classified according to both their genomic background and chromatin state. For the first analysis the annotation was obtained from public databases like UCSC (172), GenCode (176), RefSeq (177) and MirBase (133; 134); for the second analysis the chromatin state segmentation results from the ENCODE project in which a common set of states were learned by integrating ChIP-Seq data for nine factors using a Hidden Markov Model (178).
A useful way of checking whether a ChIP-Seq experiment was successful is to compare the peak list with those obtained by other scientists in same cell types: even though the overlap will not be perfect, a very poor overlap will suggest that the experiment might not have worked. Therefore, integrating NED with the experimental available data, I compared the list of peaks with which obtained by ChIP-Seq experiments on same cell lines derived from the ENCODE repository.

Considering different parameters or tool versions, the analysed experiments have generated a large number of ChIP-Seq datasets that have produced a number of hundreds of thousands significant peaks. It can be a challenging and time-

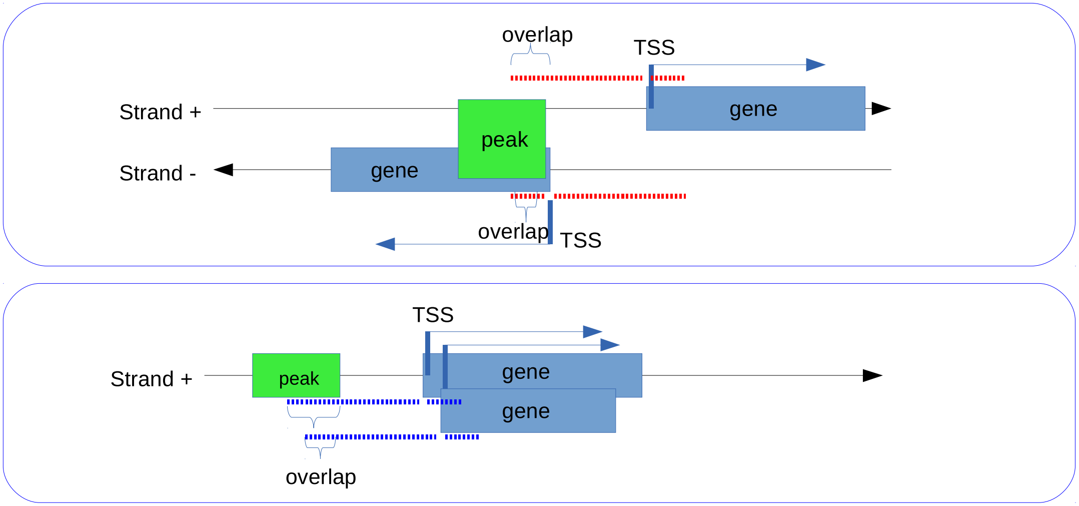


Figure 2.18: Underlying idea of NED - By choosing the size of the regions before and after the TSS of the genes, a researcher can identify all the peaks that fall in specified regions

consuming task to examine all peaks and to develop meaningful biological interpretations of their functional relevance. Motivated by this challenge, I developed a database to store all the results and take trace of the specific analysis with which I obtained them (Fig, 2.19). In this way I can perform complex queries to extract information for a specific experiment or to compare results derived by analyses performed with different parameters. In order to facilitate data comparison, interpretation and hypothesis generation for the experimentalists I also created personalized web-pages in which, for each experiment, the researcher can, for example, visualize on a genome browser a specific peak.

Motif analysis is useful for much more than just identifying the causal DNA-binding motif in ChIP-Seq peaks. For example, when the motif of the ChIPed protein is already known, motif analysis provides validation of the success of the experiment (179). For the motif analysis I assembled a set of genomic sequences corresponding to the significant ChIP-Seq peaks and then performed the motif discovery using MEME-tools (180; 181; 182) and RSATpeak-motifs (183; 184). I compared the discovered motifs with known DNA motifs using motif comparison software (185; 186). This analysis is useful to confirm the presence of the ChIPed TF motif if its (or its TF-family) binding motif is known.

Finally a functional interpretation of the genes associated with the peaks was initially done by using DAVID (40) and GREAT (175) and then the home-made tool FIDEA.

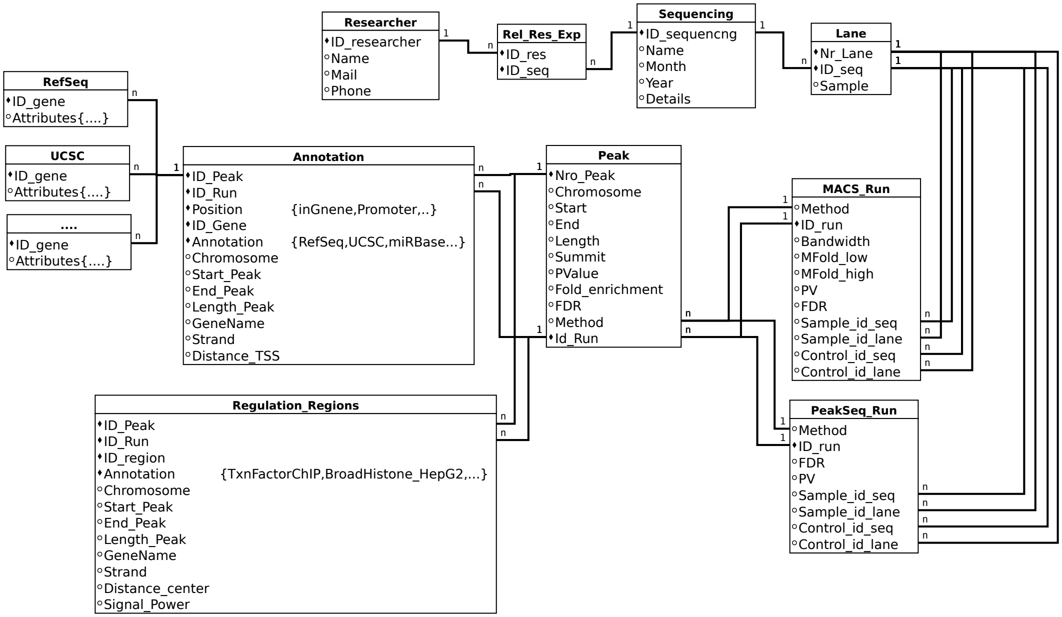


Figure 2.19: Scheme of the database used to store ChIP-Seq data - Scheme of the database used to store all the results about ChIP-Seq experiments and take trace of the specific analysis with which a researcher obtained them.

2.3.3 Application: identification of target genes for the Hepatitis B Viruses X protein

The project in collaboration with Prof. Levrero group focuses on HBV X protein (hepatitis B virus X protein (HBx)). HBx is an essential factor for transcription/viral replication and it has been considered to be one of the most important causes of HBV-induced hepatocarcinogenesis (187; 188; 189). For this reason HBx might affect viral functions, as well as host cell functions, by modulating a wide variety of cellular processes, including transcription, cell cycle progression, DNA damage repair, and apoptosis (66). Aim of this study is to use a broad chromatin immunoprecipitation approach (ChIP and ChIP-Seq) to identify target genes of HBx and to validate and to refine existing hepatocellular carcinoma molecular signatures. High-throughput sequencing of anti-HBx ChIP-enriched DNA fragments (ChIP-Seq) was performed on wild type, mock and HBx-mt monomeric linear full length HBV DNA HepG2 transfected cells. The ChIP-Seq experiments were performed on an Illumina GAIIx.

The data analysis on the ChIP-Seq experiments performed on HBx protein is a very complex problem: HBx is a small protein (154 aminoacids) and a protein complex is required for its interaction with the DNA. Therefore this situation is not an optimal system for the identification of binding sites due to the fragility of the protein complex and the multiple interactions that the protein might have.

The study was based on 2 different experiments of ChIP-Seq, each one composed by 2 samples. One experiment was performed at laboratory of Life-Nanoscience of the Sapienza University of Rome (defined below SAP, (Experiment 4 in Appendix 6.1)) and the other was performed at Istituti Fisioterapici Ospitalieri (IFO) (defined below IFO, (Experiment 5 in Appendix 6.1)).

ChIPSeq analysis of HBx chromatin recruitment revealed a specific binding to a large number of new and known target sequences. In the 4 independent ChIP-Seq samples ~16000 potential HBx binding sites were identified, 12.8% located within 10 kb upstream of a transcription start site (defined Promoter in Fig. 2.20). Several of these peaks were validated by quantitative PCR (qPCR) by the experimental group. Moreover, by observing the chromatin states, 3.3% of the peaks falls in the regions classified as Promoter and 8.1% in regions classified as Enhancer (Fig. 2.21).

Despite the high complexity of the system we detected 6212 distinct genes (considering a range of 10000 nt up and 1000 nt down the TSS), 20% of which are in common between the two experiments. Integrative analysis of these genes (which are potentially regulated by HBx) shows an enrichment in genes involved in cell metabolism, chromatin dynamics and cancer as well as in categories involved in HBV replication (for example Ras family [1], calcium transport, endocytosis, mitogen-activated protein kinases and $Wnt\beta - catenin$ signaling pathways (MAPK\WNT), proto-oncogene tyrosine-protein kinase Src, hepatocyte growth factor (HGF) and epidermal growth factor (EGF) families). Moreover, 173 miRNAs are placed in proximity of an enriched regions. For 75 of these miRNAs (Fig. 2.22), the peaks represent putative promoters since fall in a maximum distance of 5000 nt up the miRNA. Among these miRNAs there are some of particular interest, such as mir224 or mir21, that are shown to be involved in hepatocellular carcinoma (191; 192; 193; 194; 195) and, so far, 16 miRNAs have already been experimentally validated by real-time RT-PCR. The other miRNAs are potentially mirtrons since they are located in the introns of the mRNA encoding host genes (196).

[1] Discoveries made in the late 1970s and the early 1980s revealed that the transforming activities of the rat-derived Harvey and Kirsten murine sarcoma retroviruses contribute to cancer pathogenesis through a common set of genes, termed ras (for rat sarcoma virus) (190)

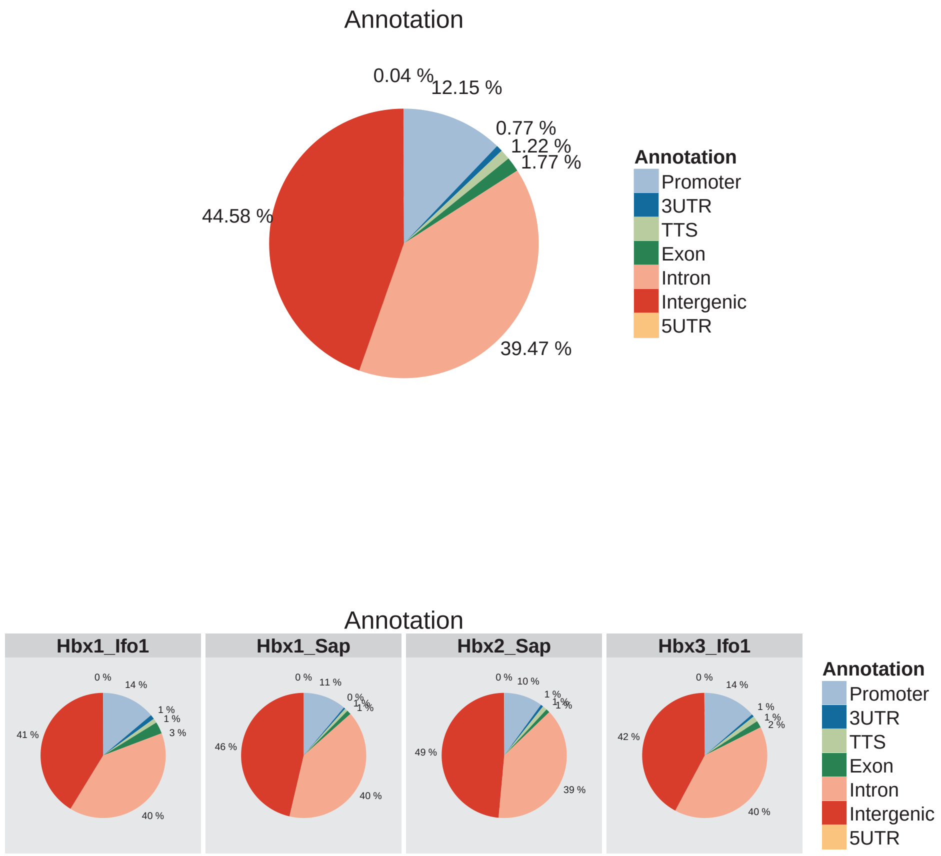


Figure 2.20: Classification of the peaks considering genomic features of interest - Distribution of peaks on genomic features of interest according to UCSC annotation. All the called peaks are considered together (top) as well as separately in line with the experiment and the replicate (bottom).

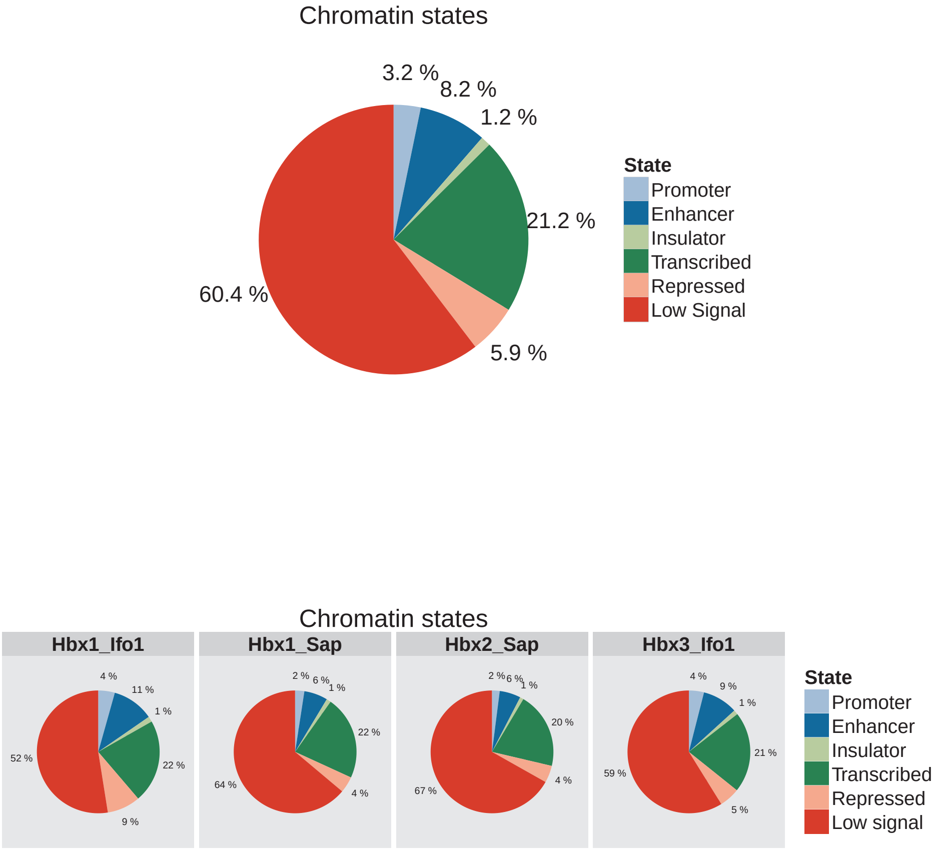
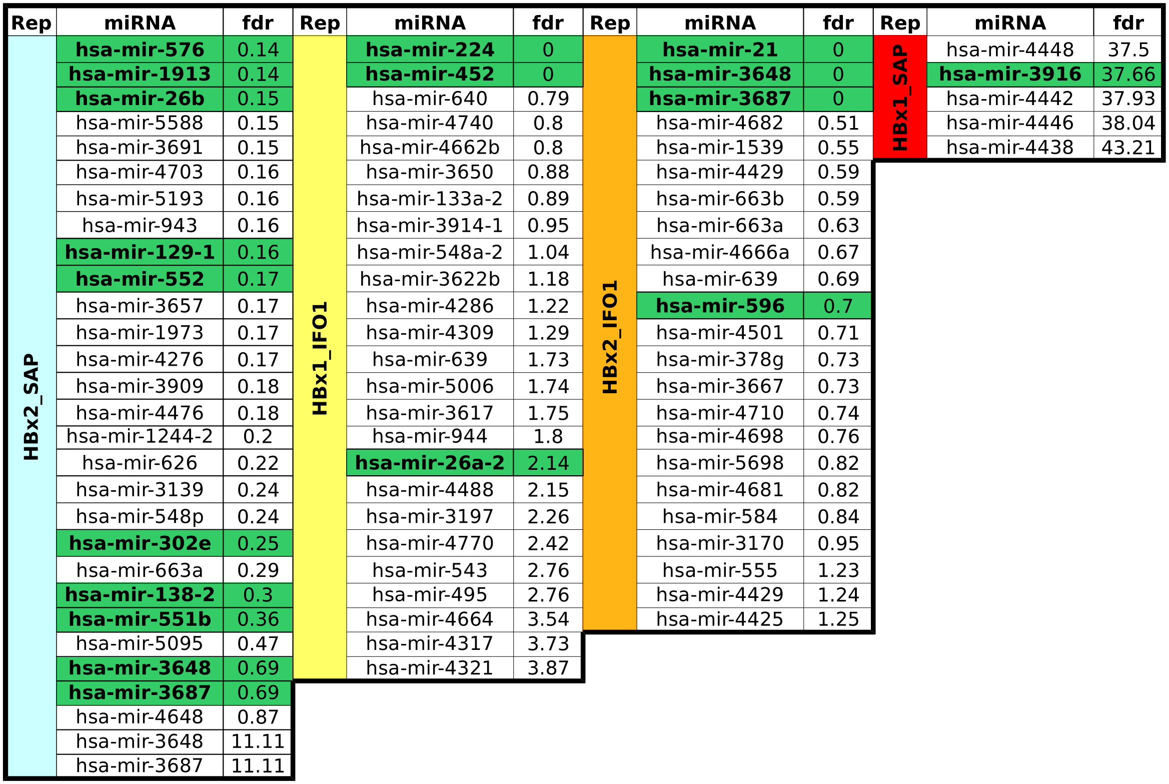


Figure 2.21: Classification of the peaks considering the chromatin states segmentation - Distribution of peaks on genomic features according to the chromatin state segmentation resulted from the ENCODE project. All the called peaks are considered together (top) as well as separately in line with the experiment and the replicate (bottom).



Rep	miRNA	fdr	Rep	miRNA	fdr	Rep	miRNA	fdr	Rep	miRNA	fdr
HBx2_SAP	**hsa-mir-576**	0.14	HBx1_IFO1	**hsa-mir-224**	0	HBx2_IFO1	**hsa-mir-21**	0	HBx1_SAP	hsa-mir-4448	37.5
	hsa-mir-1913	0.14		**hsa-mir-452**	0		**hsa-mir-3648**	0		**hsa-mir-3916**	37.66
	hsa-mir-26b	0.15		hsa-mir-640	0.79		**hsa-mir-3687**	0		hsa-mir-4442	37.93
	hsa-mir-5588	0.15		hsa-mir-4740	0.8		hsa-mir-4682	0.51		hsa-mir-4446	38.04
	hsa-mir-3691	0.15		hsa-mir-4662b	0.8		hsa-mir-1539	0.55		hsa-mir-4438	43.21
	hsa-mir-4703	0.16		hsa-mir-3650	0.88		hsa-mir-4429	0.59			
	hsa-mir-5193	0.16		hsa-mir-133a-2	0.89		hsa-mir-663b	0.59			
	hsa-mir-943	0.16		hsa-mir-3914-1	0.95		hsa-mir-663a	0.63			
	hsa-mir-129-1	0.16		hsa-mir-548a-2	1.04		hsa-mir-4666a	0.67			
	hsa-mir-552	0.17		hsa-mir-3622b	1.18		hsa-mir-639	0.69			
	hsa-mir-3657	0.17		hsa-mir-4286	1.22		**hsa-mir-596**	0.7			
	hsa-mir-1973	0.17		hsa-mir-4309	1.29		hsa-mir-4501	0.71			
	hsa-mir-4276	0.17		hsa-mir-639	1.73		hsa-mir-378g	0.73			
	hsa-mir-3909	0.18		hsa-mir-5006	1.74		hsa-mir-3667	0.73			
	hsa-mir-4476	0.18		hsa-mir-3617	1.75		hsa-mir-4710	0.74			
	hsa-mir-1244-2	0.2		hsa-mir-944	1.8		hsa-mir-4698	0.76			
	hsa-mir-626	0.22		**hsa-mir-26a-2**	2.14		hsa-mir-5698	0.82			
	hsa-mir-3139	0.24		hsa-mir-4488	2.15		hsa-mir-4681	0.82			
	hsa-mir-548p	0.24		hsa-mir-3197	2.26		hsa-mir-584	0.84			
	hsa-mir-302e	0.25		hsa-mir-4770	2.42		hsa-mir-3170	0.95			
	hsa-mir-663a	0.29		hsa-mir-543	2.76		hsa-mir-555	1.23			
	hsa-mir-138-2	0.3		hsa-mir-495	2.76		hsa-mir-4429	1.24			
	hsa-mir-551b	0.36		hsa-mir-4664	3.54		hsa-mir-4425	1.25			
	hsa-mir-5095	0.47		hsa-mir-4317	3.73						
	hsa-mir-3648	0.69		hsa-mir-4321	3.87						
	hsa-mir-3687	0.69									
	hsa-mir-4648	0.87									
	hsa-mir-3648	11.11									
	hsa-mir-3687	11.11									

Figure 2.22: MicroRNAs possibly associated to the HBx protein - MiRNAs for which the peaks represent putative promoters since fall in a maximum distance of 5000 nt up the miRNA. The microRNAs that have been validated by experimental group are highlighted in green.

These results drive to some conclusions that need to be tested and validated by experimentalists. The principal idea is that HBx can support as well as repress the expression of miRNAs that affect HBV replication defining new regulatory loops (for example in association with miR224, miR552 or miR3648) or regulating cellular functions (for example in association with miR21, miR26b, miR-502). Multiple transcription factors are involved in these circuits since they may mediate HBx binding to its target sequences (for example NFkB, E2F1, $\beta - catenin$).

2.4 Deep sequencing of phage display libraries to support antibody discovery

2.4.1 Background

Measuring the total concentration of antigen-specific serum antibodies is a fundamental step in the diagnosis of infectious and autoimmune diseases since it is used to monitor the efficacy of vaccination, which is the most powerful tool to preserve human health and to reduce the costs of medical care. However, a purely quantitative analysis of serum antibodies is a poor indicator of the complexity of the antibody response, which involves the activation of thousands of different B cell clones and the secretion of a wide variety of antibodies, each one directed against a different region of the immunizing antigen(s) (197). For reasons that are only partially understood, the antibodies induced by any immunizing antigen are not equally directed against the various portions of the antigen molecule (198). Often, within an antigen, there are regions that are strongly reactive with antibodies (i.e. immunodominant regions) flanked by domains that seem to be partially or completely ignored by the immune system. Anti-microbial vaccination induces the production of a great variety of antigen-specific antibodies, only a minority of which possesses the ability to protect against target infections (198; 199). In other words, only certain antibodies (those directed against specific "hot spots" of the antigen molecule) have immunoprotective activities. Therefore, pathogens adopt sometimes the strategy of incorporating, in the context of their virulence factors, immunodominant regions that work properly as "decoys" by preventing the immune system from targeting the "hot spots" (199). In such a scenario, selective removal of the immunodominant regions can boost the immunoprotective properties of the antigen (200). In view of these considerations it would be helpful, particularly in the course of preclinical studies and clinical trials involving vaccines, to establish whether the immune response is optimally targeted against the antigenic residues, critical for immune-mediated protection. To this end, a method capable of providing a detailed analysis of the fine specificity of vaccine-induced antibody repertoires would be useful to guide rational antigen design and selection of appropriate adjuvants. Indeed, the ability of certain adjuvants to broaden the antibody repertoire and to provide extended coverage is becoming increasingly clear (201; 202). Moreover, because the spectrum of antibody specificities varies with age and physiology, repertoire profiles may be useful to specifically tailor vaccine formulations for different age groups and for high-risk populations (203; 204; 205).

The recent development of high-throughput methods for repertoire data collection (from single cell mass spectroscopy and multicolor flow cytometry to massively parallel sequencing of immunoglobulin transcripts) offers today an opportunity to analyse large samples of lymphocyte repertoires (206; 207; 208). Although these methods provide extensive information regarding the diversity of clonotypes and immunoglobulin gene usage, they have limited usefulness, by their nature, in sampling the antibody repertoire in terms of epitope specificity. Libraries consisting of phage particles or cells expressing on their surface peptides of various lengths have also been used in epitope mapping (209; 210; 211; 212). These techniques are, however, labor-intensive, time consuming and can identify only a limited number of epitopes. The group of prof. Teti of the University of Messina in collaboration with the group of prof. Felici of the University of Molise introduced a novel approach, based on the combined use of phage-displayed antigen-specific libraries and massive parallel sequencing of the entire population of affinity-selected phages. This approach (named PROFILER, standing for Phage-based Representation OF Immuno Ligand Epitope Repertoire) generates in a short time a high-resolution profile of antigen-specific antibody repertoires.

2.4.2 Phagotto: a tool for analysing deep sequencing data of phage-displayed libraries of peptides

The experimental groups chose NadA (213), one of the 4 components of the 4CMenB anti-meningococcal vaccine (Bexsero), as the model antigen. They generated a lambda phage displayed library in which individual phage particles display on their surface NadA fragments of various lengths. Finally, after a selection with sera from Bexsero-immunized volunteers, they performed a deep sequencing by using the Illumina MiSeq platform (Experiment 8 in Appendix 6.1). In this context I am not going to describe the experimental details which are not part of my PhD work. I am focusing on the home-made tool (named Phagotto), designed and implemented in order to analyse the data derived by sequencing (Fig. 2.23).

Sequence data from the insert amplicons were processed with an ad hoc pipeline made of Perl scripts. It gives as result a list of all univocally definable (or “unique”) sequences, classified into one of the following three categories: a) *empty* i.e. sequences lacking fragments of the antigen-encoding gene (*nadA* sequences in this study); b) *natural frame*, i.e. sequences bearing fragments of the antigen-encoding gene, the products of which are predicted (because in the correct orientation and reading frame with respect to the recombinant insert sequence) to display authentic peptide fragments of the antigen on the phage surface; c) *not natural frame*, i.e. sequences bearing fragments of the antigen-encoding gene that are not expressed on the phage surface, or that express peptides corresponding to a non-authentic reading frame because of frame shifts. The paired-end reads are expected to contain in the following order: vector, adapter and insert. The pipeline checks whether these elements are present in each read in the correct order (forward in the left reads and reverse in the right ones)(see top of Fig. 2.23). Subsequently for each pair of reads, short regions, named anchors, are extracted from the insert in proximity of the adapter. The complete insert is then defined as the region of the reference gene included between the two anchors. The nucleotide sequence of the insert is translated into amino acids starting from the first ATG codon. Peptides that do not continue in frame with the phage protein sequence cannot be expressed on the phage surface. We label these as *not in downstream frame* and filter them

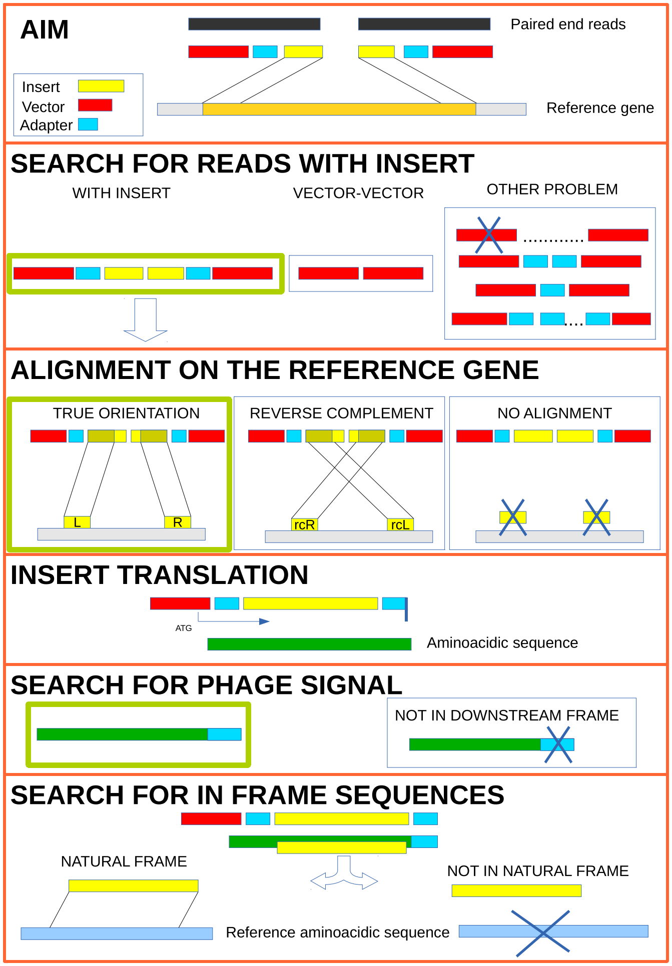


Figure 2.23: Phagotto pipeline - Pipeline of Phagotto, a tool for analysing deep sequencing data of phage-displayed libraries of peptides. The analysis is performed in 6 principal steps: (I) search for reads with insertions; (II) alignment on the reference gene in order to extract the entire insert; (III) translation of the inserts; (IV) search for phage signal in order to filter out the *not in downstream frame* inserts; (V) identification of the *natural frame* inserts; (VI) plotting.

out. The remaining sequences that can be aligned to the reference amino acid sequence of the target are classified as *natural frame*. For each of these fragments the scripts report the following attributes: a) copy number; b) amino acid sequence and length; c) start and end position of the corresponding amino acid sequence in the *nadA* protein. Anyway, after each step, the tool saves intermediate text files which can be opened and inspected to extract further informations. The results of the analysis are also given as plots in order to get an overall view of the enrichment.

In order to follow the enrichment of *natural frame* inserts during the selection process we normalized the counts of each *natural frame* insert by dividing them by the total number of sequenced reads in a given experiment and multiplying them for the mean value of sequenced reads in all the experiments. We calculated the occurrence of each amino acid of the NadA sequence by summing the counts of all inserts in the corresponding position. The "enrichment factor" for each amino acid residue of the *natural frame* sequences was calculated as the ratio between the occurrence of the residue in the affinity-selected phage population and its occurrence in the unselected library, after adding a pseudocount of 1 to the counts for each position.

2.4.3 Results

By sequencing a few hundred clones by the Sanger method, only some of the properties of a library can be inferred. Using next generation sequencing, instead, we were able to sequence thousands of clones of the *nadA*-specific library and thereby assess its quality and diversity in depth. First, we found that 4.8% percent of all sequences containing *nadA* fragments was *in natural frame* (i.e. they fulfill the requirements to be expressed on the phage surface as authentic NadA peptides). This percentage is close to the expected 1/18 (5.6%) value, calculated as the probability that a gene fragment is randomly cloned as an insert in the natural frame at the N-terminus of the lambda phage capsid protein D encoding sequence.

Natural frame fragments were evenly distributed along the entire sequence of the protein, with no major over- or under-representations of specific regions (as shown in Fig. 2.25 (first box)). Minor under-representation of short amino acid stretches at either end of the protein was expected based on library construction, which involved fragmentation of the nadA gene by DNase I digestion. To follow the process of antibody-mediated selection, two rounds of phage selection were performed on the *nadA* library using a pool of sera obtained from adults immunized with the 4C MenB Bexsero vaccine, which contains recombinant *NadA* as one of the antigens. For each selection, over 10^4 sequences were obtained and analyzed. During the selection, there was a progressive increase in the frequency of *Natural frame* sequences, as it can be appreciated from the black areas in Fig. 2.25. This indicated that phage particles displaying authentic *NadA* fragments had been selectively enriched by the *NadA*-specific antibodies present in the immune sera, while those carrying *not natural frame* or no inserts rapidly decreased in numbers. As expected in a typical phage display experiment, the copy numbers of specific polypeptides dramatically increased during subsequent rounds of selection (Fig. 2.24). To more precisely quantitate this feature, we calculated for each *Natural frame* sequence the “enrichment factor”, as the ratio between the occurrence of that sequence in the affinity-selected phage population and its occurrence in the unselected library (Fig. 2.26). By displaying cumulative enrichment factor values as a function of single amino acid positions along the *NadA* sequence (Fig. 2.27), it was possible to unambiguously identify *NadA* regions that were enriched by the antibody-mediated selection process.

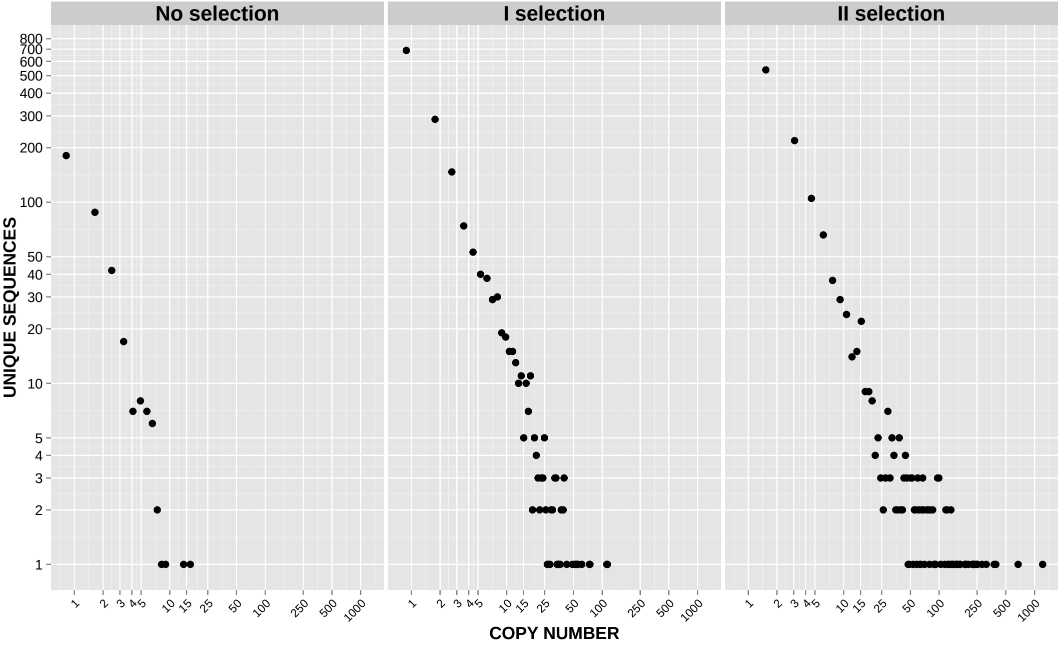


Figure 2.24: Abundance of natural frame NadA fragments - Distribution of *natural frame nadA* fragments in the unselected library and after one and two rounds of selection. Each point represents the number of unique fragments (vertical axis) for the number of copies (horizontal axis).

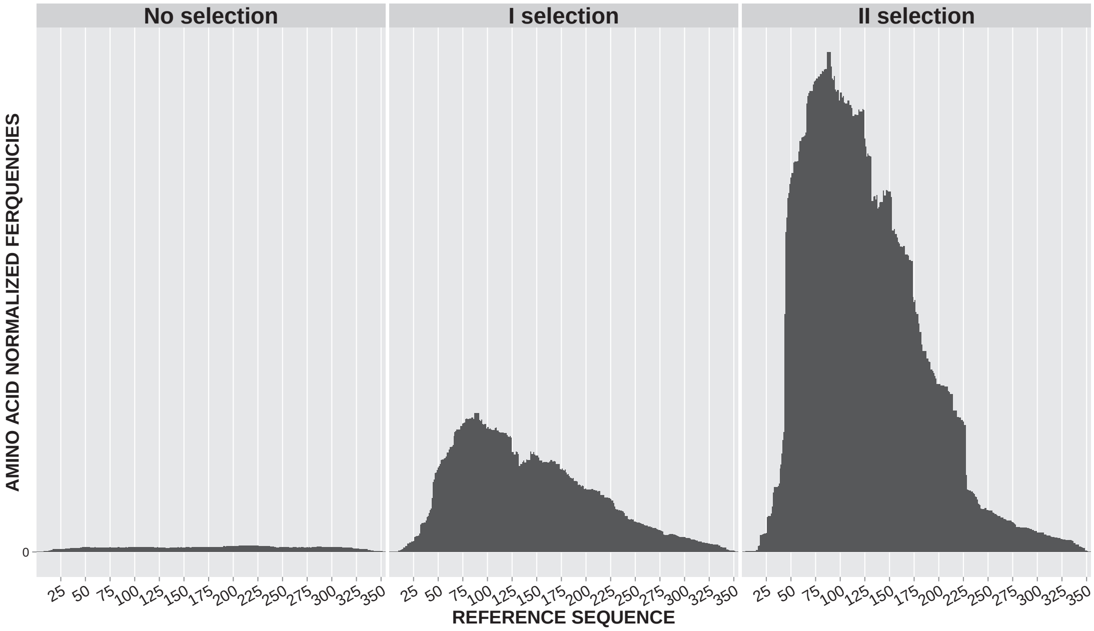


Figure 2.25: Amino acid frequencies of the natural frame NadA fragments - Cumulative occurrences, per single amino acid position, of sequences predicted to express authentic *NadA* fragments (vertical axis) in the uselected library and after one and two rounds of selection. The horizontal axis reports the amino acid positions of the translated *NadA* sequence.

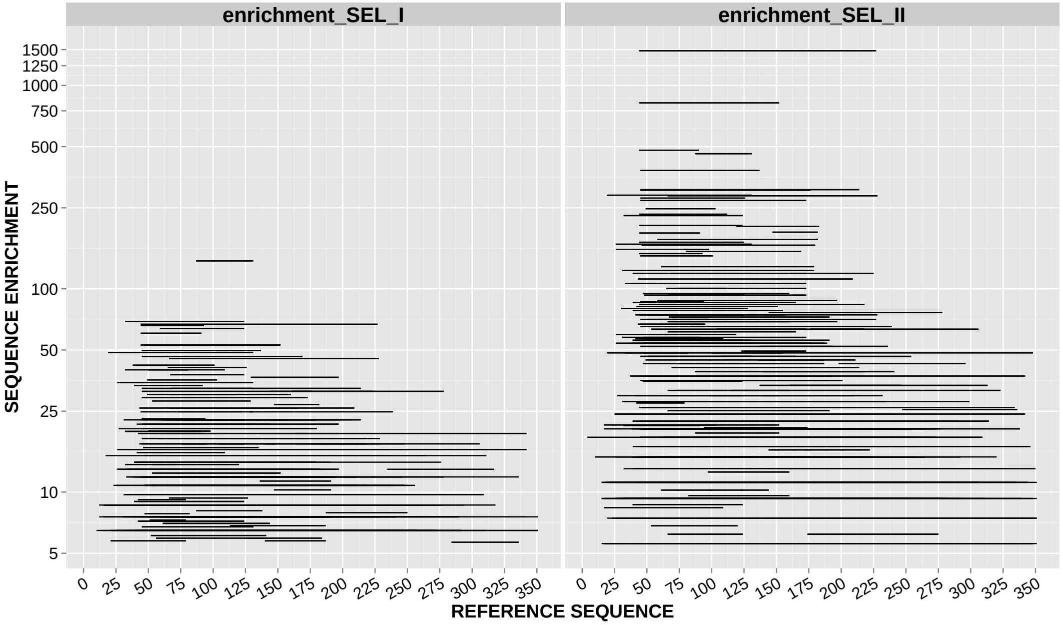


Figure 2.26: Enrichment factor of the NadA fragments - Enrichment factor of NadA fragments after one and two round of selection. Each fragment is shown as a line connecting its start and end positions. For clarity, only the enriched fragments laying in the upper quartile of enrichment factor values are shown.

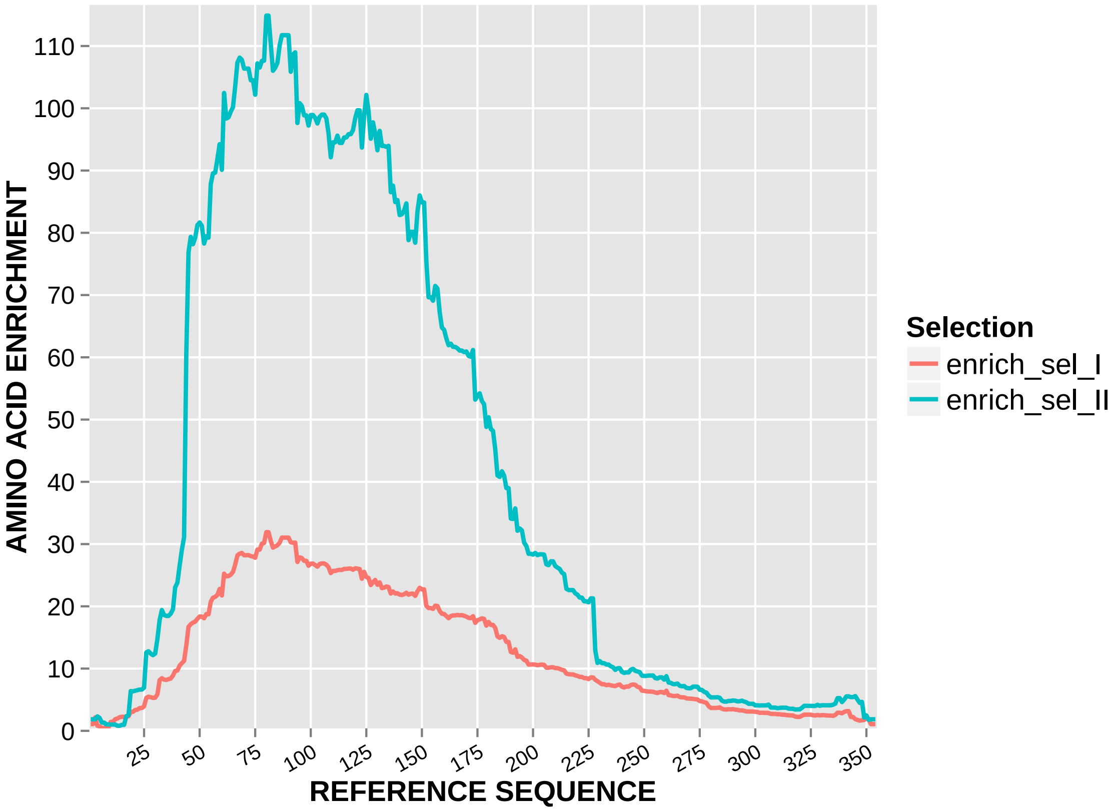


Figure 2.27: Amino acid enrichment of the natural frame NadA fragments - Cumulative enrichment factors for each amino acid position derived from NadA fragments obtained after one (red line) and two (blue line) rounds of selection.

3

FIDEA: server for the Functional Interpretation of Differential Expression Analysis

3.1 Background

Differential expression analyses typically end up with results made by hundred or thousand differentially expressed (DE) genes. Although this type of result is exhaustive and provides an essentially complete view of the analyzed transcriptomes, it is not easily readable and their functional interpretation is not always straightforward. This situation is even more evident upon the advent of high-throughput sequencing technology, that made easier the possibility of characterizing a whole transcriptome in a single experiment.

Once the differentially expressed genes have been identified and their statistical significance correctly assessed, it is essential to interpret the data in light of the biology of the specific system under study and to select the most biologically significant transcripts for further validation.

A way to make more explicative the results of a differential expression analysis is given by moving from DE genes to the process or the specific mechanism they belong (that is perform a functional analysis). The first step of the functional analysis is almost invariably an enrichment analysis (27) aimed at verifying whether a significant number of the identified genes belong to one or more specific pathways or functional categories. This is usually carried out by statistically assessing whether a pathway or process is enriched in a specific list of genes (26) and can be performed by using different classifications of the genes, for example KEGG pathways (37), Interpro (38), Gene Ontology Molecular Function, Biological Process, and Cellular Component categories (39).

The correct interpretation of the results and, especially, the identification of particularly interesting genes or functions among the differentially expressed ones, should be performed by scientists who are well acquainted with the biological problem under study since they have a wealth of knowledge about the system and can, more easily than a bioinformatician, discover less obvious and therefore more interesting relationships. However often they are not sufficiently expert to be able to effectively exploit the power of different tools and databases or to perform the comparison of the results of more than one experiment (which usually requires some scripting). This remains true even though there are a number of publicly available servers that allow functional enrichment analysis given a list of genes, the most used of which are DAVID (35; 40); g:Profiler (41; 42); Gorilla (43); High-Throughput GoMiner (44); Babelomics (45) and GeneCodis3 (46). All these tools require the user to provide lists of genes, which implies that the identification of genes the transcripts of which are up-and down-regulated needs to be performed separately and in advance. Furthermore, should the user wish to see how the results change when a different p-value or fold-change threshold is applied to identify differentially expressed transcripts. In this case the list has to be rebuilt, resubmitted to the server and the results compared. Two of the abovementioned servers, Babelomics and GeneCodis3 give the possibility of performing the analysis in parallel on two different lists of genes, for example up-regulated and down-regulated ones or deregulated in different experiments, but the burden of comparing the results is still left to the user.

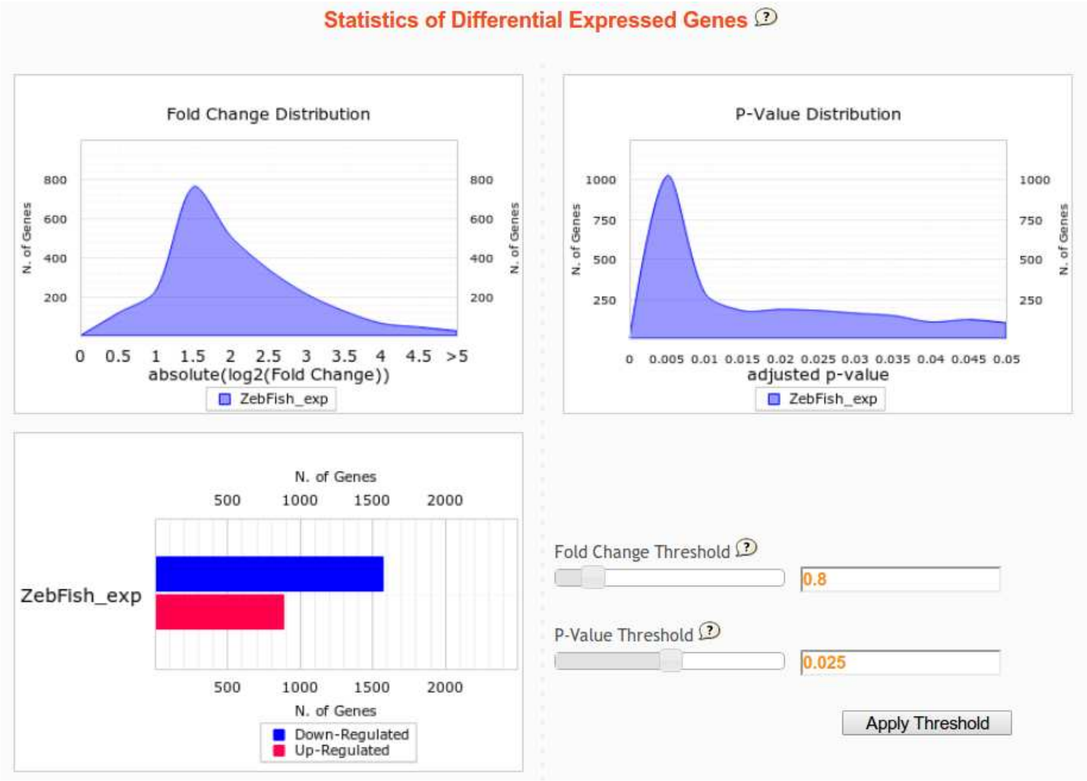
The above considerations prompted us to provide experimentalists with a more user-friendly tool for analyzing their data from a functional point of view. Therefore we designed FIDEA (67) a web server for the Functional Interpretation of Differential Expression Analysis. FIDEA was developed through a number of iterations with the experimental groups with which we collaborate and the resulting system is sufficiently easy to use and at the same time complete and flexible to be useful in the functional analysis of differential expression experiments.

The server is located at http://www.biocomputing.it/fidea, it is free and open to all users and there is no login requirement. Since FIDEA has been published (seven months ago), it has been broadly used for more than 2000 analyses, and two studies cited it (214; 215).

3.2 Description

The FIDEA server allows the user to directly input the results of a differential expression analysis, by uploading the output of Cufflinks (87; 88) or, alternatively, a formatted list of up- and down-regulated genes that can be easily obtained through other tools such as EdgeR (85) or DESeq (86). The input format in the latter case is simple and consists of the experiment ID, the gene ID, its fold change and a (corrected) P-value. The most commonly used gene IDs (Gene symbol, Entrez gene ID, Ensembl gene ID, UCSC gene ID, Refseq ID) are accepted as input. They can refer to one of the following species: *Homo sapiens, Mus musculus, Drosophila melanogaster, Danio rerio and Saccharomyces cerevisiae.*

Upon loading the input data, the system immediately shows the distribution of the fold changes in the dataset (as absolute log2 values), thereby permitting to quickly verify whether the distributions of up- and down-regulated genes are well balanced. The threshold for the minimum fold change to be considered significant can be interactively modified, leading to the direct display of the updated distributions (Fig. 3.1).



Statistics of Differential Expressed Genes

Fold Change Threshold: 0.8

P-Value Threshold: 0.025

Figure 3.1: FIDEA: statistics of Differentially Expressed Genes - The figure shows the first page of FIDEA where, upon uploading the data, the user has an overview of the distribution of fold changes for up- and down-regulated genes and can interactively modify the p-value and fold-change thresholds.

Once a fold change threshold is selected, the enrichment analysis is performed considering the up-regulated and down-regulated genes both together and separately. The statistical significance of the enrichment is computed using the hypergeometric test, the resulting P-values are corrected using the Benjamini and Yekutieli FDR method (216). The background distribution is, by default, the distribution of all the genes for the selected organism, but the user can also provide his/her own list of genes.
FIDEA analyzes both the up-regulated and down-regulated genes separately or taken together. The results of each analysis are given as:

- interactive and dynamic heat maps showing the absolute log10 of the corrected p-value;
- interactive table reporting the category name, the P-value and the corrected P-value, the fold enrichment and the number of differentially expressed genes;
- a static publication-ready heat map (Fig. 3.2) or word cloud (Fig. 3.3) reporting the enriched categories;
- a text table (csv format and downloadable).

If more than one experiment is uploaded, the user can obtain a list of genes that are in common among the experiments. These can include genes that are up-regulated in all the experiments, up-regulated in one and down-regulated in another, etc. (Fig. 3.4). The results are shown as a Venn diagram and as a list of genes that can be directly submitted to the functional enrichment analysis or downloaded.

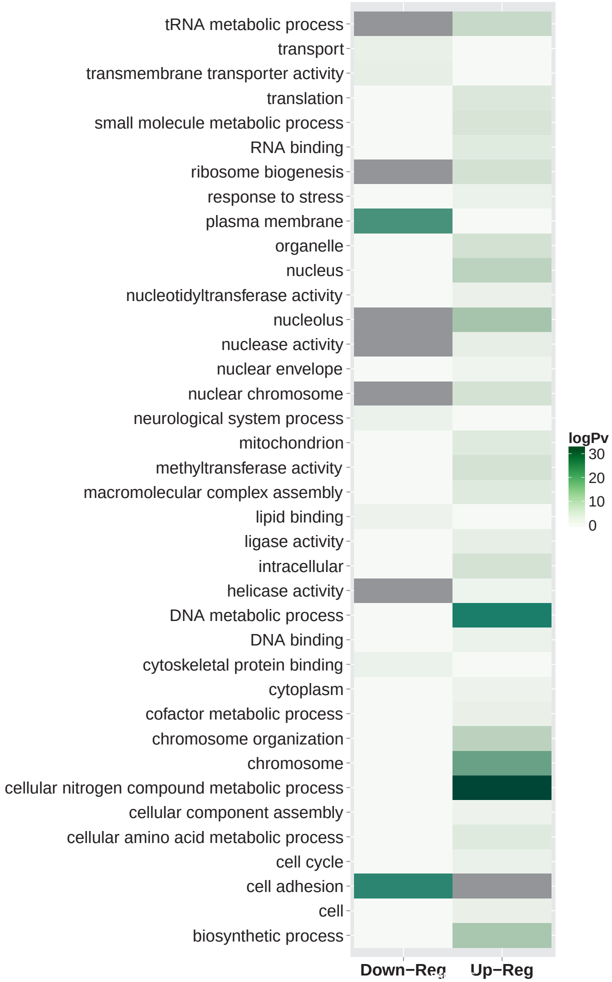


Figure 3.2: Functional analysis by using GoSlim (heat map) - The figure shows, as a heat map, the results of the GOSlim analysis considering up and down regulated genes separately. The color of the cells represents the absolute log10 of the corrected P-value.

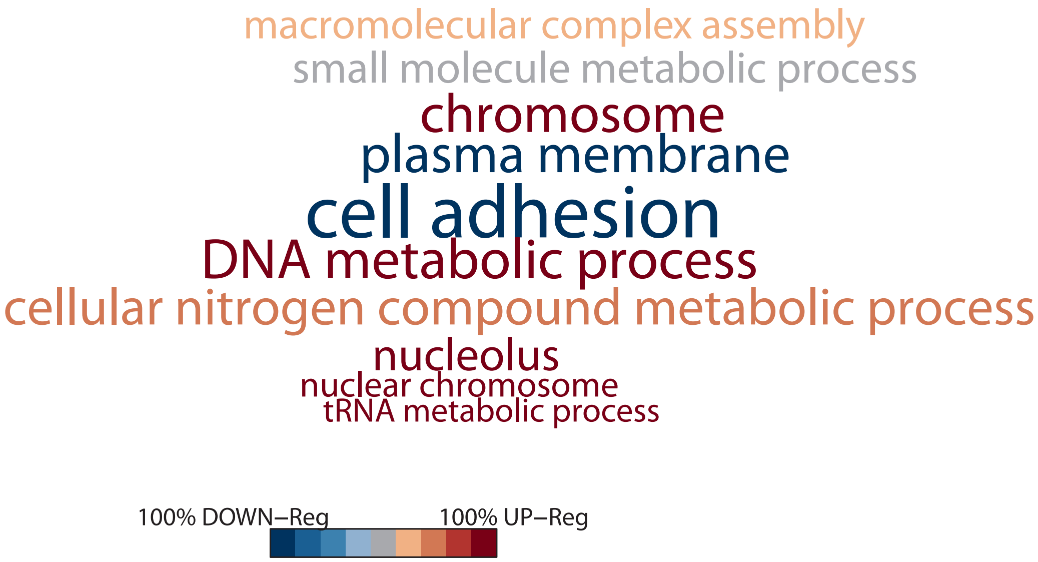
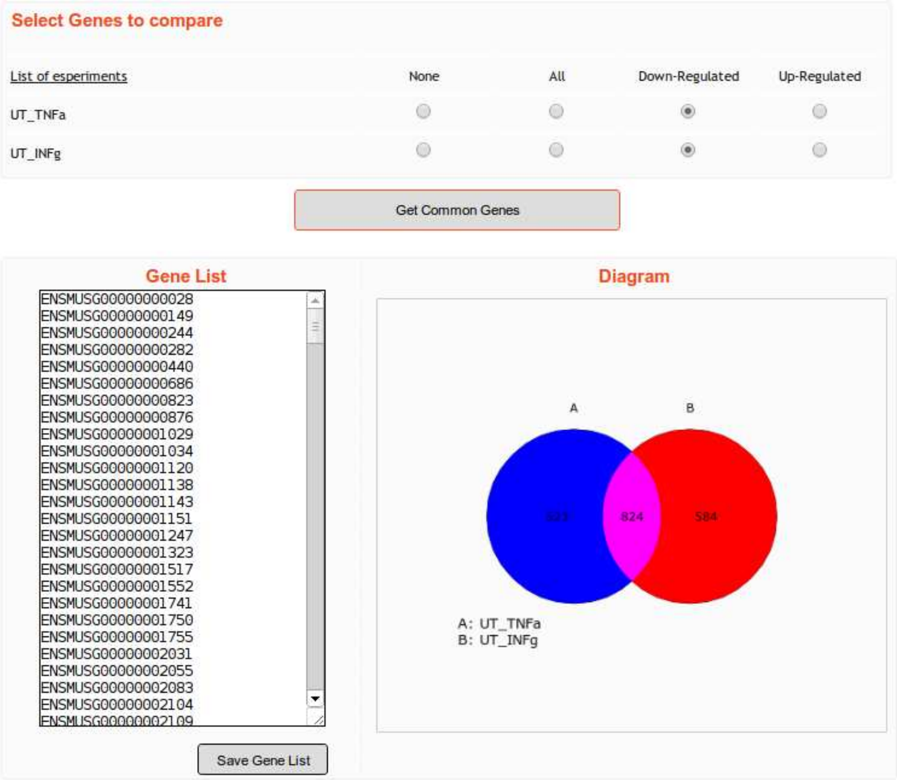


Figure 3.3: Functional analysis by using GoSlim (word cloud) - The figure shows, as a word clouds, the results of the GOSlim analysis considering up and down regulated genes taken together. The functional categories are shown with a character size related to their enrichment (according to the corrected P-value) and in different colors according to the extent by which the pathways or categories are enriched by up- or down regulated genes (red to blue, respectively).



Select Genes to compare

List of esperiments	None	All	Down-Regulated	Up-Regulated
UT_TNFa	○	○	●	○
UT_INFg	○	○	●	○

Gene List

ENSMUSG00000000028
ENSMUSG00000000149
ENSMUSG00000000244
ENSMUSG00000000282
ENSMUSG00000000440
ENSMUSG00000000686
ENSMUSG00000000823
ENSMUSG00000000876
ENSMUSG00000001029
ENSMUSG00000001034
ENSMUSG00000001120
ENSMUSG00000001138
ENSMUSG00000001143
ENSMUSG00000001151
ENSMUSG00000001247
ENSMUSG00000001323
ENSMUSG00000001517
ENSMUSG00000001552
ENSMUSG00000001741
ENSMUSG00000001750
ENSMUSG00000001755
ENSMUSG00000002031
ENSMUSG00000002055
ENSMUSG00000002083
ENSMUSG00000002104
ENSMUSG00000002109

Diagram

A: UT_TNFa
B: UT_INFg

Figure 3.4: FIDEA: intersection among lists of genes - The figure shows an example of how data from different experiments can be combined and subsequently analyzed.

3.3 Implementation

The associations between functional annotations and gene ID are stored in a local MySQL database (see Fig. 3.5 for a detailed scheme).
Due to the large number of independent databases, the identifiers are, in most cases, redundant. FIDEA maps all functional annotations to the ENSEMBL gene IDs (99) and converts all the supported gene IDs (Entrez, UCSC, Gene Symbol, Refseq) in ENSEMBL gene IDs. Specific organism IDs can also be used, namely Flybase ID, Gene and Annotation Symbol for *Drosophila melanogaster*, Gene Name, Zfin Gene ID for *Danio rerio*, and SGD Systematic Name, Primary SGD and Gene Name for *Saccharomyces cerevisiae*. Regardless of this internal conversion, the final results are given using the original gene ID provided by the user.

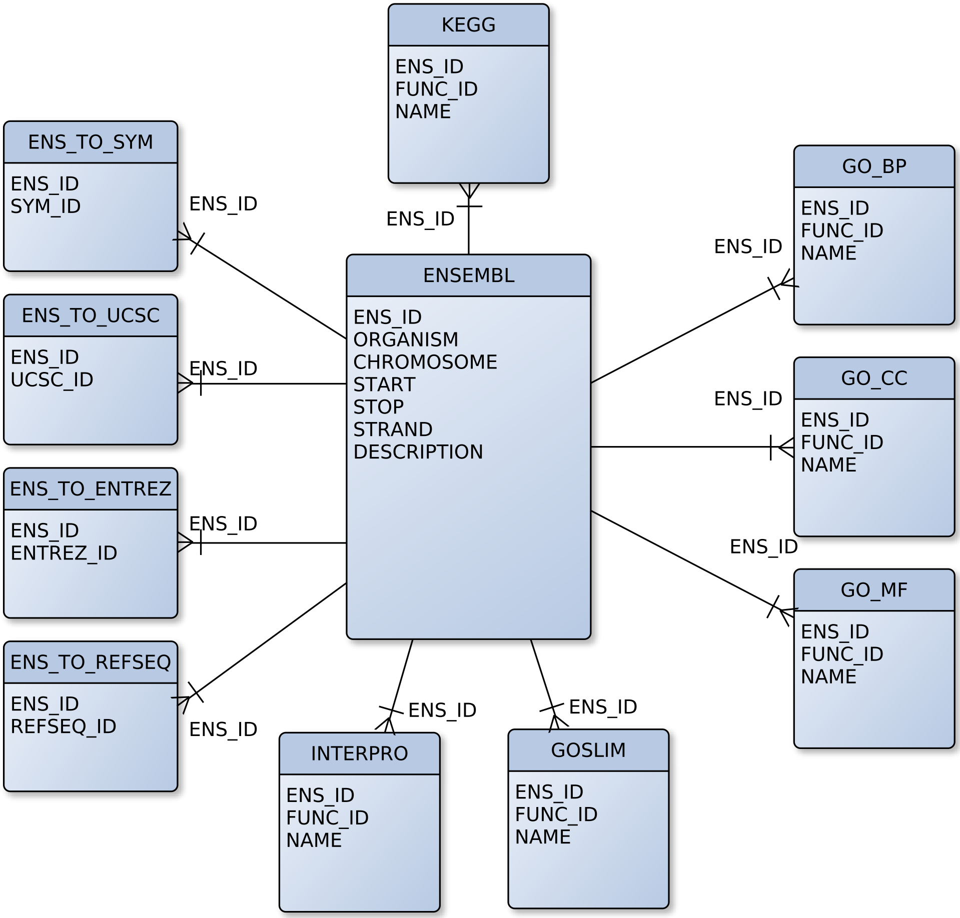


Figure 3.5: Database schema - A detailed schema of the local database integrated in FIDEA and implemented in MySQL. It stores the associations between functional annotations and gene IDs.

4

Evaluation of residue-residue contact prediction in CASP10

4.1 Background

High-throughput experiments are providing us with complete genetic blueprints for hundreds of organisms. We are now faced with assigning and understanding the functions of proteins encoded by these genomes. This task is generally facilitated by protein three-dimensional (3D) structure which, in effect, is much more informative than the amino acid sequence alone (54) and allows us to explain the biochemical mechanism by which the protein implements its functionality. Moreover if the function is unknown, a 3D structure-based similarity to other structures can reveal more about its function (62; 65). By and large, most biological functions are mediated by proteins through their 3D structures that, in turn, are mainly dictated by their respective sequences.

The protein 3D structures can be experimentally determined by X ray crystallography and NMR spectroscopy but, unfortunately, the experimental solution of all known proteins remains a challenge: existing techniques are time and resource consuming and not always successful. De facto, experimental structures are currently available for less than 1/1000 $^{\text{th}}$ of the proteins for which sequence is known and computational methods are the only viable alternative to the experimental exploration of the protein structure space. Even though hundreds of servers and tools are widely available for producing a structural model of a protein of interest by using several heuristic strategies, the protein structure prediction problem is far from being solved in general terms. For example, it is possible to derive information about a protein structure on the basis of the structure of an evolutionary related protein (comparative modeling) (217) or, when no sequence similarity between two proteins can be detected, it is possible to recognize the compatibility of the sequence of the target protein with a known fold (fold recognition) (218; 219). Alternatively, it is possible to try and assemble fragments of proteins of a known structure to reconstruct the complete structure of a target protein (220; 221).

The prediction of intra-molecular contacts in proteins can also serve as an intermediate step toward accurate prediction of the three-dimensional structure, and triggered extensive research to connect protein sequence and structure with a "two-span bridge": from sequence to contacts and from contacts to structure (68). To build such a bridge, the researchers focused on predicting contacts with accuracy sufficiently high to be useful for structure modelling on one side, and on building a structure from incomplete/inaccurate contact data, on the other.

Even though these methods for protein structure prediction have become widely available to both experimentalists and computational biologists a question remains: how good are they? How can we know which method is better suited for a specific task? In order to answer to these questions and to assess the reliability of structure prediction methods a comparison of computational models with the corresponding experimental structures is required. Effective tests and comparisons have to be performed on proteins whose structures are not yet publicly available, to avoid overestimating the reliability of protocols based on statistical observations derived from the PDB. For this purpose, since 1994 the large scale test CASP (Critical Assessment of Protein Structure Prediction) (222; 223; 224; 225; 226; 227; 228; 229; 230) has been run every two years. It aims at establishing the current state of the art in the field of protein structure prediction, measuring the progresses made and highlighting areas requiring improvements. Experimental biologists and structural genomics centres are invited to release the sequences (the CASP targets) of soon to be determined protein structures. To make all predictions blind, participating groups deposit their models before the actual target structures are made publicly available.

CASP is also testing publicly accessible servers on the same target dataset. This allows to estimate the extent to which autonomous servers can be outperformed by human predictors and meta-servers. The latter usually send the target sequence to independent servers, collect the results and compare them with each other. To make their prediction, they either select the most representative structure or assemble a hybrid model from the most frequently represented structural fragments. After model deposition ends, a huge amount of numerical data is derived from the comparison of all predictions with their actual target structures. Target domains are classified into separate categories according to their similarity to known structures and hence prediction difficulty. Independent assessors evaluate the results within each category in a critical and blind way, because predictor names are hidden until the very end. Other categories have also been established over the time for the prediction of function, domain boundaries, disordered regions and model quality.

At the end of experiment, assessors and predictors convene to discuss the results. These are also made available to the scientific community via internet and in a special issue of the journal *Proteins: Structure, Function and Bioinformatics*.

Besides showing improvements and bottlenecks in the structure prediction field, CASP has other relevant merits too. It raised the issue of objective evaluation of structure prediction methods and fostered the establishment of similar blind tests in other areas (52; 231; 232; 233).

I have had the possibility to be involved in the evaluation of residue-residue contact prediction in CASP10. In this case the participating groups were asked to submit a list of pairs of residues predicted to be in contact that hopefully can be used as restraints in constructing three dimensional models. In addiction to the measures used in previous CASPs (i.e., prediction accuracy), we decided to introduce new measures such as the rank of the first correctly and incorrectly predicted contacts and the ability to detect inter-domain contacts.

In this context I'm going to show only the methods and results related to my PhD work. The complete results can be found in the paper (68).

4.2 Methods

A pair of residues is defined to be in contact when the distance between their C_β atoms (C_α in case of GLY) is less than $8.0\mathring{A}$. Each reported contact had to be annotated with a probability score in the [0;1] range, reflecting the predictor confidence in assigning the contact. The evaluation was performed for all types of contacts with emphasis placed on long-range contacts, i.e. those involving residues separated by at least 24 residues along the sequence.

The evaluation of predictions was carried out on a per-domain basis. The domains with detectable homology to proteins of known structures were not included in the evaluation as in these cases contacts could easily be derived from the template structures. Thus, we used only the domains for which structural templates did not exist or were very difficult to identify. Therefore we assessed the performance of contact prediction methods on two sets of domains:"FM", for these domains templates did not exist or could not be reliably identified based on the target sequence; "FM + TBM_hard", an extension of the previous set obtained by adding the domains for which templates exist but are hard to identify or to properly align with the target.

Not all methods are conceptually different as often-times they rely on similar prediction techniques using similar mathematical apparatus and predictive features. To illustrate this, I clustered the methods participating in CASP10 based on the pair-wise Jaccard distance (234). In particular, the dissimilarity between two groups for each target is defined as:

$$J_{ij} = \frac{(M_{01} + M_{10})}{(M_{11} + M_{01} + M_{10})}$$

where M_{11} is the number of common contacts predicted by groups i and j, M_{01} and M_{10} are the contacts only predicted by group i or j, respectively. The J-score has values in the range of [0;1], with the value of 0 corresponding to identical predictors and 1 - to completely dissimilar ones.

I introduced also the position of the first correct prediction as well as the position of the first error for each target and each group. If the prediction contains several contacts with the same probability value, the position of the first correct/incorrect prediction is assigned regardless of whether there are incorrect/correct predictions with the same probability. In other words, if the correct prediction with the highest probability has the same probability, and therefore the same rank R, as one or more incorrect predictions, the correct prediction is assigned rank R. Analogously, the position of the first incorrect prediction is assigned regardless of whether there are correct predictions with the same probability, i.e. if the first incorrect prediction has the same rank R as a correct prediction, the first incorrect prediction is assigned rank R.

4.3 Results

In CASP10 26 groups submitted predictions of intra-molecular contacts, including 22 automated servers and four expert groups. Three groups used new methods, while others used modified techniques developed earlier and tested in previous rounds of CASP.
Here I concentrate on the results for long-range contacts on FM targets.
Fig. 4.1 shows the results of the clustering on the similarity among the groups. As one can notice, four lowest level clusters encompass two prediction groups each from the same research centres, i.e. two *Proc-S*, *Distill*, *Multicom*, and *confuzz* methods. It is apparent that the clustered groups use similar methodologies with slight modifications in the implementation of the method.

The results of the performances for each group for long-range contacts are presented in Table 4.1. The results of the PR-analysis (AUC_PR scores) and the Matthews correlation coefficient (MCC) clearly identifies the top performing group, *G489* (*Multicom*), which reaches an AUC_PR score of 9.5%. The two other groups that stand out in the PR-curve analysis are *G087* and *G072*, both from the *Distill* family of methods.

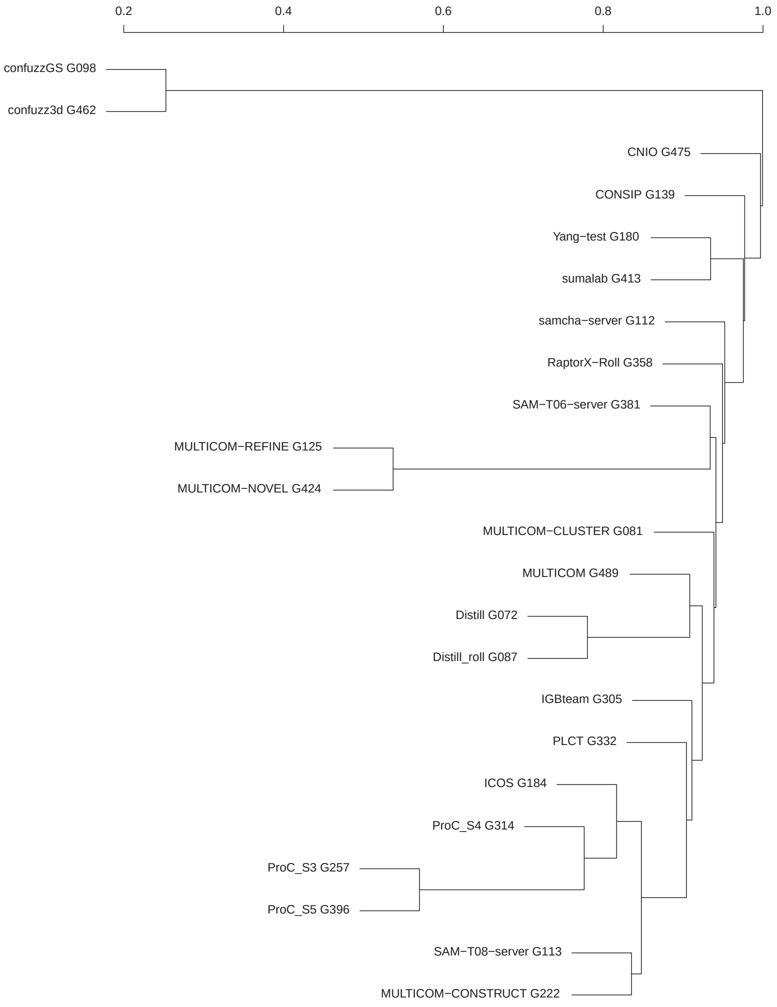


Figure 4.1: Similarity among the groups - Dendrogram illustrating the similarity among different methods as judged by the number of common predictions for all targets.

Group	No dom	MCC	precision,%	recall,%	AUC_PR
G489	15	0.131	6.9	31.4	0.095
G087	16	0.127	5	43.8	0.065
G222	16	0.12	5.7	33.7	0.043
G072	16	0.112	5.6	30.3	0.049
G396	16	0.109	4.5	37.5	0.038
G257	16	0.092	3	49.6	0.038
G113	13	0.091	4.8	23.8	0.025
G314	16	0.085	2.8	46.7	0.032
G125	16	0.083	4.6	22.2	0.034
G413	12	0.082	2.4	53.1	0.038
G424	16	0.081	4.8	20.1	0.035
G081	16	0.078	6.8	11.6	0.018
G112	13	0.076	1.7	85.1	0.027
G184	16	0.065	3.5	21.2	0.021
G139	10	0.063	1.5	67.4	0.015
G332	16	0.063	2.8	27.5	0.026
G381	13	0.061	3.3	19.1	0.019
G305	16	0.058	1.6	72.7	0.038
G358	16	0.057	7.3	5.7	0.026
G180	12	0.035	1.6	31	0.012
G434	12	0.019	6.9	0.8	0.015
G475	12	0.009	2.2	0.9	0.009
G098	12	-0.005	1.3	1	0.018
G462	12	-0.007	1.1	0.8	0.018

Table 4.1: Results of the analysis of the group performance for long-range contacts - Descriptive Statistics scores calculated for the predictions treated in the context of the complete contact maps for long-range contacts for FM domains. The results are sorted according to the MCC score.

The prediction of contacts in protein structures can be used as input for computational methods aimed at structure prediction and, in this case, the correct ranking of the contacts in terms of their probability might not be necessarily relevant. On the other hand, prediction of specific contacts in a protein might shed light on its functional or structural properties and in this case, their correctness should be experimentally tested before drawing conclusions. This is usually done by designing appropriate mutations of the residues predicted to be in contact, expressing the mutated protein(s) and testing their function (see for example Refs.(235; 236; 237; 238)). Clearly, one would like to perform as few experiments as possible.
Since contact predictions are provided together with estimates of their reliability, it is reasonable to expect that the contacts would be tested in the order they appear in the list of predictions. This raises the question of how much down the ordered list of contacts is the first correct prediction for a given method.
I computed the position of the first correct prediction as well as the position of the first error for each target and each group. Fig. 4.2(A) shows, for each group, the percentage of times in which the first correct prediction is found in a given position; Fig. 4.2(B) shows the rank of the first incorrectly predicted contact. Group *G489* that performs better than the other groups has a correct prediction in the first position on the L/5 contact lists 56% of the times and in 13% of the cases the first correct prediction is in position 2. Other groups also often have the first correct prediction ranking high in the list.
It is instructive to compare the two parts of the figure. For example, group *G184* has a correct prediction in one of the top positions about 40% of the time, but also often it has an incorrect prediction in the first positions. This is due to the fact that this group often assigns the same probability values to a set of contacts, some correct and some incorrect.

The prediction of contacts between different domains can be extremely useful in cases where multi-domain proteins are modelled using different templates for the different domains, since the step of packing together the partial models can, and often does, introduce errors. I analysed the number of cases in which different participating groups correctly predicted contacts between residues belonging to two different domains. The results for inter-domain long-range contacts in FM

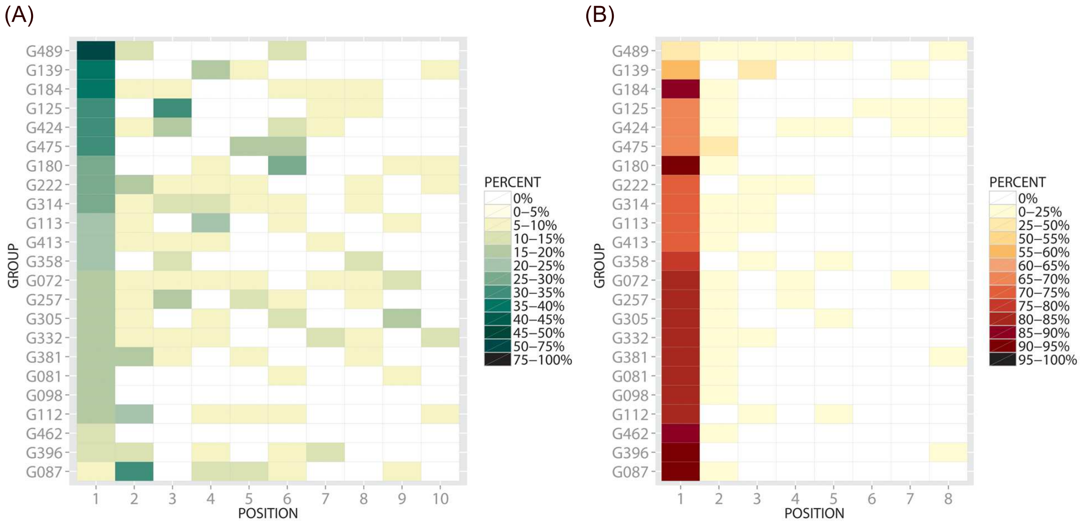


Figure 4.2: Position of the first correct and incorrect contact - Percent of cases where the first correct (A) and first incorrect (B) prediction is in the reported position for each group. Rows are ordered according to the percentage in the first column of A. The data are shown for the long-range contacts in FM domains.

targets are summarized in Table 4.2 and the example for target T0658 is shown in Fig. 4.3. Table 4.2 shows that in this analysis the best results are achieved by group *G489*, followed by groups *G112* and *G072*. Also in this case, one can ask the question of how often the contacts predicted with the highest probabilities are correct. The results, shown in Fig. 4.4 again highlight that group *G489* is particularly effective in ranking the predicted contacts.

GROUP	FP	TP	Precision (%)
G489	265	18	0.064
G087	259	7	0.026
G072	261	7	0.026
G475	84	2	0.023
G112	246	5	0.02
G381	213	3	0.014
G334	74	1	0.013
G081	217	2	0.009
G332	261	2	0.008
G139	182	1	0.005
G180	217	1	0.005
G424	231	1	0.004
G077	35	0	0
G098	221	0	0
G113	231	0	0
G125	259	0	0
G184	232	0	0
G222	249	0	0
G257	305	0	0
G305	305	0	0
G314	305	0	0
G358	212	0	0
G396	305	0	0
G413	218	0	0
G462	215	0	0

Table 4.2: Results of inter-domain predictions - Results of the Prediction of Long-Range Contacts in Which the Contacting Residues Belong to Two Different Domains
.

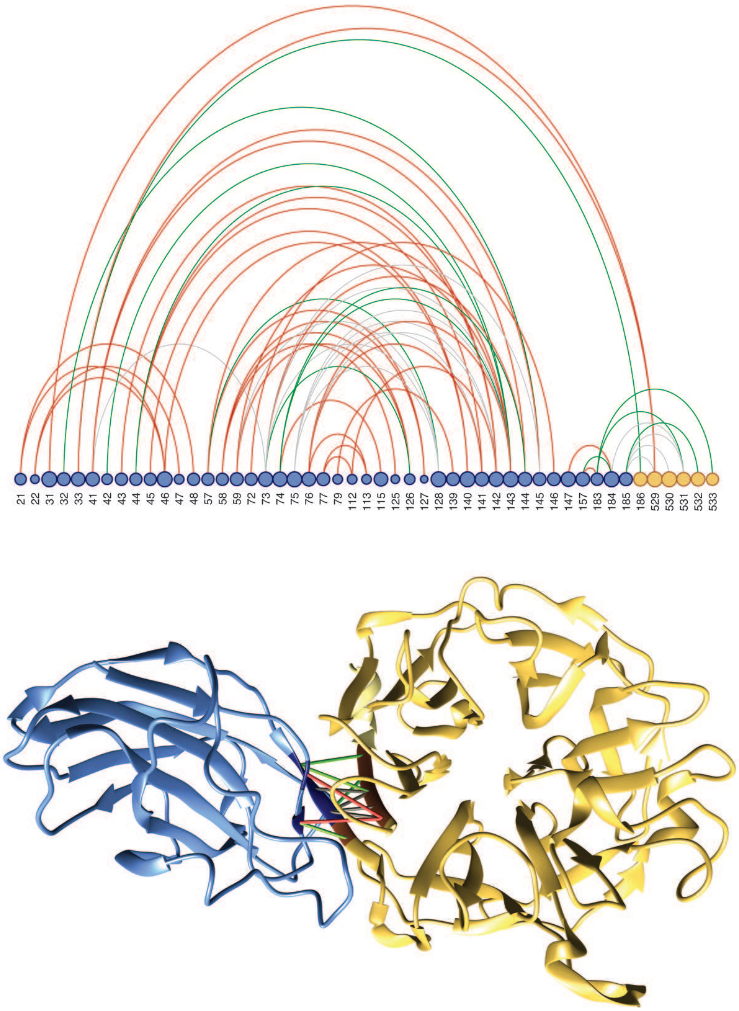


Figure 4.3: Prediction of inter-domain contacts for target T0658 - [The caption is on the next page [2]]

[0] This is a two domain protein with the first domain (residues 20185) being an FM target and the second (residues 186540)a template based target.The top panel shows L/5 contacts correctly predicted by at least one group as arcs connecting the corresponding residues indicated by circles.We show all the residues involved in correctly predicted contacts in the first (FM) domain, both intra- and inter-domain, and only the residues involved in correctly predicted inter-domain contacts for the second (TBM) domain. The size of the circle is proportional to the number of contacts the residue makes in the experimental structure. Blue and yellow circles are residues belonging to the first and second domain, respectively. The colour of the connecting arcs indicates the frequency with which the corresponding contact was predicted by the groups. Red, green, and grey lines indicate contacts predicted with a frequency below the median, between the median and the third quartile and above the third quartile, respectively. The bottom figure shows the three-dimensional structure of the protein with the first domain in blue and the second in yellow. The correctly predicted contacts are indicated by sticks with the same colour scheme as the corresponding arcs in the top panel.

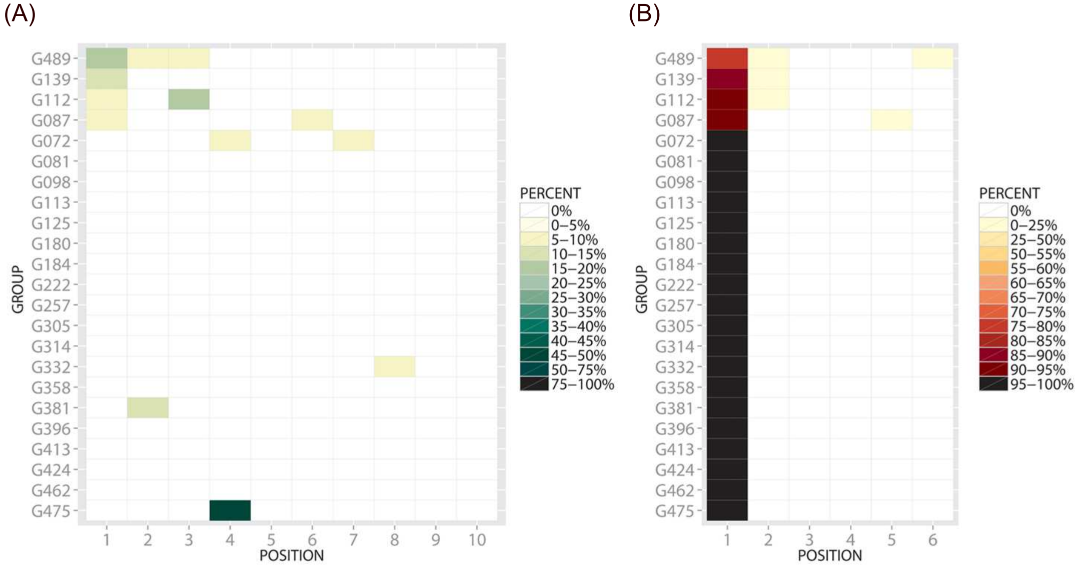


Figure 4.4: Inter-domain: position of the first correct and incorrect prediction - Percent of cases where the first correct (A) and first incorrect (B) prediction for inter-domain contacts is in the reported position for each group. Rows are ordered according to the percentage in the first column of A. The data are shown for the long-range contacts in FM domains.

5

Conclusions

Biology is undergoing technological revolutions at an unprecedented speed and often the experimental advances are not immediately amenable for existing computational tools. In this race between obtaining the data and interpreting them correctly, thereby adding value to the experimental results, it is essential to have two complementary computational approaches. On one side, the various existing methods have to be integrated and made easy to use, on the other, new methodologies need to be developed and tested. It follows that it is important to build flexible systems that also have the property to be designed in such a way to simplify the comparisons of results obtained using different parameters or by replacing tools with more recent ones.

This thesis falls in this context. During my PhD I have developed new needed tools (FIDEA, IsomirT, Phagotto), implemented pipelines endowed with the properties described above, and applied them to a number of experimental configurations.

As it is well known, the computational analysis of high throughput data is meant to provide a prioritisation list for further experiments, most of which are undergoing in the experiments analysed here. Nevertheless, the various collaborations have allowed me to understand what is needed and to make the selection of the future validation experiments more robust.

6

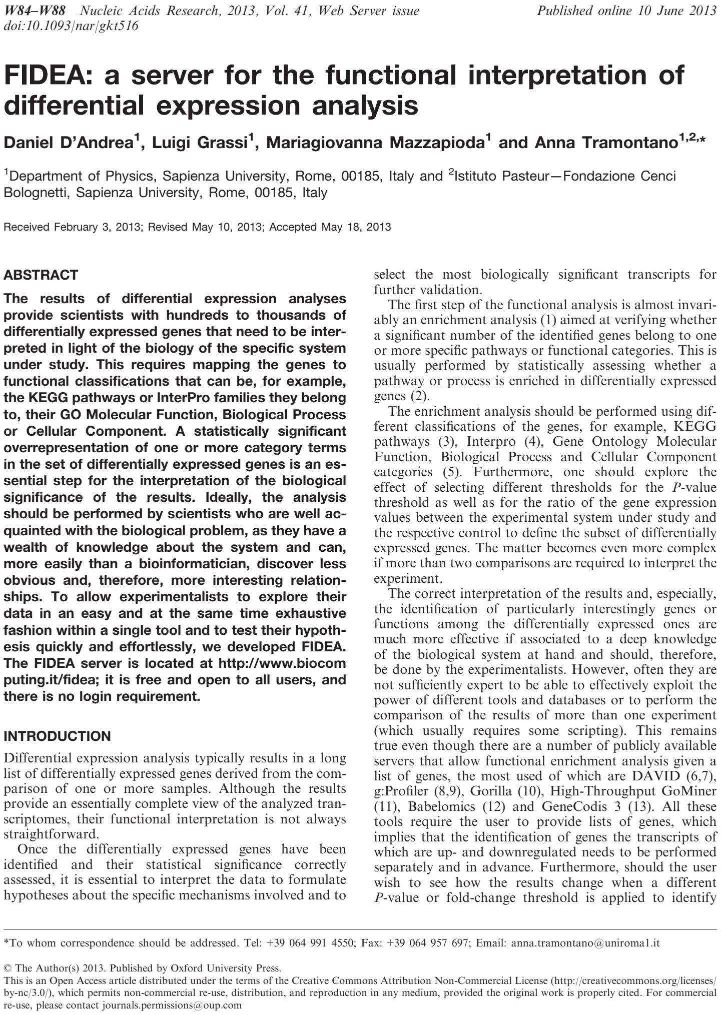
Appendices

6.1 Table of all analysed experiments

ID	Type	Sequencing	Aim	Samples
1	RNA-Seq (Prof. Bozzoni)	86 cycles in paired end	Evaluate the contribution of long non-coding RNAs in the molecular circuitries controlling myogenesis.	*C_2C_12* undifferentiated myoblasts, *C_2C_12* myoblasts at three differentiation times: one day, three days and five days.
2	RNA-Seq (Prof. Bozzoni)	74 cycles in paired end	To study how long non-coding RNAs affect on circuitries controlling proper muscle differentiation.	Proliferant cells;Cells in a early state of muscle differentiation.
3	RNA-Seq (Prof. Bozzoni)	74 cycles in paired end	Identification of differentially expressed genes and isoforms associated with familial forms of Amyotrophic Lateral Sclerosis (ALS).	Undifferentiated fibroblast Wild Type; Fibroblast Wild Type in a early state of differentiation; Differentiated fibroblast Wild Type; Differentiated fibroblast ALS patient II; Differentiated fibroblast ALS patient III.
4	ChIP-Seq (Prof. Levrero)	36 cycles in single end	Identification of target genes and microRNAs of HBV oncoprotein HBx andvalidation and refinement of existing hepatocellular carcinoma molecular signatures.	1 replicate of the DNA sample removed prior to immunoprecipitation; 2 replicates of DNA obtained from infected cells and then immunoprecipitated with anti-Hbx antibody; 2 replicates of DNA obtained from not infected cells and then immunoprecipitated with anti-Hbx antibody; DNA obtained from infected cells and then immunoprecipitated with anti-PolII antibody; DNA obtained from not infected cells and then immunoprecipitated with anti-PolII antibody.
5	ChIP-Seq (Prof. Levrero)	36 cycles in single end	Identification of target genes and microRNAs of HBV oncoprotein HBx andvalidation and refinement of existing hepatocellular carcinoma molecular signatures.	1 replicate of DNA sample removed prior to immunoprecipitation; 2 replicates of DNA obtained from infected cells and immunoprecipitated with anti-Hbx antibody; 1 replicate of DNA obtained from not infected cells and immunoprecipitated with anti-Hbx antibody.
6	ChIP-Seq (Prof. Levrero)	36 cycles in single end	Identification of target genes and microRNAs of HBV oncoprotein HBx andvalidation and refinement of existing hepatocellular carcinoma molecular signatures.	1 replicate of DNA sample removed prior to immunoprecipitation; 1 replicate of DNA obtained from infected cells and immunoprecipitated with anti-Hbx antibody; 1 replicate of DNA obtained from infected cells and immunoprecipitated without anti-Hbx antibody.
7	RNA-Seq (Prof. Levrero)	70 cycles in paired end	Identification of target genes and microRNAs of HBV oncoprotein HBx andvalidation and refinement of existing hepatocellular carcinoma molecular signatures.	Wild type and mock cells for three different system of infection.
8	DNA-Seq (Prof. Felici)	78 cycles in paired end	Introduction of a method capable of providing a detailed analysis of the fine specificity of vaccine-induced antibody repertoires in order to guide rational antigen design and selection of appropriate adjuvants.	5 no-selected libraries, 15 first selection libraries and 8 second selection libraries. These includes different adult, adolescent and infant sera pools from healthy volunteers vaccinated with the 4CMenB vaccine (Bexsero) in the course of a phase I clinical trial.
9	smallRNA-Seq (Public data (He *et al.*, 2012))	36 cycles in single end	Analysing miRNA expression in neurons and revealing the expression of a large fraction of known miRNAs with distinct profiles in glutamatergic and GABAergic neurons.	5 different embryonic stem (ES) cell lines in triplicate: Neocortex, Gad2, Camk2a, Purkinje, Cerebellum.
10	smallRNA-Seq (Public data (Chiang *et al.*, 2010))	36 cycles in single end	Experimental evaluation of novel and previously annotated mammalian microRNAs	Mouse brain, ovary, testes, embryonic stem cells, three embryonic stages, and whole newborns. All in duplicates.
11	smallRNA-Seq (Public data (Kuchen *et al.*, 2010))	36 cycles in single end	Regulation of microRNA Expression and Abundance during Lymphopoiesis	27 well defined cell types from the mouse immune system, hematopoietic progenitor cells, embryonic stem cells, as well as 12 tissues.

Table 6.1: All the analysed experiments during my Ph.D.. The experiments were performed on Illumina GAIIx except exp. 8 that was performed on Illumina MiSeq.

6.2 Pubblications



W84–W88 Nucleic Acids Research, 2013, Vol. 41, Web Server issue *Published online 10 June 2013*
doi:10.1093/nar/gkt516

FIDEA: a server for the functional interpretation of differential expression analysis

Daniel D'Andrea[1], Luigi Grassi[1], Mariagiovanna Mazzapioda[1] and Anna Tramontano[1,2,*]

[1]Department of Physics, Sapienza University, Rome, 00185, Italy and [2]Istituto Pasteur—Fondazione Cenci Bolognetti, Sapienza University, Rome, 00185, Italy

Received February 3, 2013; Revised May 10, 2013; Accepted May 18, 2013

ABSTRACT

The results of differential expression analyses provide scientists with hundreds to thousands of differentially expressed genes that need to be interpreted in light of the biology of the specific system under study. This requires mapping the genes to functional classifications that can be, for example, the KEGG pathways or InterPro families they belong to, their GO Molecular Function, Biological Process or Cellular Component. A statistically significant overrepresentation of one or more category terms in the set of differentially expressed genes is an essential step for the interpretation of the biological significance of the results. Ideally, the analysis should be performed by scientists who are well acquainted with the biological problem, as they have a wealth of knowledge about the system and can, more easily than a bioinformatician, discover less obvious and, therefore, more interesting relationships. To allow experimentalists to explore their data in an easy and at the same time exhaustive fashion within a single tool and to test their hypothesis quickly and effortlessly, we developed FIDEA. The FIDEA server is located at http://www.biocomputing.it/fidea; it is free and open to all users, and there is no login requirement.

INTRODUCTION

Differential expression analysis typically results in a long list of differentially expressed genes derived from the comparison of one or more samples. Although the results provide an essentially complete view of the analyzed transcriptomes, their functional interpretation is not always straightforward.

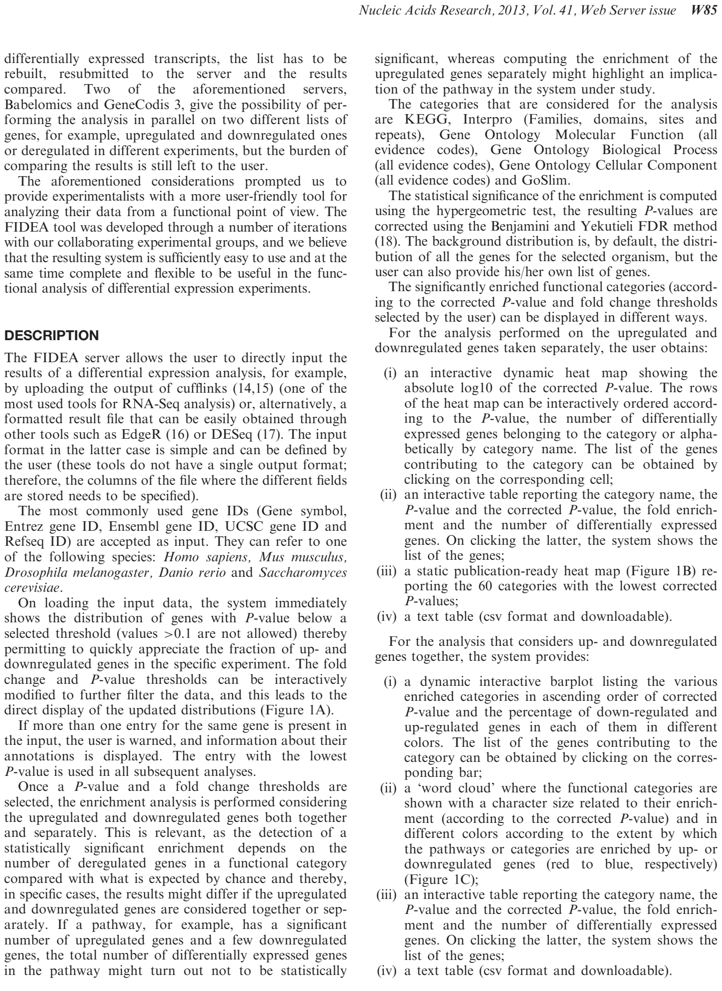
Once the differentially expressed genes have been identified and their statistical significance correctly assessed, it is essential to interpret the data to formulate hypotheses about the specific mechanisms involved and to select the most biologically significant transcripts for further validation.

The first step of the functional analysis is almost invariably an enrichment analysis (1) aimed at verifying whether a significant number of the identified genes belong to one or more specific pathways or functional categories. This is usually performed by statistically assessing whether a pathway or process is enriched in differentially expressed genes (2).

The enrichment analysis should be performed using different classifications of the genes, for example, KEGG pathways (3), Interpro (4), Gene Ontology Molecular Function, Biological Process and Cellular Component categories (5). Furthermore, one should explore the effect of selecting different thresholds for the *P*-value threshold as well as for the ratio of the gene expression values between the experimental system under study and the respective control to define the subset of differentially expressed genes. The matter becomes even more complex if more than two comparisons are required to interpret the experiment.

The correct interpretation of the results and, especially, the identification of particularly interestingly genes or functions among the differentially expressed ones are much more effective if associated to a deep knowledge of the biological system at hand and should, therefore, be done by the experimentalists. However, often they are not sufficiently expert to be able to effectively exploit the power of different tools and databases or to perform the comparison of the results of more than one experiment (which usually requires some scripting). This remains true even though there are a number of publicly available servers that allow functional enrichment analysis given a list of genes, the most used of which are DAVID (6,7), g:Profiler (8,9), Gorilla (10), High-Throughput GoMiner (11), Babelomics (12) and GeneCodis 3 (13). All these tools require the user to provide lists of genes, which implies that the identification of genes the transcripts of which are up- and downregulated needs to be performed separately and in advance. Furthermore, should the user wish to see how the results change when a different *P*-value or fold-change threshold is applied to identify

*To whom correspondence should be addressed. Tel: +39 064 991 4550; Fax: +39 064 957 697; Email: anna.tramontano@uniroma1.it



differentially expressed transcripts, the list has to be rebuilt, resubmitted to the server and the results compared. Two of the aforementioned servers, Babelomics and GeneCodis 3, give the possibility of performing the analysis in parallel on two different lists of genes, for example, upregulated and downregulated ones or deregulated in different experiments, but the burden of comparing the results is still left to the user.

The aforementioned considerations prompted us to provide experimentalists with a more user-friendly tool for analyzing their data from a functional point of view. The FIDEA tool was developed through a number of iterations with our collaborating experimental groups, and we believe that the resulting system is sufficiently easy to use and at the same time complete and flexible to be useful in the functional analysis of differential expression experiments.

DESCRIPTION

The FIDEA server allows the user to directly input the results of a differential expression analysis, for example, by uploading the output of cufflinks (14,15) (one of the most used tools for RNA-Seq analysis) or, alternatively, a formatted result file that can be easily obtained through other tools such as EdgeR (16) or DESeq (17). The input format in the latter case is simple and can be defined by the user (these tools do not have a single output format; therefore, the columns of the file where the different fields are stored needs to be specified).

The most commonly used gene IDs (Gene symbol, Entrez gene ID, Ensembl gene ID, UCSC gene ID and Refseq ID) are accepted as input. They can refer to one of the following species: *Homo sapiens, Mus musculus, Drosophila melanogaster, Danio rerio* and *Saccharomyces cerevisiae*.

On loading the input data, the system immediately shows the distribution of genes with *P*-value below a selected threshold (values >0.1 are not allowed) thereby permitting to quickly appreciate the fraction of up- and downregulated genes in the specific experiment. The fold change and *P*-value thresholds can be interactively modified to further filter the data, and this leads to the direct display of the updated distributions (Figure 1A).

If more than one entry for the same gene is present in the input, the user is warned, and information about their annotations is displayed. The entry with the lowest *P*-value is used in all subsequent analyses.

Once a *P*-value and a fold change thresholds are selected, the enrichment analysis is performed considering the upregulated and downregulated genes both together and separately. This is relevant, as the detection of a statistically significant enrichment depends on the number of deregulated genes in a functional category compared with what is expected by chance and thereby, in specific cases, the results might differ if the upregulated and downregulated genes are considered together or separately. If a pathway, for example, has a significant number of upregulated genes and a few downregulated genes, the total number of differentially expressed genes in the pathway might turn out not to be statistically significant, whereas computing the enrichment of the upregulated genes separately might highlight an implication of the pathway in the system under study.

The categories that are considered for the analysis are KEGG, Interpro (Families, domains, sites and repeats), Gene Ontology Molecular Function (all evidence codes), Gene Ontology Biological Process (all evidence codes), Gene Ontology Cellular Component (all evidence codes) and GoSlim.

The statistical significance of the enrichment is computed using the hypergeometric test, the resulting *P*-values are corrected using the Benjamini and Yekutieli FDR method (18). The background distribution is, by default, the distribution of all the genes for the selected organism, but the user can also provide his/her own list of genes.

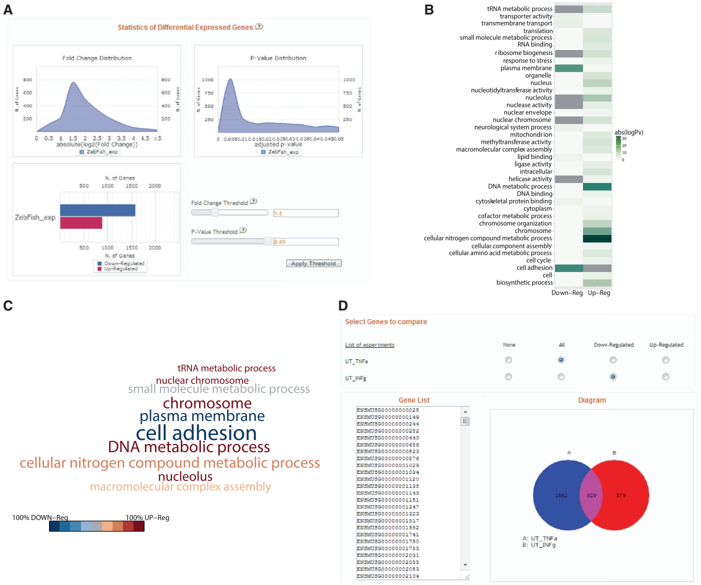
The significantly enriched functional categories (according to the corrected *P*-value and fold change thresholds selected by the user) can be displayed in different ways.

For the analysis performed on the upregulated and downregulated genes taken separately, the user obtains:

(i) an interactive dynamic heat map showing the absolute log10 of the corrected *P*-value. The rows of the heat map can be interactively ordered according to the *P*-value, the number of differentially expressed genes belonging to the category or alphabetically by category name. The list of the genes contributing to the category can be obtained by clicking on the corresponding cell;
(ii) an interactive table reporting the category name, the *P*-value and the corrected *P*-value, the fold enrichment and the number of differentially expressed genes. On clicking the latter, the system shows the list of the genes;
(iii) a static publication-ready heat map (Figure 1B) reporting the 60 categories with the lowest corrected *P*-values;
(iv) a text table (csv format and downloadable).

For the analysis that considers up- and downregulated genes together, the system provides:

(i) a dynamic interactive barplot listing the various enriched categories in ascending order of corrected *P*-value and the percentage of down-regulated and up-regulated genes in each of them in different colors. The list of the genes contributing to the category can be obtained by clicking on the corresponding bar;
(ii) a 'word cloud' where the functional categories are shown with a character size related to their enrichment (according to the corrected *P*-value) and in different colors according to the extent by which the pathways or categories are enriched by up- or downregulated genes (red to blue, respectively) (Figure 1C);
(iii) an interactive table reporting the category name, the *P*-value and the corrected *P*-value, the fold enrichment and the number of differentially expressed genes. On clicking the latter, the system shows the list of the genes;
(iv) a text table (csv format and downloadable).



List of esperiments	None	All	Down-Regulated	Up-Regulated
UT_TNFa	○	●	○	○
UT_INFg	○	○	●	○

ENSMUSG00000000028
ENSMUSG00000000149
ENSMUSG00000000244
ENSMUSG00000000282
ENSMUSG00000000440
ENSMUSG00000000686
ENSMUSG00000000823
ENSMUSG00000000876
ENSMUSG00000001029
ENSMUSG00000001034
ENSMUSG00000001120
ENSMUSG00000001138
ENSMUSG00000001143
ENSMUSG00000001151
ENSMUSG00000001247
ENSMUSG00000001323
ENSMUSG00000001517
ENSMUSG00000001552
ENSMUSG00000001741
ENSMUSG00000001750
ENSMUSG00000001755
ENSMUSG00000002031
ENSMUSG00000002055
ENSMUSG00000002083
ENSMUSG00000002104

A: UT_TNFa
B: UT_INFg

Figure 1. The figure shows an example of the results of FIDEA. Panel (**A**) is the first page where, on uploading the data, the user has an overview of the distributions of fold changes and *P*-values for up- and downregulated genes and can interactively modify the *P*-value and fold-change thresholds. The results of the GOSlim analysis are shown as both a heat map (**B**) and a word cloud (**C**). Panel (**D**) shows an example of how data from different experiments can be combined and subsequently analyzed.

When more than one experiment is uploaded, the user can obtain a list of genes that are in common among experiments. These can include genes that are upregulated in all the experiments, upregulated in one and downregulated in another and so forth (Figure 1D). The results are shown as a Venn diagram and as a list of genes that can be directly submitted to the functional enrichment analysis or downloaded.

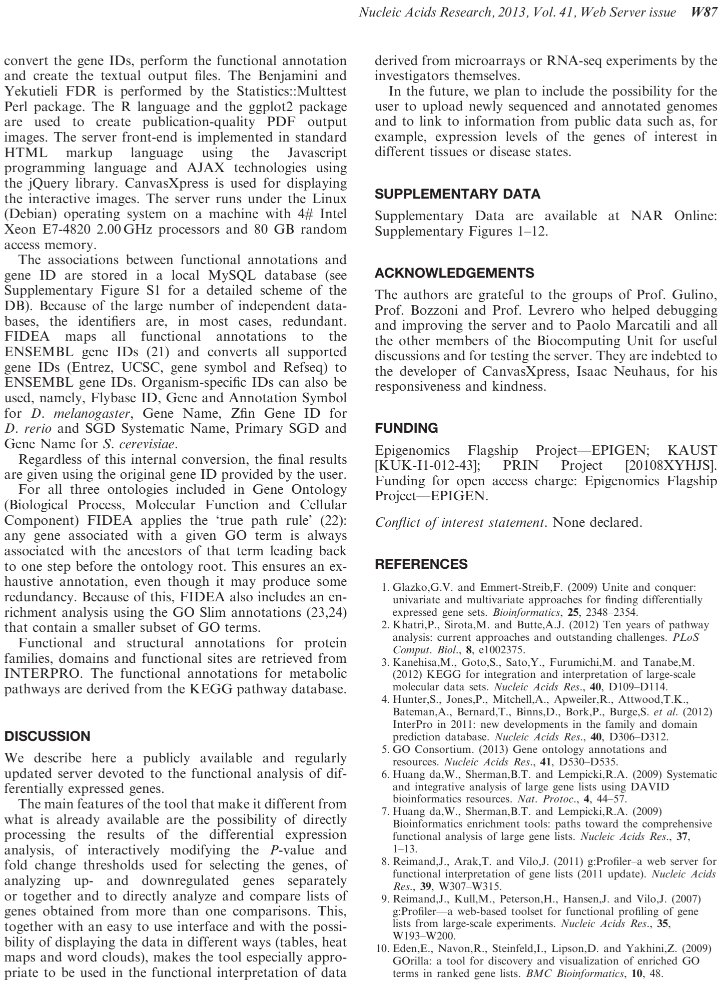
The server is regularly updated in parallel with new releases of Ensembl and of the functional classification annotations.

As an example, Figure 1 and Supplementary Figures 2–6 show some of the results obtained using the server for the data described previously (19), the authors of which performed an RNA-seq experiment of Id2a-deficient retinae obtained from zebrafish embryos in which Id2a expression was blocked by morpholino-mediated knockdown. As described by the authors, the data show an enrichment of downregulated genes belonging to the "cell adhesion" GO biological process and an enrichment of upregulated genes in the 'RNA processing' and 'nitrogen compound biosynthesis' processes.

The different results that can be obtained from the system according to the different functional categories are shown in the Supplementary Figures. These were obtained using the data from RNA sequencing experiments aimed at identifying transcripts expressed in human islets of Langerhans under control conditions or following exposure to pro-inflammatory cytokines (20).

IMPLEMENTATION

The FIDEA core consists in a Perl code and a set of Perl modules. The Perl modules are used to process the input,



convert the gene IDs, perform the functional annotation and create the textual output files. The Benjamini and Yekutieli FDR is performed by the Statistics::Multtest Perl package. The R language and the ggplot2 package are used to create publication-quality PDF output images. The server front-end is implemented in standard HTML markup language using the Javascript programming language and AJAX technologies using the jQuery library. CanvasXpress is used for displaying the interactive images. The server runs under the Linux (Debian) operating system on a machine with 4# Intel Xeon E7-4820 2.00 GHz processors and 80 GB random access memory.

The associations between functional annotations and gene ID are stored in a local MySQL database (see Supplementary Figure S1 for a detailed scheme of the DB). Because of the large number of independent databases, the identifiers are, in most cases, redundant. FIDEA maps all functional annotations to the ENSEMBL gene IDs (21) and converts all supported gene IDs (Entrez, UCSC, gene symbol and Refseq) to ENSEMBL gene IDs. Organism-specific IDs can also be used, namely, Flybase ID, Gene and Annotation Symbol for *D. melanogaster*, Gene Name, Zfin Gene ID for *D. rerio* and SGD Systematic Name, Primary SGD and Gene Name for *S. cerevisiae*.

Regardless of this internal conversion, the final results are given using the original gene ID provided by the user.

For all three ontologies included in Gene Ontology (Biological Process, Molecular Function and Cellular Component) FIDEA applies the 'true path rule' (22): any gene associated with a given GO term is always associated with the ancestors of that term leading back to one step before the ontology root. This ensures an exhaustive annotation, even though it may produce some redundancy. Because of this, FIDEA also includes an enrichment analysis using the GO Slim annotations (23,24) that contain a smaller subset of GO terms.

Functional and structural annotations for protein families, domains and functional sites are retrieved from INTERPRO. The functional annotations for metabolic pathways are derived from the KEGG pathway database.

DISCUSSION

We describe here a publicly available and regularly updated server devoted to the functional analysis of differentially expressed genes.

The main features of the tool that make it different from what is already available are the possibility of directly processing the results of the differential expression analysis, of interactively modifying the *P*-value and fold change thresholds used for selecting the genes, of analyzing up- and downregulated genes separately or together and to directly analyze and compare lists of genes obtained from more than one comparisons. This, together with an easy to use interface and with the possibility of displaying the data in different ways (tables, heat maps and word clouds), makes the tool especially appropriate to be used in the functional interpretation of data derived from microarrays or RNA-seq experiments by the investigators themselves.

In the future, we plan to include the possibility for the user to upload newly sequenced and annotated genomes and to link to information from public data such as, for example, expression levels of the genes of interest in different tissues or disease states.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–12.

ACKNOWLEDGEMENTS

The authors are grateful to the groups of Prof. Gulino, Prof. Bozzoni and Prof. Levrero who helped debugging and improving the server and to Paolo Marcatili and all the other members of the Biocomputing Unit for useful discussions and for testing the server. They are indebted to the developer of CanvasXpress, Isaac Neuhaus, for his responsiveness and kindness.

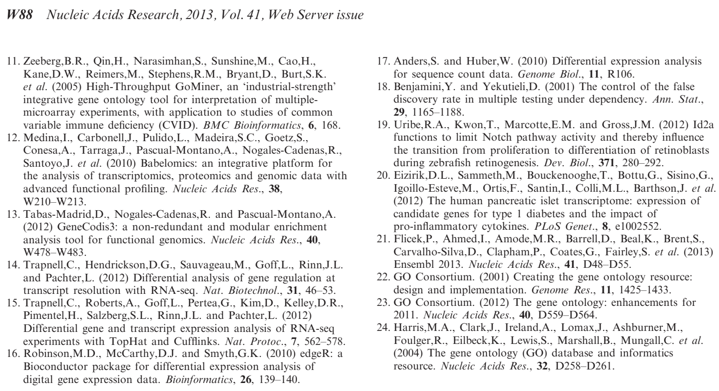
FUNDING

Epigenomics Flagship Project—EPIGEN; KAUST [KUK-I1-012-43]; PRIN Project [20108XYHJS]. Funding for open access charge: Epigenomics Flagship Project—EPIGEN.

Conflict of interest statement. None declared.

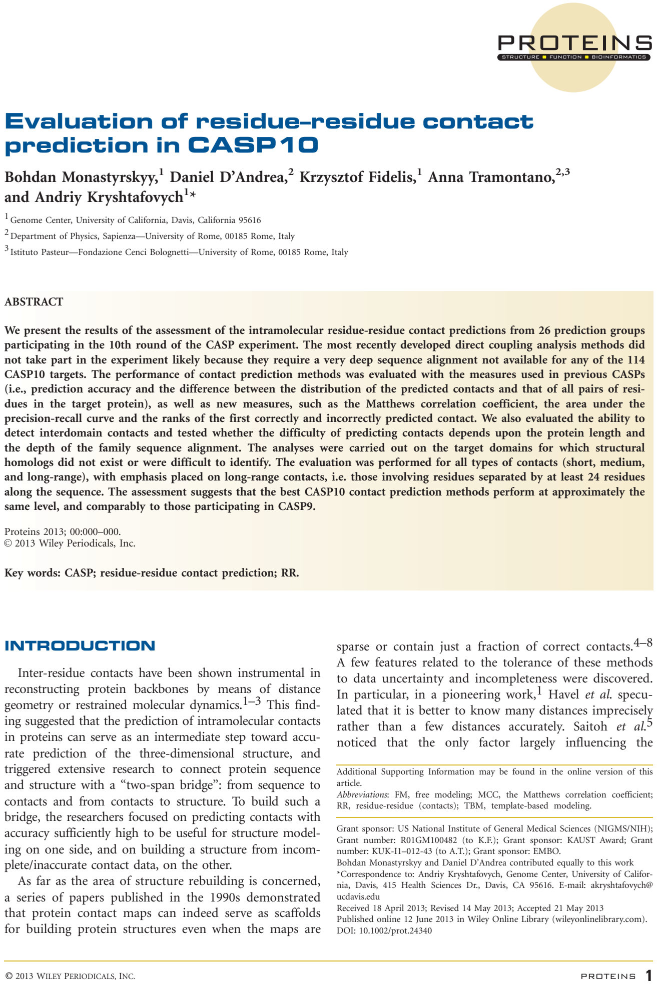
REFERENCES

1. Glazko,G.V. and Emmert-Streib,F. (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, **25**, 2348–2354.
2. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
3. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
4. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
5. GO Consortium. (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
6. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
7. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
8. Reimand,J., Arak,T. and Vilo,J. (2011) g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.*, **39**, W307–W315.
9. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–W200.
10. Eden,E., Navon,R., Steinfeld,I., Lipson,D. and Yakhini,Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.



11. Zeeberg,B.R., Qin,H., Narasimhan,S., Sunshine,M., Cao,H., Kane,D.W., Reimers,M., Stephens,R.M., Bryant,D., Burt,S.K. *et al*. (2005) High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, **6**, 168.
12. Medina,I., Carbonell,J., Pulido,L., Madeira,S.C., Goetz,S., Conesa,A., Tarraga,J., Pascual-Montano,A., Nogales-Cadenas,R., Santoyo,J. *et al*. (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res*., **38**, W210–W213.
13. Tabas-Madrid,D., Nogales-Cadenas,R. and Pascual-Montano,A. (2012) GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res*., **40**, W478–W483.
14. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol*., **31**, 46–53.
15. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc*., **7**, 562–578.
16. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
17. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.
18. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.
19. Uribe,R.A., Kwon,T., Marcotte,E.M. and Gross,J.M. (2012) Id2a functions to limit Notch pathway activity and thereby influence the transition from proliferation to differentiation of retinoblasts during zebrafish retinogenesis. *Dev. Biol*., **371**, 280–292.
20. Eizirik,D.L., Sammeth,M., Bouckenooghe,T., Bottu,G., Sisino,G., Igoillo-Esteve,M., Ortis,F., Santin,I., Colli,M.L., Barthson,J. *et al*. (2012) The human pancreatic islet transcriptome: expression of candidate genes for type 1 diabetes and the impact of pro-inflammatory cytokines. *PLoS Genet*., **8**, e1002552.
21. Flicek,P., Ahmed,I., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S. *et al*. (2013) Ensembl 2013. *Nucleic Acids Res*., **41**, D48–D55.
22. GO Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.
23. GO Consortium. (2012) The gene ontology: enhancements for 2011. *Nucleic Acids Res*., **40**, D559–D564.
24. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.

6. APPENDICES



PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Evaluation of residue-residue contact prediction in CASP10

Bohdan Monastyrskyy,[1] Daniel D'Andrea,[2] Krzysztof Fidelis,[1] Anna Tramontano,[2,3] and Andriy Kryshtafovych[1]*

[1] Genome Center, University of California, Davis, California 95616
[2] Department of Physics, Sapienza—University of Rome, 00185 Rome, Italy
[3] Istituto Pasteur—Fondazione Cenci Bolognetti—University of Rome, 00185 Rome, Italy

ABSTRACT

We present the results of the assessment of the intramolecular residue-residue contact predictions from 26 prediction groups participating in the 10th round of the CASP experiment. The most recently developed direct coupling analysis methods did not take part in the experiment likely because they require a very deep sequence alignment not available for any of the 114 CASP10 targets. The performance of contact prediction methods was evaluated with the measures used in previous CASPs (i.e., prediction accuracy and the difference between the distribution of the predicted contacts and that of all pairs of residues in the target protein), as well as new measures, such as the Matthews correlation coefficient, the area under the precision-recall curve and the ranks of the first correctly and incorrectly predicted contact. We also evaluated the ability to detect interdomain contacts and tested whether the difficulty of predicting contacts depends upon the protein length and the depth of the family sequence alignment. The analyses were carried out on the target domains for which structural homologs did not exist or were difficult to identify. The evaluation was performed for all types of contacts (short, medium, and long-range), with emphasis placed on long-range contacts, i.e. those involving residues separated by at least 24 residues along the sequence. The assessment suggests that the best CASP10 contact prediction methods perform at approximately the same level, and comparably to those participating in CASP9.

Proteins 2013; 00:000–000.

Key words: CASP; residue-residue contact prediction; RR.

INTRODUCTION

Inter-residue contacts have been shown instrumental in reconstructing protein backbones by means of distance geometry or restrained molecular dynamics.[1–3] This finding suggested that the prediction of intramolecular contacts in proteins can serve as an intermediate step toward accurate prediction of the three-dimensional structure, and triggered extensive research to connect protein sequence and structure with a "two-span bridge": from sequence to contacts and from contacts to structure. To build such a bridge, the researchers focused on predicting contacts with accuracy sufficiently high to be useful for structure modeling on one side, and on building a structure from incomplete/inaccurate contact data, on the other.

As far as the area of structure rebuilding is concerned, a series of papers published in the 1990s demonstrated that protein contact maps can indeed serve as scaffolds for building protein structures even when the maps are sparse or contain just a fraction of correct contacts.[4–8] A few features related to the tolerance of these methods to data uncertainty and incompleteness were discovered. In particular, in a pioneering work,[1] Havel *et al.* speculated that it is better to know many distances imprecisely rather than a few distances accurately. Saitoh *et al.*[5] noticed that the only factor largely influencing the

Additional Supporting Information may be found in the online version of this article.

Abbreviations: FM, free modeling; MCC, the Matthews correlation coefficient; RR, residue-residue (contacts); TBM, template-based modeling.

Grant sponsor: US National Institute of General Medical Sciences (NIGMS/NIH); Grant number: R01GM100482 (to K.F.); Grant sponsor: KAUST Award; Grant number: KUK-I1–012-43 (to A.T.); Grant sponsor: EMBO.
Bohdan Monastyrskyy and Daniel D'Andrea contributed equally to this work
*Correspondence to: Andriy Kryshtafovych, Genome Center, University of California, Davis, 415 Health Sciences Dr., Davis, CA 95616. E-mail: akryshtafovych@ucdavis.edu
Received 18 April 2013; Revised 14 May 2013; Accepted 21 May 2013
Published online 12 June 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24340

quality of the reconstructed structures is the long-range geometrical constraint. Skolnick *et al*. suggested[7] that knowing contacts for one in every seven residues would be sufficient to recover the structure of short proteins. Later, Vassura *et al*.[9] claimed that knowing one in four actual contacts might be enough to facilitate rebuilding tertiary structure with 5 Å accuracy. Although in general it is still unclear what accuracy, coverage, and distribution of contacts along the sequence are needed to be useful in practice, it has become common knowledge that information on just a few correct contacts can be valuable for improving structure prediction. This is especially true for the long-range contacts, which impose strong constraints on the three-dimensional structure and effectively narrow the search space of possible conformations. The usefulness of the contact approach was illustrated in the current edition of CASP, where predictors in the newly introduced contact-assisted structure prediction category (see the contact-assisted assessment article, this issue) were able to build substantially better models using information provided by the organizers on some of the long-range contacts in the target structures. Other studies also report that incorporating contact information into protein folding programs such as Rosetta and I-TASSER leads to improvement of the 3D models.[10,11]

Returning to the first bridge span in the "two-span bridge" analogy, substantial attention was dedicated to the prediction of intramolecular contacts. Much of the research in this area stemmed from the hypothesis of correlated mutations, suggesting that pairs of residues that mutate in a coordinated fashion during evolution are likely to be in contact. In the 1990s, the first articles demonstrating the applicability of this idea to contact prediction were published.[12–14] After these promising results, a series of contact prediction methods developing this concept further appeared in the literature.[15] Quite recently, the 20-year-old idea received a new twist as several articles claimed improved accuracy of contact prediction through disentangling the direct pairwise couplings from the background network of coordinately mutating positions.[15–22] Besides the coordinated mutations approaches, many other contact prediction methods were developed based on different or hybrid methodological concepts. In general, they are based on machine-learning techniques incorporating sequence-related features such as the sequence evolutionary profile of the target, secondary structure, and solvent accessibility—to name just a few. These methods use neural networks,[23–29] support vector machines,[30–32] hidden Markov models,[33–35] genetic algorithms,[36] random forest models,[37] and learning classifier systems.[38] Many of the methods mentioned above were tested in CASP experiments achieving different levels of success.

The prediction of residue-residue contacts has been a part of the CASP experiment since CASP2[39] (1996), however, the prediction format and the assessment procedures have been standardized only in CASP6–CASP9.[40–43] For CASP10, we developed an infrastructure for an automatic evaluation of the RR predictions and visual analysis of the results.[44] Here we analyze the results obtained by groups participating in CASP10 and quantify progress in the area compared with the previous CASPs.

MATERIALS AND METHODS

RR prediction format and definition of a contact

The RR prediction format and definition of intramolecular contacts in CASP10 have not changed since previous rounds of CASP. A pair of residues is defined to be in contact when the distance between their C_β atoms (C_α in case of GLY) is less than 8.0 Å. Depending on the separation along the sequence, short-, medium- and long-range contacts are between residues separated by 6 to 11, 12 to 23, and at least 24 residues, respectively. The contacts with a separation of less than six residues are not considered as they typically correspond to contacts within secondary structure elements. The participating groups were asked to submit a list of pairs of residues predicted to be in contact. Each reported contact had to be annotated with a probability score in the [0;1] range, reflecting the predictor confidence in assigning the contact. Unlike the previous rounds of CASP, only one set of contact predictions per target was allowed in CASP10 for each participating group.

Sets of domains evaluated

The evaluation of predictions was carried out on a per-domain basis. The domains with detectable homology to proteins of known structures were not included in the evaluation as in these cases contacts could easily be derived from the template structures. Thus, we used only the domains for which structural templates did not exist or were very difficult to identify, that is, the domains classified in the FM, TBM/FM, or TBM_hard categories.[45] The complete list of CASP10 domains with their classifications is available at http:/predictioncenter.org/casp10/domains_summary.cgi.

We assessed the performance of contact prediction methods on two sets of domains.

Set 1 (denoted as "FM") comprises 15 FM and 1 FM/TBM domains. For these domains templates did not exist or could not be reliably identified based on the target sequence. Set 1 is our main evaluation set and is consistent with the sets used in previous rounds of CASP.

Set 2 (hereinafter referred to as "FM + TBM_hard") is an extension of the previous set obtained by adding the domains from the TBM_hard category (13 entries). These are the hardest TBM targets, for which templates exist but are hard to identify or to properly align with the target. As a consequence, the scores of all submitted three-dimensional models for these targets were rather poor, not exceeding 50 GDT_TS units.[45]

We also performed the assessment on two sets of targets generated from the original two sets by eliminating non-globular proteins consisting of repeated structural blocks: Set 1R = Set 1 – {T0653-D1, T0695-D1}, and Set 2R = Set 2 – {T0653-D1, T0671-D2, T0690-D1, T0695-D1}.

The first three targets removed from Set 2 are the well-known leucine-rich repeats,[46] while the last one is a three-helical spectrin bundle repeated five times.[62] All four structures are built with repeated structural blocks for which good templates exist. Since the majority of contacts for these domains could be derived from the templates, their inclusion could introduce a bias in the evaluation. In practice, differences in the results on the original and the reduced sets were minor for the majority of analyses, and therefore we present here the results only for the original datasets, except for the domain-length dependence analysis, where using the reduced sets is more appropriate.

An estimate of the difficulty of individual domains for contact prediction is shown in Supporting Information Figure S1.

Sets of evaluated contacts

To compare the performance of contact prediction methods we used two different approaches. In the first approach, we trimmed the predicted lists of contacts to the same number of contacts per target (see the Reduced contact lists subsection below); in the second, we "padded" the lists by assigning a probability value of 0 to all non-listed contacts. The both procedures ensure that the participating groups are compared on the same number of contacts.

Preprocessing of predictions

For multidomain targets, we extracted the lists of inter-residue contacts for each individual domain. This step was necessary as predictions were submitted for the entire targets, but evaluated on a per-domain basis (see above). We also considered contacts between residues from different domains as their correct prediction can be useful in predicting the orientation of the interacting domains.

For each prediction, we separated short-, medium-, and long-range contacts and assessed them independently. The medium and long-range contacts were also assessed together.

Reduced contact lists

For every domain, the lists described above were trimmed to the *L*/5 and *L*/10 contacts predicted with higher probability (*L* is the length of the domain). The number *L*/5 (or *L*/10) is rounded to the closest integer, and if there are multiple entries corresponding to the same probability they are considered in the order provided by the predictor. To be included in the evaluation, the filtered list of contacts had to comprise at least *L*/5 or *L*/10 contacts. In order to assess also the groups that submitted only very small numbers of contacts, we also evaluated predictions on the five contacts with the highest assigned probability values, regardless of the domain length.

Thus, for every group we generated 12 reduced lists of contacts per predicted domain, whenever possible. The results for all lists of contacts and all contact range categories are available at http:/predictioncenter.org/casp10/rr_results.cgi. In this paper we focus on the results for the *L*/5 lists of long-range contacts. The numbers of domains predicted on these datasets for each of the participating groups are summarized in Figure 1. Two groups (G334 and G077) submitted just a few predictions for the evaluated domains and one (G246) did none, so we excluded them from the analysis and present the results on the reduced lists for the remaining 23 groups. For every group, the final scores on the reduced datasets are averages of the per-domain scores.

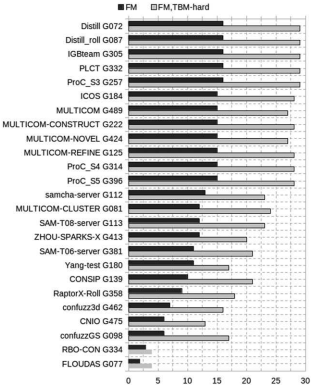


Figure 1

Number of domains per group for which the L/5 list of long-range contacts were evaluated. Two groups RBO-CON (G334) and FLOUDAS (G077) submitted too few predictions and are not included in the subsequent analyses.

Padded contact maps

As contact probability maps generated from submitted predictions are sparse, they are usually unsuitable for many analyses that require complete predictions (i.e. we need each pair of residues to be predicted either in contact or not). We remediate the "sparseness" problem here by setting the values of the empty cells of contact probability maps to zero ("padded" lists). In other words, pairs of residues that are missing in predictions are considered as non-contacts. Under such assumption, each prediction list classifies every pair of residues within the selected range to one of the four cases: TP, correctly predicted contact; FP, non-contact predicted as contact; TN, correctly "predicted" non-contact (i.e., the non-contact not included in the predicted contact list); and FN, contact "predicted" as non-contact (i.e., the contact missing in the submitted list).

We only assessed the groups that submitted predictions for at least 10 domains on the "padded" datasets—these are the same 23 groups as above, plus group G334. As in the case of the reduced contact lists, in this article we concentrate on the analysis of the performance of the participating groups for the long-range contacts only. Differently from the assessment on the reduced contact lists, the final group scores on the padded datasets are calculated from the data on all domains pooled together.

Evaluation procedure

In CASP10 we have substantially expanded the set of evaluation tools to assess residue-residue contact predictions. Besides the methods used in the previous CASPs, we introduced several new evaluations providing an alternative point of view on methods' performance. While in previous CASPs the assessors analyzed the results exclusively on the "reduced" datasets, implicitly concentrating on two aspects of contact prediction: (1) how good are methods in identifying the most reliable predicted contacts and (2) how accurate are the methods in predicting contacts with the highest reliability, in this CASP we complemented the assessment with analyses on the full sets of contacts addressing the issue of how accurate are all submitted contact predictions, including those predicted with lower reliability. Below, we briefly outline all evaluation procedures, focusing in more detail on the new evaluation measures.

Basic scoring functions and group performance on the reduced datasets

Since CASP6, predictions in the RR category have been evaluated on the reduced contact lists using two main scores: precision = TP/(TP + FP), and Xd. The detailed description of these scores can be found in the previous CASP contact assessment articles.[40–43] Note, that in those papers the measure defined by the formula TP/(TP + FP) was called "accuracy" (Acc); here we have changed its name to "precision" to be consistent with the classic descriptive statistics definition. The precision-based results are discussed in the main text of this article, while the Xd-based results are shown in the Supporting Information.

Based on these two scores, the performance of groups was further compared with two strategies: cumulative *z*-score ranking (sum of precision-based and Xd-based *z*-scores) and "head-to-head" comparisons.[43]

Evaluation measures for the padded datasets

Matthews' correlation coefficient and other binary descriptive statistics measures

For the assessment of the effectiveness of the predictive methods as binary classifiers we used four evaluation measures.

The first two are precision and recall, a.k.a. sensitivity:

$$precision=\frac{TP}{TP+FP}, \qquad recall=sensitivity=\frac{TP}{TP+FN}.$$

They were already used in previous CASPs, but were shown to be equivalent on the reduced prediction sets.[41] On the complete datasets, precision and recall are not inter-dependent any more as the number of predicted contacts is different for different predictions. Based on the formulae, one can notice that each of these measures takes into account only two out of the four parameters of prediction quality (TP, FP, TN, and FN) and therefore focuses on the specific aspects of predicting contacts only (ignoring non-contacts).

The *F*-score is a more comprehensive measure as it combines precision and recall

$$F_1=2\frac{precision*recall}{precision+recall}$$

and inherits useful features typical to both measures. However, the *F*-measure still does not take the true negative rate into account.

Even though employing measures that take all parameters of contact prediction into account may seem beneficial, it should be approached with caution, as in our case two binary classes of prediction (contacts and non-contacts) are disproportionally distributed in the structure (contacts constitute just a small fraction of all pairs of residues). As it was discussed in the CASP9 disorder assessment article,[47] the Matthews correlation coefficient (MCC)

$$MCC=\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

is a well-suited measure for handling cases with imbalanced class frequencies. The MCC was shown to provide a more appropriate account of the skewed data than many other methods, and not to favor over-prediction of

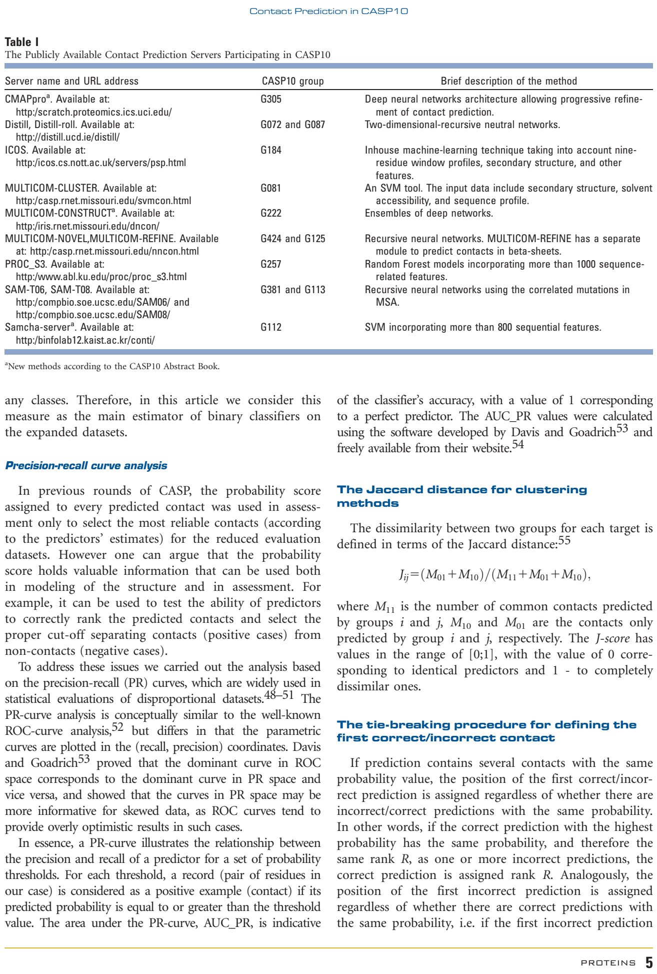


Table I
The Publicly Available Contact Prediction Servers Participating in CASP10

Server name and URL address	CASP10 group	Brief description of the method
CMAPpro[a]. Available at: http:/scratch.proteomics.ics.uci.edu/	G305	Deep neural networks architecture allowing progressive refinement of contact prediction.
Distill, Distill-roll. Available at: http://distill.ucd.ie/distill/	G072 and G087	Two-dimensional-recursive neutral networks.
ICOS. Available at: http:/icos.cs.nott.ac.uk/servers/psp.html	G184	Inhouse machine-learning technique taking into account nine-residue window profiles, secondary structure, and other features.
MULTICOM-CLUSTER. Available at: http:/casp.rnet.missouri.edu/svmcon.html	G081	An SVM tool. The input data include secondary structure, solvent accessibility, and sequence profile.
MULTICOM-CONSTRUCT[a]. Available at: http:/iris.rnet.missouri.edu/dncon/	G222	Ensembles of deep networks.
MULTICOM-NOVEL,MULTICOM-REFINE. Available at: http:/casp.rnet.missouri.edu/nncon.html	G424 and G125	Recursive neural networks. MULTICOM-REFINE has a separate module to predict contacts in beta-sheets.
PROC_S3. Available at: http:/www.abl.ku.edu/proc/proc_s3.html	G257	Random Forest models incorporating more than 1000 sequence-related features.
SAM-T06, SAM-T08. Available at: http:/compbio.soe.ucsc.edu/SAM06/ and http:/compbio.soe.ucsc.edu/SAM08/	G381 and G113	Recursive neural networks using the correlated mutations in MSA.
Samcha-server[a]. Available at: http:/binfolab12.kaist.ac.kr/conti/	G112	SVM incorporating more than 800 sequential features.

[a]New methods according to the CASP10 Abstract Book.

any classes. Therefore, in this article we consider this measure as the main estimator of binary classifiers on the expanded datasets.

Precision-recall curve analysis

In previous rounds of CASP, the probability score assigned to every predicted contact was used in assessment only to select the most reliable contacts (according to the predictors' estimates) for the reduced evaluation datasets. However one can argue that the probability score holds valuable information that can be used both in modeling of the structure and in assessment. For example, it can be used to test the ability of predictors to correctly rank the predicted contacts and select the proper cut-off separating contacts (positive cases) from non-contacts (negative cases).

To address these issues we carried out the analysis based on the precision-recall (PR) curves, which are widely used in statistical evaluations of disproportional datasets.[48–51] The PR-curve analysis is conceptually similar to the well-known ROC-curve analysis,[52] but differs in that the parametric curves are plotted in the (recall, precision) coordinates. Davis and Goadrich[53] proved that the dominant curve in ROC space corresponds to the dominant curve in PR space and vice versa, and showed that the curves in PR space may be more informative for skewed data, as ROC curves tend to provide overly optimistic results in such cases.

In essence, a PR-curve illustrates the relationship between the precision and recall of a predictor for a set of probability thresholds. For each threshold, a record (pair of residues in our case) is considered as a positive example (contact) if its predicted probability is equal to or greater than the threshold value. The area under the PR-curve, AUC_PR, is indicative of the classifier's accuracy, with a value of 1 corresponding to a perfect predictor. The AUC_PR values were calculated using the software developed by Davis and Goadrich[53] and freely available from their website.[54]

The Jaccard distance for clustering methods

The dissimilarity between two groups for each target is defined in terms of the Jaccard distance:[55]

$$J_{ij} = (M_{01} + M_{10}) / (M_{11} + M_{01} + M_{10}),$$

where M_{11} is the number of common contacts predicted by groups i and j, M_{10} and M_{01} are the contacts only predicted by group i and j, respectively. The *J-score* has values in the range of [0;1], with the value of 0 corresponding to identical predictors and 1 - to completely dissimilar ones.

The tie-breaking procedure for defining the first correct/incorrect contact

If prediction contains several contacts with the same probability value, the position of the first correct/incorrect prediction is assigned regardless of whether there are incorrect/correct predictions with the same probability. In other words, if the correct prediction with the highest probability has the same probability, and therefore the same rank *R*, as one or more incorrect predictions, the correct prediction is assigned rank *R*. Analogously, the position of the first incorrect prediction is assigned regardless of whether there are correct predictions with the same probability, i.e. if the first incorrect prediction

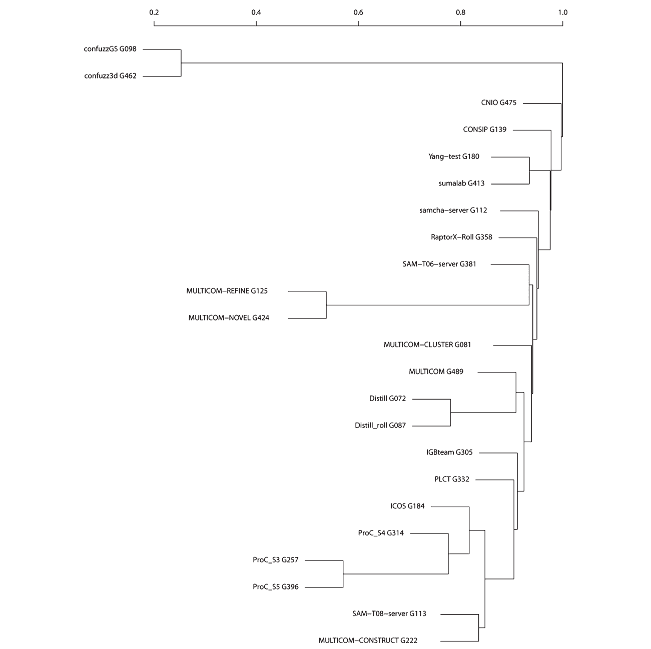


Figure 2
Dendrogram illustrating the similarity among different methods as judged by the number of common predictions for all targets.

has the same rank *R* as a correct prediction, the first incorrect prediction is assigned rank *R*.

RESULTS

Participating methods: Brief description and similarity

In CASP10 26 groups submitted predictions of intra-molecular contacts, including 22 automated servers and four expert groups. Three groups used new methods, while others used modified techniques developed earlier and tested in previous rounds of CASP. Table I presents a short description of the participating publicly available contact prediction servers. A more detailed overview of all the methods participating in CASP10 can be found in the CASP10 Abstract Book.[56]

Not all methods are conceptually different as oftentimes they rely on similar prediction techniques using similar mathematical apparatus and predictive features.

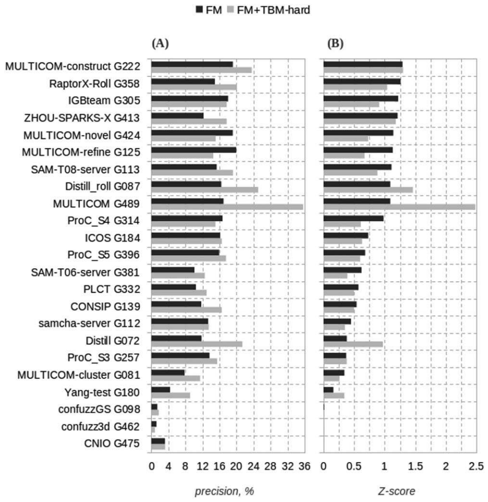


Figure 3

Precision (**A**) and cumulative *z*-score (**B**) for the participating groups on the two sets of the evaluated domains (FM and FM + TBM_hard). The data are shown for the top *L*/5 long-range contacts. Groups in both panels are ordered according to their cumulative *z*-score on FM targets.

To illustrate this, we clustered the methods participating in CASP10 based on the pair-wise Jaccard distance (see Materials). Figure 2 shows the results of the method clustering. As one can notice, four lowest level clusters encompass two prediction groups each from the same research centers, i.e. two Proc-S, Distill, Multicom, and confuzz methods. It is apparent that the clustered groups use similar methodologies with slight modifications in the implementation of the method.

Group performance on the reduced datasets: Precision and Xd

The results of the analysis of the group performance for long-range contacts in the *L*/5 contact lists are presented in Figure 3. For each group we show the values of precision and cumulative *z*-score (sum of precision-based and Xd-based *z*-scores) averaged over all predicted domains from the "FM" and "FM + TBM_hard" datasets (see Materials for a detailed description of the datasets and evaluation measures).

Panel A of Figure 3 demonstrates that the precision of the current prediction methods on FM targets does not exceed 20%. The three best performing groups on the FM targets (G125, G222, and G424) attain precision of 19% and belong to the same family of methods (Multicom, group leader J. Cheng, University of Missouri). Multicom-construct method (G222) was also shown to reach the highest score according to the Xd measure (see Fig. S2 in Supporting Information), and is ranked first according to the cumulative *z*-score (Fig. 3, panel B). It should be mentioned, though, that the difference in performance of this method and the others is marginal, as Student's *t*-tests did not reveal statistically significant difference in the performance of the top ten methods (see Table II for precision and Table S1 in Supporting Information for Xd). This statement is supported by the results of the "head-to-head"

Table II
Results of the Paired Student's *t*-Test on the Precision Score for (A) FM and (B) FM, TBM-Hard Domains for Top 10 Groups According to the Cumulative *z*-Score Ranking

A	G222	G358	G305	G413	G424	G125	G113	G087	G489	G314
G222	x	9	15	11	14	14	12	15	14	15
G358	0.49	x	9	6	8	8	9	9	8	9
G305	0.18	0.3	x	12	15	15	12	16	15	15
G413	0.11	0.19	0.19	x	11	11	8	12	11	11
G424	0.07	0.09	0.44	0.13	X	15	11	15	14	14
G125	0.16	0.1	0.39	0.13	0.26	x	11	15	14	14
G113	0.5	0.34	0.39	0.24	0.08	0.08	x	12	11	12
G087	0.26	0.46	0.4	0.46	0.36	0.32	0.44	x	15	15
G489	0.33	0.1	0.37	0.4	0.34	0.32	0.31	0.49	x	14
G314	0.19	0.21	0.41	0.09	0.21	0.34	0.37	0.48	0.48	x

B	G489	G087	G222	G413	G358	G072	G305	G113	G424	G125
G489	x	27	26	18	16	27	27	21	25	26
G087	0.03	x	28	20	18	29	29	23	27	28
G222	0.02	0.34	x	19	18	28	28	23	26	27
G413	0.05	0.34	0.08	x	10	20	20	14	18	19
G358	<0.01	0.03	0.26	0.12	x	18	18	18	17	17
G072	<0.01	0.08	0.36	0.43	0.11	x	29	23	27	28
G305	0.01	0.07	0.01	0.18	0.25	0.21	x	23	27	28
G113	<0.01	0.12	0.18	0.38	0.47	0.24	0.14	x	21	22
G424	0.01	0.03	<0.01	0.36	<0.01	0.11	0.17	0	x	27
G125	0.01	0.03	<0.01	0.33	<0.01	0.11	0.19	0	0.39	x

The tables show the *P* values (cells below the diagonal) of the Student's *t*-tests performed for each pair of the groups on the common set of domains (the numbers above the diagonal). Shaded cells indicate statistically indistinguishable results at the significance level of 0.05.

Table III
The "Head-to-Head" Comparison of the Performance of the Groups Based on the *precision* Score for (A) FM and (B) FM, TBM-Hard Domains for the Top 10 Groups According to the Cumulative *z*-Score Ranking

		Group 2									
A		G222	G358	G305	G413	G424	G125	G113	G087	G489	G314
Group 1	G222	X	44.4%	46.7%	63.6%	50.0%	50.0%	41.7%	66.7%	64.3%	46.7%
	G358	44.4%	x	44.4%	16.7%	75.0%	75.0%	33.3%	33.3%	37.5%	55.6%
	G305	33.3%	33.3%	x	58.3%	40.0%	40.0%	33.3%	50.0%	60.0%	40.0%
	G413	27.3%	66.7%	25.0%	x	36.4%	36.4%	25.0%	58.3%	45.5%	27.3%
	G424	21.4%	12.5%	53.3%	45.5%	x	13.3%	18.2%	46.7%	35.7%	28.6%
	G125	28.6%	12.5%	46.7%	45.5%	13.3%	x	18.2%	53.3%	35.7%	21.4%
	G113	33.3%	44.4%	50.0%	50.0%	63.6%	63.6%	x	66.7%	45.5%	50.0%
	G087	26.7%	55.6%	37.5%	41.7%	33.3%	26.7%	33.3%	x	26.7%	40.0%
	G489	28.6%	50.0%	40.0%	45.5%	50.0%	50.0%	45.5%	60.0%	x	35.7%
	G314	33.3%	33.3%	46.7%	54.5%	64.3%	57.1%	33.3%	60.0%	57.1%	x

B		Group 2									
		G489	G087	G222	G413	G358	G072	G305	G113	G424	G125
Group 1	G489	x	63.0%	53.8%	55.6%	75.0%	77.8%	66.7%	71.4%	72.0%	73.1%
	G087	25.9%	x	35.7%	45.0%	61.1%	55.2%	51.7%	43.5%	55.6%	50.0%
	G222	42.3%	60.7%	x	63.2%	55.6%	57.1%	60.7%	56.5%	69.2%	74.1%
	G413	38.9%	55.0%	31.6%	x	70.0%	50.0%	50.0%	35.7%	61.1%	63.2%
	G358	18.8%	27.8%	38.9%	20.0%	x	50.0%	50.0%	33.3%	76.5%	70.6%
	G072	14.8%	24.1%	35.7%	45.0%	44.4%	x	44.8%	39.1%	59.3%	57.1%
	G305	33.3%	31.0%	25.0%	40.0%	33.3%	41.4%	x	30.4%	51.9%	50.0%
	G113	23.8%	47.8%	26.1%	35.7%	44.4%	52.2%	47.8%	x	71.4%	72.7%
	G424	20.0%	29.6%	15.4%	27.8%	17.6%	33.3%	40.7%	14.3%	x	29.6%
	G125	19.2%	35.7%	14.8%	26.3%	17.6%	35.7%	39.3%	13.6%	14.8%	x

The rows show the fraction of common domains for which the precision score of the group in the row is higher than that of the group in the column. Cases of equal scores are not counted.

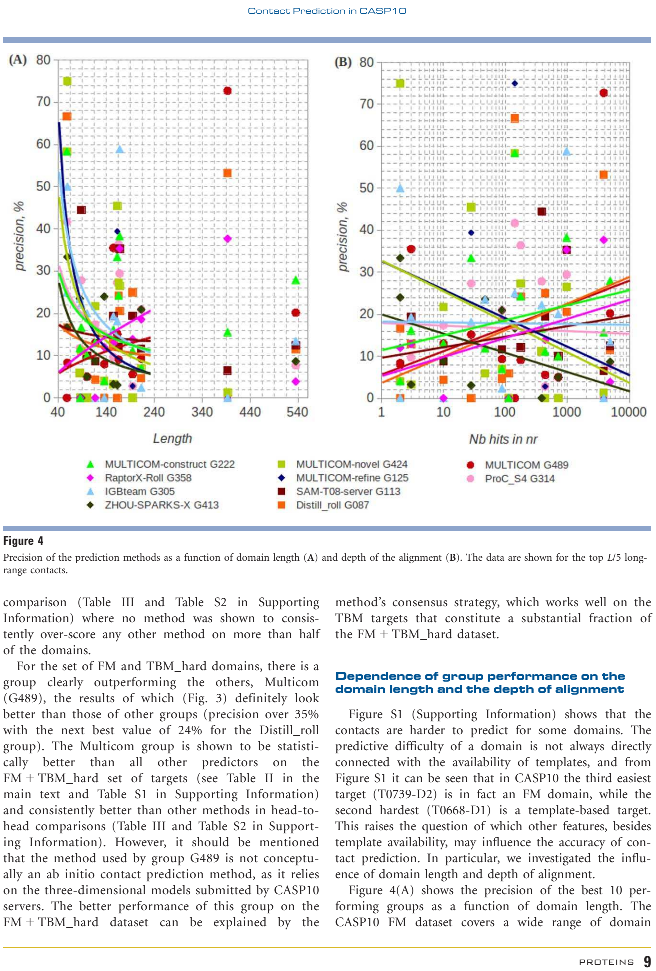


Figure 4

Precision of the prediction methods as a function of domain length (**A**) and depth of the alignment (**B**). The data are shown for the top *L*/5 long-range contacts.

comparison (Table III and Table S2 in Supporting Information) where no method was shown to consistently over-score any other method on more than half of the domains.

For the set of FM and TBM_hard domains, there is a group clearly outperforming the others, Multicom (G489), the results of which (Fig. 3) definitely look better than those of other groups (precision over 35% with the next best value of 24% for the Distill_roll group). The Multicom group is shown to be statistically better than all other predictors on the FM + TBM_hard set of targets (see Table II in the main text and Table S1 in Supporting Information) and consistently better than other methods in head-to-head comparisons (Table III and Table S2 in Supporting Information). However, it should be mentioned that the method used by group G489 is not conceptually an ab initio contact prediction method, as it relies on the three-dimensional models submitted by CASP10 servers. The better performance of this group on the FM + TBM_hard dataset can be explained by the method's consensus strategy, which works well on the TBM targets that constitute a substantial fraction of the FM + TBM_hard dataset.

Dependence of group performance on the domain length and the depth of alignment

Figure S1 (Supporting Information) shows that the contacts are harder to predict for some domains. The predictive difficulty of a domain is not always directly connected with the availability of templates, and from Figure S1 it can be seen that in CASP10 the third easiest target (T0739-D2) is in fact an FM domain, while the second hardest (T0668-D1) is a template-based target. This raises the question of which other features, besides template availability, may influence the accuracy of contact prediction. In particular, we investigated the influence of domain length and depth of alignment.

Figure 4(A) shows the precision of the best 10 performing groups as a function of domain length. The CASP10 FM dataset covers a wide range of domain

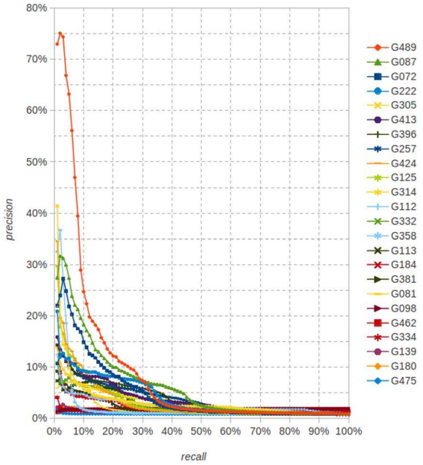


Figure 5

PR-curves for all predicted long-range contacts on FM domains.

length spanning from 58 to 535 residues. Two domains are short (under 60 residues), two rather long (over 390 residues) and the remaining 12 are of medium length (80–220 residues). On four of the domains (the shortest two and one from each of the medium and long sub-ranges), the best groups reach a very high precision (over 50%). It should be noticed, though, that the two longest domains in this graph (T0653-D1 and T0695-D1) represent non-globular targets with a repeated topology (see the description of Set 2R in Materials), and this may introduce bias in the analysis. Therefore, we analyzed per-group trends in the results excluding these two domains. Inspection of the graph reveals that the vast majority of groups reach better precision on shorter targets.

To analyze the dependence of group performance on the depth of the target alignments, we searched for sequence homologs for each target with PSI-BLAST[57] running five iterations against the non-redundant database with parameters "-h 0.05 -v 1000 -b 1000." The number of hits covering at least 75% of target's sequence was used as a measure of the alignment depth. The depth of the alignment for CASP10 FM targets varied from just a few hits (for T0726-D3, T0741-D1, T0740-D1) to more than a thousand for two repeat-topology domains (T0653-D1 and T0695-D1). Figure 4(B) shows that CASP10 methods are in general insensitive to the alignment depth, as no trend in the data can be detected. As precision of group performance depends on target length, we also tested a hypothesis that length can be a contributing factor in how precision depends on depth of alignment. Our additional analysis showed that this is not the case.

Group performance on the untrimmed contact lists: PR-curve and MCC analyses

Figure 5 and Table IV present a different perspective on the methods' performance based on the PR-curve analysis, *MCC* and other descriptive statistics measures (see Materials).

The PR-curve analysis clearly identifies the top performing group, G489 (Multicom), which reaches an *AUC_PR* score of 9.5%. Again, we remind here that this

Table IV
Descriptive Statistics Scores Calculated for the Predictions Treated in the Context of the Complete Contact Maps for Long-Range Contacts for FM Domains

Group	No dom	TP	FP	TN	FN	MCC	Precision (%)	Recall (%)	F_1	AUC_PR
G489	15	841	11,331	27,5798	1838	0.131	6.9	31.4	0.113	0.095
G087	16	1175	22,200	266,876	1510	0.127	5.0	43.8	0.090	0.065
G222	16	905	15,025	274,051	1780	0.120	5.7	33.7	0.097	0.043
G072	16	814	13,674	275,402	1871	0.112	5.6	30.3	0.095	0.049
G396	16	1006	21,195	267,881	1679	0.109	4.5	37.5	0.081	0.038
G257	16	1331	43,378	245,698	1354	0.092	3.0	49.6	0.056	0.038
G113	13	559	11,025	266,176	1785	0.091	4.8	23.8	0.080	0.025
G314	16	1255	42,987	246,089	1430	0.085	2.8	46.7	0.053	0.032
G125	16	597	12,454	276,622	2088	0.083	4.6	22.2	0.076	0.034
G413	12	841	34,071	152,543	742	0.082	2.4	53.1	0.046	0.038
G424	16	540	10,788	278,288	2145	0.081	4.8	20.1	0.077	0.035
G081	16	312	4287	284,789	2373	0.078	6.8	11.6	0.086	0.018
G112	13	1542	89,198	107,628	269	0.076	1.7	85.1	0.033	0.027
G184	16	570	15,934	273,142	2115	0.065	3.5	21.2	0.059	0.021
G139	10	1294	82,600	167,456	625	0.063	1.5	67.4	0.030	0.015
G332	16	738	25,302	263,774	1947	0.063	2.8	27.5	0.051	0.026
G381	13	448	13,232	263,969	1896	0.061	3.3	19.1	0.056	0.019
G305	16	1952	123,658	165,418	733	0.058	1.6	72.7	0.030	0.038
G358	16	154	1969	287,107	2531	0.057	7.3	5.7	0.064	0.026
G180	12	490	30,968	155,646	1093	0.035	1.6	31.0	0.030	0.012
G334	12	18	241	154,081	2166	0.019	6.9	0.8	0.014	0.015
G475	12	19	857	261,349	2169	0.009	2.2	0.9	0.012	0.009
G098	12	13	982	70,457	1345	−0.005	1.3	1.0	0.011	0.018
G462	12	11	981	70,458	1347	−0.007	1.1	0.8	0.009	0.018

The results are sorted according to the MCC score.

group does not predict contacts directly from the sequence but relies on the submitted three-dimensional models. The two other groups that stand out in the PR-curve analysis are G087 and G072, both from the Distill family of methods (group leader G. Pollastri, University College Dublin).

The results of the PR-analysis (AUC_PR scores) are shown to be well correlated with the MCC and F_1 scores

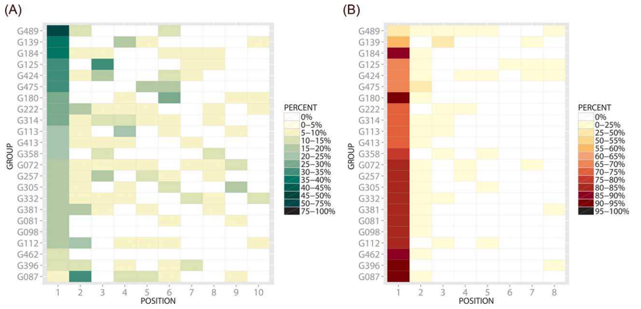


Figure 6
Percent of cases where the first correct (**A**) and first incorrect (**B**) prediction is in the reported position for each group. Rows are ordered according to the percentage in the first column of A. The data are shown for the top *L*/5 long-range contacts in FM domains.

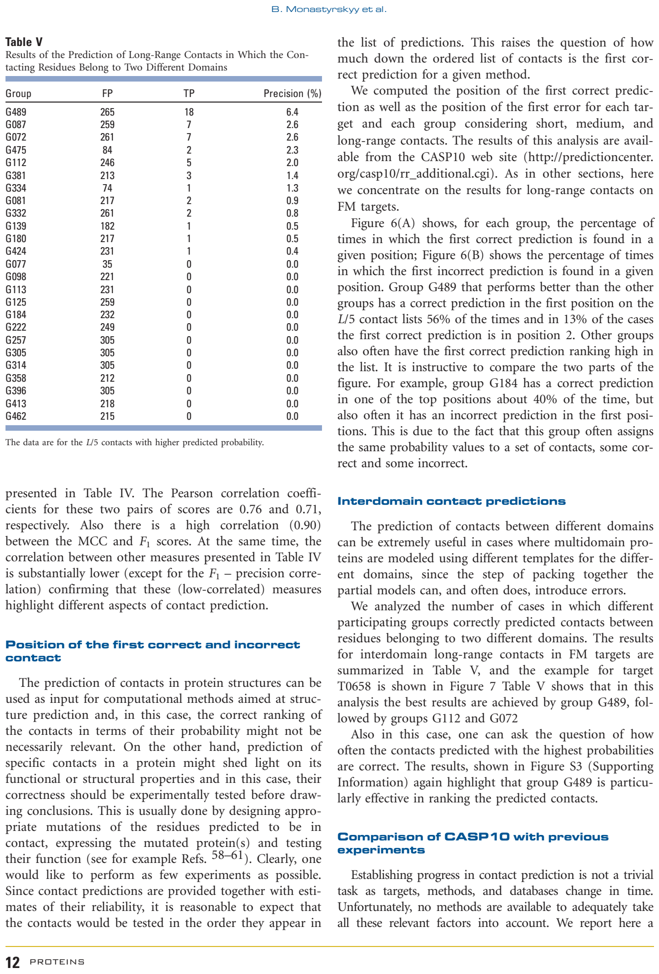


Table V
Results of the Prediction of Long-Range Contacts in Which the Contacting Residues Belong to Two Different Domains

Group	FP	TP	Precision (%)
G489	265	18	6.4
G087	259	7	2.6
G072	261	7	2.6
G475	84	2	2.3
G112	246	5	2.0
G381	213	3	1.4
G334	74	1	1.3
G081	217	2	0.9
G332	261	2	0.8
G139	182	1	0.5
G180	217	1	0.5
G424	231	1	0.4
G077	35	0	0.0
G098	221	0	0.0
G113	231	0	0.0
G125	259	0	0.0
G184	232	0	0.0
G222	249	0	0.0
G257	305	0	0.0
G305	305	0	0.0
G314	305	0	0.0
G358	212	0	0.0
G396	305	0	0.0
G413	218	0	0.0
G462	215	0	0.0

The data are for the *L*/5 contacts with higher predicted probability.

presented in Table IV. The Pearson correlation coefficients for these two pairs of scores are 0.76 and 0.71, respectively. Also there is a high correlation (0.90) between the MCC and F_1 scores. At the same time, the correlation between other measures presented in Table IV is substantially lower (except for the F_1 – precision correlation) confirming that these (low-correlated) measures highlight different aspects of contact prediction.

Position of the first correct and incorrect contact

The prediction of contacts in protein structures can be used as input for computational methods aimed at structure prediction and, in this case, the correct ranking of the contacts in terms of their probability might not be necessarily relevant. On the other hand, prediction of specific contacts in a protein might shed light on its functional or structural properties and in this case, their correctness should be experimentally tested before drawing conclusions. This is usually done by designing appropriate mutations of the residues predicted to be in contact, expressing the mutated protein(s) and testing their function (see for example Refs. [58–61]). Clearly, one would like to perform as few experiments as possible. Since contact predictions are provided together with estimates of their reliability, it is reasonable to expect that the contacts would be tested in the order they appear in the list of predictions. This raises the question of how much down the ordered list of contacts is the first correct prediction for a given method.

We computed the position of the first correct prediction as well as the position of the first error for each target and each group considering short, medium, and long-range contacts. The results of this analysis are available from the CASP10 web site (http://predictioncenter.org/casp10/rr_additional.cgi). As in other sections, here we concentrate on the results for long-range contacts on FM targets.

Figure 6(A) shows, for each group, the percentage of times in which the first correct prediction is found in a given position; Figure 6(B) shows the percentage of times in which the first incorrect prediction is found in a given position. Group G489 that performs better than the other groups has a correct prediction in the first position on the *L*/5 contact lists 56% of the times and in 13% of the cases the first correct prediction is in position 2. Other groups also often have the first correct prediction ranking high in the list. It is instructive to compare the two parts of the figure. For example, group G184 has a correct prediction in one of the top positions about 40% of the time, but also often it has an incorrect prediction in the first positions. This is due to the fact that this group often assigns the same probability values to a set of contacts, some correct and some incorrect.

Interdomain contact predictions

The prediction of contacts between different domains can be extremely useful in cases where multidomain proteins are modeled using different templates for the different domains, since the step of packing together the partial models can, and often does, introduce errors.

We analyzed the number of cases in which different participating groups correctly predicted contacts between residues belonging to two different domains. The results for interdomain long-range contacts in FM targets are summarized in Table V, and the example for target T0658 is shown in Figure 7 Table V shows that in this analysis the best results are achieved by group G489, followed by groups G112 and G072

Also in this case, one can ask the question of how often the contacts predicted with the highest probabilities are correct. The results, shown in Figure S3 (Supporting Information) again highlight that group G489 is particularly effective in ranking the predicted contacts.

Comparison of CASP10 with previous experiments

Establishing progress in contact prediction is not a trivial task as targets, methods, and databases change in time. Unfortunately, no methods are available to adequately take all these relevant factors into account. We report here a

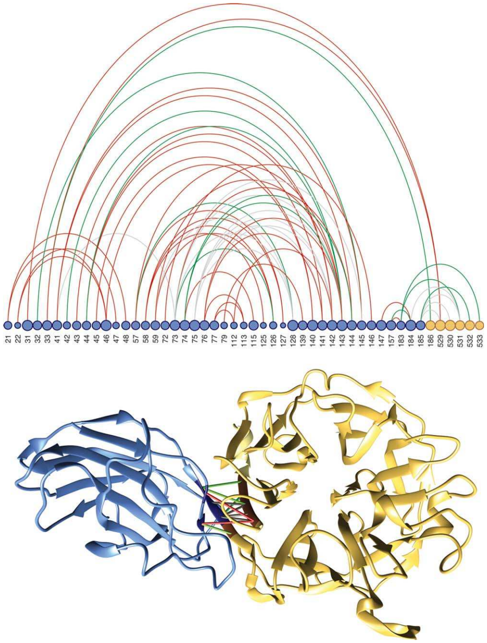


Figure 7

Example of the prediction of inter-domain contacts for target T0658. This is a two domain protein with the first domain (residues 20–185) being an FM target and the second (residues 186–540)—a template based target. The top panel shows *L*/5 contacts correctly predicted by at least one group as arcs connecting the corresponding residues indicated by circles. We show all the residues involved in correctly predicted contacts in the first (FM) domain, both intra- and inter-domain, and only the residues involved in correctly predicted inter-domain contacts for the second (TBM) domain. The size of the circle is proportional to the number of contacts the residue makes in the experimental structure. Blue and yellow circles are residues belonging to the first and second domain, respectively. The color of the connecting arcs indicates the frequency with which the corresponding contact was predicted by the groups. Red, green, and gray lines indicate contacts predicted with a frequency below the median, between the median and the third quartile and above the third quartile, respectively. The bottom figure shows the three-dimensional structure of the protein with the first domain in blue and the second in yellow. The correctly predicted contacts are indicated by sticks with the same color scheme as the corresponding arcs in the top panel.

comparison of the results without attempting to make any claim about the presence of real and measurable progress.

Figure 8 shows the results of the top 10 groups in the latest three CASPs on FM domains for the *L*/5 lists of long-range contacts (CASP10 results for the FM + TBM _hard domains are also included for comparison). On average, the CASP8 predictions (12 domains) have the highest precision—24.6%, followed by CASP9 (29

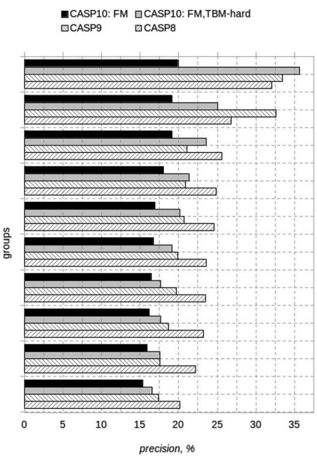


Figure 8

Precision of prediction for the top 10 groups in latest three CASPs.

domains)—21.4%, CASP10 (FM + TBM_hard, 28 domains)—21.4%, and CASP10: (FM, 16 domains)—17.4%. These results may indicate lack of substantial progress or, alternatively, be a consequence of the growing difficulty of targets in subsequent CASPs.[63]

CONCLUSIONS

The assessment of the state-of-the-art in contact prediction shows that the current precision of the best contact prediction methods on long-range contacts averages around 20%—the same limit observed in several previous CASPs. We look forward to seeing the results of the new methods that have recently appeared. Their published results in tests other than CASP have certainly stirred a lot of attention and it is therefore likely that we will see a renewed interest in the development of novel methods in contact prediction that will lead to improved results. We believe that progress in the field is objectively offset by the increased difficulty of the targets in CASP10 and that the depth of the alignments available for these targets made them less attractive for these new methods. At the same time, it should be mentioned that the list of CASP targets does mirror the proteins that the biological community considers interesting and worth an effort.

The predictions submitted by the best performing groups are statistically indistinguishable on the set of free-modeling domains. When hard template-based targets are added to the dataset, the results of the Multicom group, which uses consensus strategy to extract the contacts from predicted three-dimensional structures, are better than the others. Among the remaining groups, two implementations of the Distill method and ab initio predictors from the Multicom series of methods quite consistently perform better.

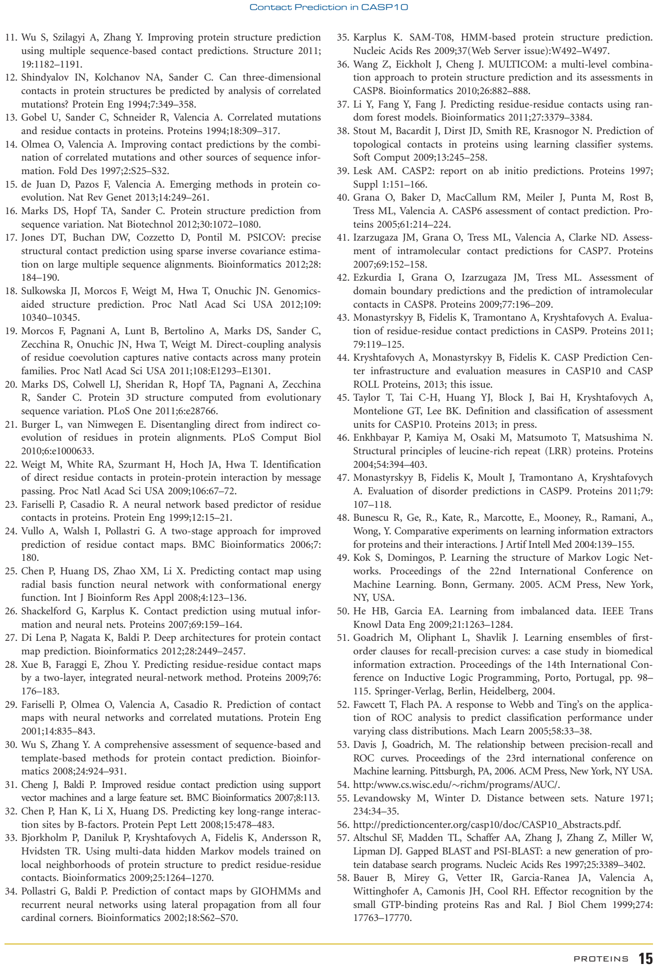
Based on the CASP10 data, we show that shorter domains are in general easier targets for contact prediction, and that the difficulty of predicting contacts in domains is not correlated with the depth of target sequence alignment.

ACKNOWLEDGMENTS

The authors thank John Moult and David Jones for useful suggestions.

REFERENCES

1. Havel TF, Crippen GM, Kuntz ID. Effects of distance constraints on macromolecular conformation. 2. Simulation of experimental results and theoretical predictions. Biopolymers 1979;18:73–81.
2. Brunger AT, Clore GM, Gronenborn AM, Karplus M. Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. Proc Natl Acad Sci USA 1986;83:3801–3805.
3. Clore GM, Nilges M, Brunger AT, Karplus M, Gronenborn AM. A comparison of the restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances. FEBS Lett 1987;213: 269–277.
4. Bohr J, Bohr H, Brunak S, Cotterill RM, Fredholm H, Lautrup B, Petersen SB. Protein structures from distance inequalities. J Mol Biol 1993;231:861–869.
5. Saitoh S, Nakai T, Nishikawa K. A geometrical constraint approach for reproducing the native backbone conformation of a protein. Proteins 1993;15:191–204.
6. Taylor WR. Protein fold refinement: building models from idealized folds using motif constraints and multiple sequence data. Protein Eng 1993;6:593–604.
7. Skolnick J, Kolinski A, Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 1997;265:217–241.
8. Vendruscolo M, Kussell E, Domany E. Recovery of protein structure from contact maps. Fold Des 1997;2:295–306.
9. Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. Bioinformatics 2008;24:1313–1315.
10. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.



11. Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure 2011; 19:1182–1191.
12. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
13. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
14. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
15. de Juan D, Pazos F, Valencia A. Emerging methods in protein coevolution. Nat Rev Genet 2013;14:249–261.
16. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–1080.
17. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28: 184–190.
18. Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. Genomics-aided structure prediction. Proc Natl Acad Sci USA 2012;109: 10340–10345.
19. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 2011;108:E1293–E1301.
20. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.
21. Burger L, van Nimwegen E. Disentangling direct from indirect coevolution of residues in protein alignments. PLoS Comput Biol 2010;6:e1000633.
22. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 2009;106:67–72.
23. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
24. Vullo A, Walsh I, Pollastri G. A two-stage approach for improved prediction of residue contact maps. BMC Bioinformatics 2006;7: 180.
25. Chen P, Huang DS, Zhao XM, Li X. Predicting contact map using radial basis function neural network with conformational energy function. Int J Bioinform Res Appl 2008;4:123–136.
26. Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins 2007;69:159–164.
27. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28:2449–2457.
28. Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins 2009;76: 176–183.
29. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
30. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 2008;24:924–931.
31. Cheng J, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. BMC Bioinformatics 2007;8:113.
32. Chen P, Han K, Li X, Huang DS. Predicting key long-range interaction sites by B-factors. Protein Pept Lett 2008;15:478–483.
33. Bjorkholm P, Daniluk P, Kryshtafovych A, Fidelis K, Andersson R, Hvidsten TR. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. Bioinformatics 2009;25:1264–1270.
34. Pollastri G, Baldi P. Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. Bioinformatics 2002;18:S62–S70.
35. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res 2009;37(Web Server issue):W492–W497.
36. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010;26:882–888.
37. Li Y, Fang Y, Fang J. Predicting residue-residue contacts using random forest models. Bioinformatics 2011;27:3379–3384.
38. Stout M, Bacardit J, Dirst JD, Smith RE, Krasnogor N. Prediction of topological contacts in proteins using learning classifier systems. Soft Comput 2009;13:245–258.
39. Lesk AM. CASP2: report on ab initio predictions. Proteins 1997; Suppl 1:151–166.
40. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61:214–224.
41. Izarzugaza JM, Grana O, Tress ML, Valencia A, Clarke ND. Assessment of intramolecular contact predictions for CASP7. Proteins 2007;69:152–158.
42. Ezkurdia I, Grana O, Izarzugaza JM, Tress ML. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 2009;77:196–209.
43. Monastyrskyy B, Fidelis K, Tramontano A, Kryshtafovych A. Evaluation of residue-residue contact predictions in CASP9. Proteins 2011; 79:119–125.
44. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL Proteins, 2013; this issue.
45. Taylor T, Tai C-H, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee BK. Definition and classification of assessment units for CASP10. Proteins 2013; in press.
46. Enkhbayar P, Kamiya M, Osaki M, Matsumoto T, Matsushima N. Structural principles of leucine-rich repeat (LRR) proteins. Proteins 2004;54:394–403.
47. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. Evaluation of disorder predictions in CASP9. Proteins 2011;79: 107–118.
48. Bunescu R, Ge, R., Kate, R., Marcotte, E., Mooney, R., Ramani, A., Wong, Y. Comparative experiments on learning information extractors for proteins and their interactions. J Artif Intell Med 2004:139–155.
49. Kok S, Domingos, P. Learning the structure of Markov Logic Networks. Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany. 2005. ACM Press, New York, NY, USA.
50. He HB, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–1284.
51. Goadrich M, Oliphant L, Shavlik J. Learning ensembles of first-order clauses for recall-precision curves: a case study in biomedical information extraction. Proceedings of the 14th International Conference on Inductive Logic Programming, Porto, Portugal, pp. 98–115. Springer-Verlag, Berlin, Heidelberg, 2004.
52. Fawcett T, Flach PA. A response to Webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. Mach Learn 2005;58:33–38.
53. Davis J, Goadrich, M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, PA, 2006. ACM Press, New York, NY USA.
54. http:/www.cs.wisc.edu/∼richm/programs/AUC/.
55. Levandowsky M, Winter D. Distance between sets. Nature 1971; 234:34–35.
56. http://predictioncenter.org/casp10/doc/CASP10_Abstracts.pdf.
57. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
58. Bauer B, Mirey G, Vetter IR, Garcia-Ranea JA, Valencia A, Wittinghofer A, Camonis JH, Cool RH. Effector recognition by the small GTP-binding proteins Ras and Ral. J Biol Chem 1999;274: 17763–17770.

59. Domanski TL, Halpert JR. Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. Curr Drug Metab 2001;2:117–137.
60. Thompson D, Lazennec C, Plateau P, Simonson T. Probing electrostatic interactions and ligand binding in aspartyl-tRNA synthetase through site-directed mutagenesis and computer simulations. Proteins 2008;71:1450–1460.
61. Fremgen SA, Burke NS, Hartzell PL. Effects of site-directed mutagenesis of mglA on motility and swarming of Myxococcus xanthus. BMC Microbiol 2010;10:295.
62. Kryshtafovych A, Moult J, Bales P, Bazan JF, Burgin A, Chen C, Cochran FV, Craig TK, Das R, Fass D, Garcia-Doval C, Herzberg O, Lorimer D, Luecke H, Ma X, Nelson D, van Raaij MJ, Rohwer F, Segall A, Seguritan V, Zeth K, Schwede T. Challenging the state-of-the-art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. Proteins 2013; this issue.
63. Kryshtafovych A, Fidelis K., Moult J. CASP10 results compared to those of previous CASP experiments, Proteins 2013; this issue.

References

[1] Mardis, E.R.: A decade's perspective on DNA sequencing technology. Nature **470**(7333) (February 2011) 198–203 1

[2] Levy, S., Sutton, G., Ng, P.C., et al.: The diploid genome sequence of an individual human. PLoS biology **5**(10) (September 2007) e254 1

[3] Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., et al.: Accurate whole human genome sequencing using reversible terminator chemistry. Nature **456**(7218) (November 2008) 53–9 1

[4] Wheeler, D.a., Srinivasan, M., Egholm, M., et al.: The complete genome of an individual by massively parallel DNA sequencing. Nature **452**(7189) (April 2008) 872–6 1

[5] Mardis, E.R.: The impact of next-generation sequencing technology on genetics. Trends in genetics : TIG **24**(3) (March 2008) 133–41 1

[6] Metzker, M.L.: Sequencing technologies - the next generation. Nature reviews. Genetics **11**(1) (January 2010) 31–46 1, 11

[7] Kircher, M., Heyn, P., Kelso, J.: Addressing challenges in the production and analysis of Illumina sequencing data. BMC Genomics **12**(1) (2011) 382 1, 2

[8] Hawkins, R.D., Hon, G.C., Ren, B.: Next-generation genomics: an integrative approach. Nature reviews. Genetics **11**(7) (July 2010) 476–86 2

[9] Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics **10**(1) (January 2009) 57–63 2, 11

[10] Asmann, Y.W., Klee, E.W., Thompson, E.A., et al.: 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. BMC genomics **10** (January 2009) 531 2

[11] Serteyn, D., Piquemal, D., Vanderheyden, L., et al.: Gene expression profiling from leukocytes of horses affected by osteochondrosis. Journal of orthopaedic research : official publication of the Orthopaedic Research Society **28**(7) (July 2010) 965–70 2

[12] 't Hoen, P.a.C., Ariyurek, Y., Thygesen, H.H., et al.: Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic acids research **36**(21) (December 2008) e141 2

[13] Bloom, J.S., Khan, Z., Kruglyak, L., Singh, M., Caudy, A.a.: Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC genomics **10** (January 2009) 221 2

[14] Maher, C.a., Kumar-Sinha, C., Cao, X., et al.: Transcriptome sequencing to detect gene fusions in cancer. Nature **458**(7234) (March 2009) 97–101 2

[15] Guttman, M., Amit, I., Garber, M., et al.: Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature **458**(7235) (March 2009) 223–7 2

[16] Lu, C., Meyers, B.C., Green, P.J.: Construction of small RNA cDNA libraries for deep sequencing. Methods (San Diego, Calif.) **43**(2) (October 2007) 110–7 2, 10

[17] Schmidt, D., Wilson, M.D., Spyrou, C., et al.: ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods (San Diego, Calif.) **48**(3) (July 2009) 240–8 2, 40

[18] Cullum, R., Alder, O., Hoodless, P.a.: The next generation: using new sequencing technologies to analyse gene regulation. Respirology (Carlton, Vic.) **16**(2) (February 2011) 210–22 2

[19] Dekker, J., Rippe, K., Dekker, M., Kleckner, N.: Capturing chromosome conformation. Science (New York, N.Y.) **295**(5558) (February 2002) 1306–11 2

[20] Gobeil, S., Zhu, X., Doillon, C.J., Green, M.R.: A genome-wide shRNA screen identifies GAS1 as a novel melanoma metastasis suppressor gene. Genes & development **22**(21) (November 2008) 2932–40 2

[21] Gazin, C., Wajapeyee, N., Gobeil, S., Virbasius, C.M., Green, M.R.: An elaborate pathway required for Ras-mediated epigenetic silencing. Nature **449**(7165) (October 2007) 1073–7 2

[22] Bric, A., Miething, C., Bialucha, C.U., et al.: NIH Public Access. Cancer cell **16**(4) (2010) 324–335 2

[23] Meacham, C.E., Ho, E.E., Dubrovsky, E., Gertler, F.B., Hemann, M.T.: In vivo RNAi screening identifies regulators of actin dynamics as key determinants of lymphoma progression. Nature genetics **41**(10) (October 2009) 1133–7 2

[24] Luo, J., Emanuele, M.J., Li, D., et al.: A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. Cell **137**(5) (May 2009) 835–48 2

[25] Nekrutenko, A., Taylor, J.: Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature Reviews Genetics **13**(9) (August 2012) 667–672 2

[26] Khatri, P., Sirota, M., Butte, A.J.: Ten Years of Pathway Analysis : Current Approaches and Outstanding Challenges. PLoS Computational Biology **8**(2) (2012) 3, 62

[27] Glazko, G.V., Emmert-Streib, F.: Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. Bioinformatics (Oxford, England) **25**(18) (September 2009) 2348–54 3, 62

[28] Martin, D., Brun, C., Remy, E., et al.: GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome biology **5**(12) (January 2004) R101 3

[29] Doniger, S.W., Salomonis, N., Dahlquist, K.D., et al.: MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome biology **4**(1) (January 2003) R7 3

[30] Castillo-Davis, C.I., Hartl, D.L.: GeneMerge–post-genomic analysis, data mining, and hypothesis testing. Bioinformatics (Oxford, England) **19**(7) (May 2003) 891–2 3

[31] Boyle, E.I., Weng, S., Gollub, J., et al.: GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics (Oxford, England) **20**(18) (December 2004) 3710–5 3

[32] Berriz, G.F., King, O.D., Bryant, B., Sander, C., Roth, F.P.: Characterizing gene sets with FuncAssociate. Bioinformatics (Oxford, England) **19**(18) (December 2003) 2502–4 3

[33] Beissbarth, T., Speed, T.P.: GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics (Oxford, England) **20**(9) (June 2004) 1464–5 3

[34] Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C., Krawetz, S.A.: Global functional profiling of gene expression. Genomics **81**(2) (February 2003) 98–104 3

[35] Huang, D.W., Sherman, B.T., Lempicki, R.a.: Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research **37**(1) (January 2009) 1–13 3, 62

[36] Khatri, P., Drghici, S.: Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics (Oxford, England) **21**(18) (September 2005) 3587–95 3

[37] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M.: KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research **40**(Database issue) (January 2012) D109–14 3, 62

[38] Hunter, S., Jones, P., Mitchell, A., et al.: InterPro in 2011: new developments in the family and domain prediction database. Nucleic acids research **40**(Database issue) (January 2012) D306–12 3, 62

[39] Blake, J.A., Dolan, M., Drabkin, H., et al.: Gene Ontology annotations and resources. Nucleic acids research **41**(Database issue) (January 2013) D530–5 3, 62

[40] Huang, D.W., Sherman, B.T., Lempicki, R.a.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols **4**(1) (January 2009) 44–57 3, 43, 62

[41] Reimand, J., Arak, T., Vilo, J.: g:Profiler–a web server for functional interpretation of gene lists (2011 update). Nucleic acids research **39**(Web Server issue) (July 2011) W307–15 3, 62

[42] Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J.: g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic acids research **35**(Web Server issue) (July 2007) W193–200 3, 62

[43] Eden, E., Navon, R., Steinfeld, I., Lipson, D., Yakhini, Z.: GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC bioinformatics **10** (January 2009) 48 3, 62

[44] Zeeberg, B.R., Qin, H., Narasimhan, S., et al.: High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). BMC bioinformatics **6**(1) (January 2005) 168 3, 62

[45] Medina, I., Carbonell, J., Pulido, L., et al.: Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic acids research **38**(Web Server issue) (July 2010) W210–3 3, 62

[46] Tabas-Madrid, D., Nogales-Cadenas, R., Pascual-Montano, A.: GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. Nucleic acids research **40**(Web Server issue) (July 2012) W478–83 3, 62

[47] Rhee, S.Y., Wood, V., Dolinski, K., Draghici, S.: Use and misuse of the gene ontology annotations. Nature reviews. Genetics **9**(7) (July 2008) 509–15 4

[48] Baumgartner, W.A., Cohen, K.B., Fox, L.M., Acquaah-Mensah, G., Hunter, L.: Manual curation is not sufficient for annotation of genomic databases. Bioinformatics (Oxford, England) **23**(13) (July 2007) i41–8 4

[49] Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., Ofran, Y.: Automatic prediction of protein function. Cellular and molecular life sciences : CMLS **60**(12) (December 2003) 2637–50 4

[50] Whisstock, J.C., Lesk, A.M.: Prediction of protein function from protein sequence and structure. Quarterly reviews of biophysics **36**(3) (August 2003) 307–40 4

[51] Skolnick, J., Fetrow, J.S.: From genes to protein structure and function: novel applications of computational approaches in the genomic era. Trends in biotechnology **18**(1) (January 2000) 34–9 4

[52] Radivojac, P., Clark, W.T., Oron, T.R., et al.: A large-scale evaluation of computational protein function prediction. Nature methods **10**(3) (January 2013) 4, 5, 73

[53] Liolios, K., Chen, I.M.A., Mavromatis, K., et al.: The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research **38**(Database issue) (January 2010) D346–54 4

[54] Friedberg, I.: Automated protein function predictionçthe genomic challenge. Briefings in bioinformatics **7**(3) (2006) 225–242 4, 5, 71

[55] Bork, P., Dandekar, T., Diaz-lazcoz, Y., et al.: Predicting Function : From Genes to Genomes and Back. Journal of molecular biology **283** (1998) 707–725 4

[56] Watson, J.D., Laskowski, R.A., Thornton, J.M.: Predicting protein function from sequence and structural data. Current Opinion in Structural Biology **15** (2005) 275–284 4

[57] Lee, D., Redfern, O., Orengo, C.: Predicting protein function from sequence and structure. Nature reviews. Molecular cell biology **8** (2007) 995–1005 4

[58] Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular systems biology **3** (January 2007) 88 4

[59] Punta, M., Ofran, Y.: The Rough Guide to In Silico Function Prediction , or How To Use Sequence and Structure Information To Predict Protein Function. PLoS computational biology **4**(10) (2008) 4

[60] Rentzsch, R., Orengo, C.A.: Protein function prediction the power of multiplicity. Trends in Biotechnology **27**(4) (2009) 210–219 4

[61] Altschul, S.F., Madden, T.L., Schäffer, A.A., et al.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25**(17) (September 1997) 3389–402 5

[62] Rost, B.: Protein structures sustain evolutionary drift. Folding & design **2**(3) (January 1997) S19–24 5, 71

[63] Hegyi, H., Gerstein, M.: The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. Journal of molecular biology **288**(1) (April 1999) 147–64 5

[64] Aloy, P., Oliva, B., Querol, E., Aviles, F.X., Russell, R.B.: Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. Protein science : a publication of the Protein Society **11**(5) (May 2002) 1101–16 5

[65] Brenner, S.E., Chothia, C., Hubbard, T.J., Murzin, A.G.: Understanding protein structure: using scop for fold interpretation. Methods in enzymology **266** (January 1996) 635–43 5, 71

[66] Tang, H., Oishi, N., Kaneko, S., Murakami, S.: Molecular functions and biological roles of hepatitis B virus x protein. Cancer science **97**(10) (October 2006) 977–83 6, 45

[67] D'Andrea, D., Grassi, L., Mazzapioda, M., Tramontano, A.: FIDEA: a server for the functional interpretation of differential expression analysis. Nucleic acids research **41**(Web Server issue) (July 2013) W84–8 6, 17, 63

[68] Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A., Kryshtafovych, A.: Evaluation of residue-residue contact prediction in CASP10. Proteins (June 2013) 7, 72, 73

[69] Park, P.J.: ChIP-seq: advantages and challenges of a maturing technology. Nature reviews. Genetics **10**(10) (October 2009) 669–80 10, 40

[70] Farnham, P.J.: Insights from genomic profiling of transcription factors. Nature reviews. Genetics **10**(9) (September 2009) 605–16 10

[71] Shendure, J., Ji, H.: Next-generation DNA sequencing. Nature biotechnology **26**(10) (October 2008) 1135–45 10

[72] Oshlack, A., Robinson, M.D., Young, M.D.: From RNA-seq reads to differential expression results. Genome biology **11**(12) (January 2010) 220 11

[73] Mortazavi, A., Williams, B.A., Mccue, K., Schaeffer, L., Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods **5**(7) (2008) 621–628 11, 21

[74] Wilhelm, B.T., Landry, J.R.: RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods (San Diego, Calif.) **48**(3) (July 2009) 249–57 11

[75] Mardis, E.R.: Next-generation sequencing platforms. Annual review of analytical chemistry (Palo Alto, Calif.) **6** (January 2013) 287–303 11

[76] Wang, K., Singh, D., Zeng, Z., et al.: MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic acids research **38**(18) (October 2010) e178 11

[77] Au, K.F., Jiang, H., Lin, L., Xing, Y., Wong, W.H.: Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic acids research **38**(14) (August 2010) 4570–8 11

[78] Wu, T.D., Nacu, S.: Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics (Oxford, England) **26**(7) (April 2010) 873–81 11, 34

[79] Guttman, M., Garber, M., Levin, J.Z., et al.: Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature biotechnology **28**(5) (May 2010) 503–10 12

[80] Griffith, M., Griffith, O.L., Mwenifumbo, J., et al.: Alternative expression analysis by RNA sequencing. Nature methods **7**(10) (October 2010) 843–7 12

[81] Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., Dewey, C.N.: RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics (Oxford, England) **26**(4) (February 2010) 493–500 12

[82] Nicolae, M., Mangul, S., Mndoiu, I.I., Zelikovsky, A.: Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms for molecular biology : AMB **6**(1) (January 2011) 9 12

[83] Katz, Y., Wang, E.T., Airoldi, E.M., Burge, C.B.: Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature methods **7**(12) (December 2010) 1009–15 12

[84] Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X.: DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics (Oxford, England) **26**(1) (January 2010) 136–8 12

[85] Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) **26**(1) (January 2010) 139–40 12, 63

[86] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome biology **11**(10) (January 2010) R106 12, 63

[87] Trapnell, C., Roberts, A., Goff, L., et al.: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols **7**(3) (2012) 562–578 12, 63

[88] Trapnell, C., Hendrickson, D.G., Sauvageau, M., et al.: Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature biotechnology **31**(1) (January 2013) 46–53 12, 63

[89] Dunham, I., Kundaje, A., Aldred, S.F., et al.: An integrated encyclopedia of DNA elements in the human genome. Nature **489**(7414) (September 2012) 57–74 12

[90] Trapnell, C., Pachter, L., Salzberg, S.L.: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) **25**(9) (May 2009) 1105–11 12

[91] Trapnell, C., Williams, B.a., Pertea, G., et al.: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology **28**(5) (May 2010) 511–5 12

[92] Fisher, R.A.: The design of experiments. Hafner Publishing Company, New York, New York, USA (1935) 12

[93] Amaral, P.P., Mattick, J.S.: Noncoding RNA in development. Mammalian genome : official journal of the International Mammalian Genome Society **19**(7-8) (August 2008) 454–92 14

[94] Qureshi, I.A., Mattick, J.S., Mehler, M.F.: Long non-coding RNAs in nervous system function and disease. Brain research **1338** (June 2010) 20–35 14

[95] Buckingham, M., Vincent, S.D.: Distinct and dynamic myogenic populations in the vertebrate embryo. Current opinion in genetics & development **19**(5) (October 2009) 444–53 14

[96] Cacchiarelli, D., Martone, J., Girardi, E., et al.: MicroRNAs involved in molecular circuitries relevant for the Duchenne muscular dystrophy pathogenesis are controlled by the dystrophin/nNOS pathway. Cell metabolism **12**(4) (October 2010) 341–51 14

[97] Cacchiarelli, D., Incitti, T., Martone, J., et al.: miR-31 modulates dystrophin expression: new implications for Duchenne muscular dystrophy therapy. EMBO reports **12**(2) (February 2011) 136–41 14

[98] Cesana, M., Cacchiarelli, D., Legnini, I., et al.: A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell **147**(2) (October 2011) 358–69 14

[99] Hubbard, T.: The Ensembl genome database project. Nucleic Acids Research **30**(1) (January 2002) 38–41 15, 69

[100] Flicek, P., Ahmed, I., Amode, M.R., et al.: Ensembl 2013. Nucleic acids research **41**(Database issue) (January 2013) D48–55 15

[101] Carthew, R.W., Sontheimer, E.J.: Origins and Mechanisms of miRNAs and siRNAs. Cell **136**(4) (February 2009) 642–55 21

[102] He, L., Hannon, G.J.: MicroRNAs: small RNAs with a big role in gene regulation. Nature reviews. Genetics **5**(7) (July 2004) 522–31 21

[103] Bartel, D.P.: MicroRNAs: target recognition and regulatory functions. Cell **136**(2) (January 2009) 215–33 21

[104] Krol, J., Loedige, I., Filipowicz, W.: The widespread regulation of microRNA biogenesis, function and decay. Nature Publishing Group **11**(9) (2010) 597–610 21

[105] Mullokandov, G., Baccarini, A., Ruzo, A., et al.: High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. Nature methods **9**(8) (August 2012) 840–6 21

[106] Selbach, M., Schwanhäusser, B., Thierfelder, N., et al.: Widespread changes in protein synthesis induced by microRNAs. Nature **455**(7209) (September 2008) 58–63 21

[107] Friedman, R.C., Farh, K.K.H., Burge, C.B., Bartel, D.P.: Most mammalian mRNAs are conserved targets of microRNAs. Genome research **19**(1) (January 2009) 92–105 21

[108] Davis, B.N., Hata, A.: Cell Communication and Signaling Regulation of MicroRNA Biogenesis : A miRiad of mechanisms. Cell Communication and Signaling **22** (2009) 1–22 21

[109] Pasquinelli, A.E.: MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nature reviews. Genetics **13**(4) (April 2012) 271–82 21

[110] Shin, C., Nam, J.W., Farh, K.K.H., et al.: Expanding the microRNA targeting code: functional sites with centered pairing. Molecular cell **38**(6) (June 2010) 789–802 21

[111] Reinhart, B.J., Slack, F.J., Basson, M., et al.: The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature **403**(6772) (February 2000) 901–6 21

[112] Slack, F.J., Basson, M., Liu, Z., et al.: The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. Molecular cell **5**(4) (April 2000) 659–69 21

[113] Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K., Slack, F.J.: The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. Genes & development **18**(2) (January 2004) 132–7 21

[114] Cloonan, N., Forrest, A.R.R., Kolle, G., et al.: Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature methods **5**(7) (July 2008) 613–9 21

[115] Cloonan, N., Grimmond, S.M.: Transcriptome content and dynamics at single-nucleotide resolution. Genome biology **9**(9) (January 2008) 234 21

[116] Cloonan, N., Wani, S., Xu, Q., et al.: MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome biology **12**(12) (January 2011) R126 21, 22

[117] Creighton, C.J., Reid, J.G., Gunaratne, P.H.: Expression profiling of microRNAs by deep sequencing. Briefings in Bioinformatics **10**(5) (2009) 490–497 22

[118] Landgraf, P., Rusu, M., Sheridan, R., et al.: A mammalian microRNA expression atlas based on small RNA library sequencing. Cell **129**(7) (June 2007) 1401–14 22

[119] Morin, R.D., O'Connor, M.D., Griffith, M., et al.: Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome research **18**(4) (April 2008) 610–21 22

[120] Kim, J., Cho, I.S., Hong, J.S., et al.: Identification and characterization of new microRNAs from pig. Mammalian genome : official journal of the International Mammalian Genome Society **19**(7-8) (August 2008) 570–80 22

[121] Reese, T.A., Xia, J., Johnson, L.S., et al.: Identification of novel microRNA-like molecules generated from herpesvirus and host tRNA transcripts. Journal of virology **84**(19) (October 2010) 10344–53 22

[122] de Hoon, M.J.L., Taft, R.J., Hashimoto, T., et al.: Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. Genome research **20**(2) (February 2010) 257–64 22

[123] Sdassi, N., Silveri, L., Laubier, J., et al.: Identification and characterization of new miRNAs cloned from normal mouse mammary gland. BMC genomics **10** (January 2009) 149 22

[124] Ruby, J.G., Stark, A., Johnston, W.K., et al.: Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome research **17**(12) (December 2007) 1850–64 22

[125] Ruby, J.G., Jan, C., Player, C., et al.: Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell **127**(6) (December 2006) 1193–207 22

[126] Cummins, J.M., He, Y., Leary, R.J., et al.: The colorectal microRNAome. Proceedings of the National Academy of Sciences of the United States of America **103**(10) (March 2006) 3687–92 22

[127] Ebhardt, H.A., Tsang, H.H., Dai, D.C., et al.: Meta-analysis of small RNA-sequencing errors

reveals ubiquitous post-transcriptional RNA modifications. Nucleic acids research **37**(8) (May 2009) 2461–70 22

[128] Pantano, L., Estivill, X., Martí, E.: SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. Nucleic acids research **38**(5) (March 2010) e34 22

[129] Lee, L.W., Zhang, S., Etheridge, A., et al.: Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA (New York, N.Y.) **16**(11) (November 2010) 2170–80 22, 34

[130] Burroughs, A.M., Ando, Y., de Hoon, M.J.L., et al.: A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. Genome research **20**(10) (October 2010) 1398–410 22

[131] Fernandez-Valverde, S.L., Taft, R.J., Mattick, J.S.: Dynamic isomiR regulation in Drosophila development. RNA (New York, N.Y.) **16**(10) (October 2010) 1881–8 22

[132] Neilsen, C.T., Goodall, G.J., Bracken, C.P.: IsomiRs–the overlooked repertoire in the dynamic microRNAome. Trends in genetics : TIG **28**(11) (November 2012) 544–9 22, 34

[133] Kozomara, A., Griffiths-Jones, S.: miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research **39**(Database issue) (January 2011) D152–7 22, 41

[134] Griffiths-Jones, S., Saini, H.K., van Dongen, S., Enright, A.J.: miRBase: tools for microRNA genomics. Nucleic acids research **36**(Database issue) (January 2008) D154–8 22, 41

[135] de Oliveira, L.F.V., Christoff, A.P., Margis, R.: isomiRID: a framework to identify microRNA isoforms. Bioinformatics (Oxford, England) **29**(20) (October 2013) 2521–3 24, 34

[136] Sablok, G., Milev, I., Minkov, G., et al.: isomiRex: web-based identification of microRNAs, isomiR variations and differential expression using next-generation sequencing datasets. FEBS letters **587**(16) (August 2013) 2629–34 24

[137] Humphreys, D.T., Suter, C.M.: miRspring : a compact standalone research tool for analyzing miRNA-seq data. Nucleic acids research (2013) 1–8 24

[138] Friedländer, M.R., Chen, W., Adamidi, C., et al.: Discovering microRNAs from deep sequencing data using miRDeep. Nature biotechnology **26**(4) (April 2008) 407–15 24

[139] He, M., Liu, Y., Wang, X., et al.: Cell-type-based analysis of microRNA profiles in the mouse brain. Neuron **73**(1) (January 2012) 35–48 25, 30

[140] Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., et al.: Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes & development **24**(10) (May 2010) 992–1009 25, 30, 34

[141] Kuchen, S., Resch, W., Yamane, A., et al.: Regulation of microRNA expression and abundance during lymphopoiesis. Immunity **32**(6) (June 2010) 828–39 25, 30

[142] Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America **101**(12) (March 2004) 4164–9 30, 32

[143] Azuma-Mukai, A., Oguri, H., Mituyama, T., et al.: Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. Proceedings of the National Academy of Sciences of the United States of America **105**(23) (June 2008) 7964–9 34

[144] Venter, J.C., Adams, M.D., Myers, E.W., et al.: The sequence of the human genome. Science (New York, N.Y.) **291**(5507) (February 2001) 1304–51 39

[145] Reik, W., Römer, I., Barton, S.C., et al.: Adult phenotype in the mouse can be affected by epigenetic events in the early embryo. Development (Cambridge, England) **119**(3) (November 1993) 933–42 39

[146] Ptashne, M., Gann, A.: Transcriptional activation by recruitment. Nature **386**(6625) (April 1997) 569–77 39

[147] Cheung, P., Lau, P.: Epigenetic regulation by histone methylation and histone variants. Molecular endocrinology (Baltimore, Md.) **19**(3) (March 2005) 563–73 39

[148] Pepke, S., Wold, B., Mortazavi, A.: Computation for chIP-seq and rNA-seq studies. Nature Methods **6**(11) (2009) S22–S32 39, 40

[149] Janitz, M.: Next-Generation Genome Sequencing: Towards Personalized Medicine. Wiley-Blackwell (2008) 39

[150] Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B.: Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, N.Y.) **316**(5830) (June 2007) 1497–502 39, 40

[151] Leleu, M., Lefebvre, G., Rougemont, J.: Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. Briefings in functional genomics **9**(5-6) (December 2010) 466–76 39

[152] Schones, D.E., Cui, K., Cuddapah, S., et al.: Dynamic regulation of nucleosome positioning in the human genome. Cell **132**(5) (March 2008) 887–98 39

[153] Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., et al.: The DNA-encoded nucleosome organization of a eukaryotic genome. Nature **458**(7236) (March 2009) 362–6 39

[154] Barski, A., Cuddapah, S., Cui, K., et al.: Resource High-Resolution Profiling of Histone Methylations in the Human Genome. Cell (2007) 823–837 39, 40

[155] Mikkelsen, T.S., Ku, M., Jaffe, D.B., et al.: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448**(7153) (August 2007) 553–60 39

[156] Wang, Z., Zang, C., Rosenfeld, J.A., et al.: Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics **40**(7) (July 2008) 897–903 39

[157] Schmidt, D., Wilson, M.D., Ballester, B., et al.: Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (New York, N.Y.) **328**(5981) (May 2010) 1036–40 39

[158] McDaniell, R., Lee, B.K., Song, L., et al.: Heritable individual-specific and allele-specific chromatin signatures in humans. Science (New York, N.Y.) **328**(5975) (April 2010) 235–9 39

[159] Preti, M., Ribeyre, C., Pascali, C., et al.: The telomere-binding protein Tbf1 demarcates snoRNA gene promoters in Saccharomyces cerevisiae. Molecular cell **38**(4) (May 2010) 614–20 39

[160] MacIsaac, K.D., Lo, K.A., Gordon, W., et al.: A quantitative model of transcriptional regulation reveals the influence of binding location on expression. PLoS computational biology **6**(4) (April 2010) e1000773 39

[161] Kuo, M.H., Allis, C.D.: In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. Methods (San Diego, Calif.) **19**(3) (November 1999) 425–33 39

[162] Mardis, E.R.: ChIP-seq : welcome to the new frontier Prion biology : the quest for the test. Nature Methods **4**(8) (2007) 613–614 40

[163] Trapnell, C., Salzberg, S.L.: How to map billions of short reads onto genomes. Nature biotechnology **27**(5) (May 2009) 455–7 40

[164] Zhang, Y., Liu, T., Meyer, C.a., et al.: Model-based analysis of ChIP-Seq (MACS). Genome biology **9**(9) (January 2008) R137 40

[165] Feng, J., Liu, T., Zhang, Y.: Using MACS to identify peaks from ChIP-Seq data. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 2** (June 2011) Unit 2.14 40

[166] Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S.: Identifying ChIP-seq enrichment using MACS. Nature Protocols **7**(9) (August 2012) 1728–1740 40

[167] Rozowsky, J., Euskirchen, G., Auerbach, R.K., et al.: PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature biotechnology **27**(1) (January 2009) 66–75 40

[168] Valouev, A., Johnson, D.S., Sundquist, A., et al.: Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature methods **5**(9) (September 2008) 829–34 40, 41

[169] Kharchenko, P.V., Tolstorukov, M.Y., Park, P.J.: Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature biotechnology **26**(12) (December 2008) 1351–9 40

[170] Laajala, T.D., Raghav, S., Tuomela, S., et al.: A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC genomics **10** (January 2009) 618 40

[171] Landt, S.G., Marinov, G.K., Kundaje, a., et al.: ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research **22**(9) (September 2012) 1813–1831 40

[172] Kent, W.J., Sugnet, C.W., Furey, T.S., et al.: The human genome browser at UCSC. Genome research **12**(6) (June 2002) 996–1006 40, 41

[173] Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., Li, L.: Application Note PAVIS : a tool for Peak Annotation and Visualization. Bioinformatics (2013) 4–5 41

[174] Zhu, L.J., Gazin, C., Lawson, N.D., et al.: ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC bioinformatics **11** (January 2010) 237 41

[175] McLean, C.Y., Bristor, D., Hiller, M., et al.: GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology **28**(5) (May 2010) 495–501 41, 43

[176] Harrow, J., Frankish, A., Gonzalez, J.M., et al.: GENCODE: the reference human genome annotation for The ENCODE Project. Genome research **22**(9) (September 2012) 1760–74 41

[177] Pruitt, K.D., Tatusova, T., Maglott, D.R.: NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **33**(Database issue) (January 2005) D501–4 41

[178] Ernst, J., Kheradpour, P., Mikkelsen, T.S., et al.: Mapping and analysis of chromatin state dynamics in nine human cell types. Nature **473**(7345) (May 2011) 43–9 41

[179] Bailey, T., Krajewski, P., Ladunga, I., et al.: Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Computational Biology **9**(11) (November 2013) e1003326 43

[180] Bailey, T.L., Williams, N., Misleh, C., Li, W.W.: MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research **34**(Web Server issue) (July 2006) W369–73 43

[181] Bailey, T.L., Boden, M., Buske, F.a., et al.: MEME SUITE: tools for motif discovery and searching. Nucleic acids research **37**(Web Server issue) (July 2009) W202–8 43

[182] Machanick, P., Bailey, T.L.: MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics (Oxford, England) **27**(12) (June 2011) 1696–7 43

[183] Thomas-Chollier, M., Herrmann, C., Defrance, M., et al.: RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic acids research **40**(4) (February 2012) e31 43

[184] Thomas-Chollier, M., Darbo, E., Herrmann, C., et al.: A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nature protocols **7**(8) (August 2012) 1551–68 43

[185] Mahony, S., Benos, P.V.: STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic acids research **35**(Web Server issue) (July 2007) W253–8 43

[186] Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S.: Quantifying similarity between motifs. Genome biology **8**(2) (January 2007) R24 43

[187] Andrisani, O.M., Barnabas, S.: The transcriptional function of the hepatitis B virus X protein and its role in hepatocarcinogenesis (Review). International journal of oncology **15**(2) (August 1999) 373–9 45

[188] Arbuthnot, P., Capovilla, A., Kew, M.: Putative role of hepatitis B virus X protein in hepatocarcinogenesis: effects on apoptosis, DNA repair, mitogen-activated protein kinase and JAK/STAT pathways. Journal of gastroenterology and hepatology **15**(4) (April 2000) 357–68 45

[189] Park, I.Y., Sohn, B.H., Yu, E., et al.: Aberrant epigenetic modifications in hepatocarcinogenesis induced by hepatitis B virus X protein. Gastroenterology **132**(4) (April 2007) 1476–94 45

[190] Karnoub, A.E., Weinberg, R.A.: Ras oncogenes: split personalities. Nature reviews. Molecular cell biology **9**(7) (July 2008) 517–31 46

[191] Varnholt, H.: The role of microRNAs in primary liver cancer. Annals of hepatology **7**(2) (2008) 104–13 46

[192] Wang, Y., Toh, H.C., Chow, P., et al.: MicroRNA-224 is up-regulated in hepatocellular carcinoma through epigenetic mechanisms.

FASEB journal : official publication of the Federation of American Societies for Experimental Biology **26**(7) (July 2012) 3032–41 46

[193] Ma, D., Tao, X., Gao, F., Fan, C., Wu, D.: miR-224 functions as an onco-miRNA in hepatocellular carcinoma cells by activating AKT signaling. Oncology letters **4**(3) (September 2012) 483–488 46

[194] Jin, X.L., Sun, Q.S., Liu, F., et al.: microRNA 21-mediated suppression of Sprouty1 by Pokemon affects liver cancer cell growth and proliferation. Journal of cellular biochemistry **114**(7) (July 2013) 1625–33 46

[195] Xu, G., Zhang, Y., Wei, J., et al.: MicroRNA-21 promotes hepatocellular carcinoma HepG2 cell proliferation through repression of mitogen-activated protein kinase-kinase 3. BMC cancer **13** (January 2013) 469 46

[196] Ruby, J.G., Jan, C.H., Bartel, D.P.: Intronic microRNA precursors that bypass Drosha processing. Nature **448**(July) (2007) 46

[197] Paige, C.J., Wu, G.E.: The B cell repertoire. FASEB journal : official publication of the Federation of American Societies for Experimental Biology **3**(7) (May 1989) 1818–24 50

[198] Holtappels, R.: Immunodominance and its significance in immunity. B.i.f. fut edn. Volume 20. (2005) 50

[199] Nara, P.L., Garrity, R.: Deceptive imprinting: a cosmopolitan strategy for complicating vaccination. Vaccine **16**(19) (November 1998) 1780–7 50, 51

[200] Stå lhammar Carlemalm, M., Waldemarsson, J., Johnsson, E., Areschoug, T., Lindahl, G.: Non-immunodominant regions are effective as building blocks in a streptococcal fusion protein vaccine. Cell host & microbe **2**(6) (December 2007) 427–34 51

[201] Khurana, S., Verma, N., Talaat, K.R., Karron, R.A., Golding, H.: Immune response following H1N1pdm09 vaccination: differences in antibody repertoire and avidity in young adults and elderly populations stratified by age and gender. The Journal of infectious diseases **205**(4) (February 2012) 610–20 51

[202] Khurana, S., Chearwae, W., Castellino, F., et al.: Vaccines with MF59 adjuvant expand the antibody repertoire to target protective sites of pandemic avian H5N1 influenza virus. Science translational medicine **2**(15) (January 2010) 15ra5 51

[203] Bagnoli, F., Baudner, B., Mishra, R.P.N., et al.: Designing the next generation of vaccines for global public health. Omics : a journal of integrative biology **15**(9) (September 2011) 545–66 51

[204] Sette, A., Rappuoli, R.: Reverse vaccinology: developing vaccines in the era of genomics. Immunity **33**(4) (October 2010) 530–41 51

[205] Jiang, N., He, J., Weinstein, J.A., et al.: Lineage structure of the human antibody repertoire in response to influenza vaccination. Science translational medicine **5**(171) (February 2013) 171ra19 51

[206] Mehr, R., Sternberg-Simon, M., Michaeli, M., Pickman, Y.: Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. Immunology letters **148**(1) (2012) 11–22 51

[207] Mathonet, P., Ullman, C.G.: The Application of Next Generation Sequencing to the Understanding of Antibody Repertoires. Frontiers in immunology **4** (January 2013) 265 51

[208] Busse, C.E., Czogiel, I., Braun, P., Arndt, P.F., Wardemann, H.: Single-cell based high-throughput sequencing of full-length immunoglobulin heavy and light chain genes. European journal of immunology (September 2013) 51

[209] Rowley, M.J., O'Connor, K., Wijeyewickrema, L.: Phage display for epitope determination: a paradigm for identifying receptor-ligand interactions. Biotechnology annual review **10** (January 2004) 151–88 51

[210] Felici, F., Galfrè, G., Luzzago, A., et al.: Phage-displayed peptides as tools for characterization of human sera. Methods in enzymology **267** (January 1996) 116–29 51

[211] Pizzi, E., Cortese, R., Tramontano, A.: Mapping epitopes on protein surfaces. Biopolymers **36**(5) (November 1995) 675–80 51

[212] Rockberg, J., Löfblom, J., Hjelm, B., Uhlén, M., Stå hl, S.: Epitope mapping of antibodies using bacterial surface display. Nature methods **5**(12) (December 2008) 1039–45 51

[213] Comanducci, M., Bambini, S., Brunelli, B., et al.: NadA, a novel vaccine candidate of Neisseria meningitidis. The Journal of experimental medicine **195**(11) (June 2002) 1445–54 52

[214] Thomas-Chollier, M., Watson, L.C., Cooper, S.B., et al.: A naturally occurring insertion of a single amino acid rewires transcriptional regulation by glucocorticoid receptor isoforms. Proceedings of the National Academy of Sciences of the United States of America **110**(44) (October 2013) 17826–31 63

[215] Natella, F., Leoni, G., Maldini, M., et al.: Absorbtion, Metabolisation And Effects At Transcriptome Level Of A Standardised French Oak Wood Extract Robuvit In Healthy Volunteers: A Pilot Study. Journal of agricultural and food chemistry (December 2013) 63

[216] Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics **29**(4) (August 2001) 1165–1188 65

[217] Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. The EMBO journal **5**(4) (April 1986) 823–6 72

[218] Sippl, M.J., Weitckus, S.: Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins **13**(3) (July 1992) 258–71 72

[219] Gribskov, M., McLachlan, A.D., Eisenberg, D.: Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences of the United States of America **84**(13) (July 1987) 4355–8 72

[220] Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D.: Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins **Suppl 3** (January 1999) 171–6 72

[221] Bayley, M.J., Jones, G., Willett, P., Williamson, M.P.: GENFOLD: a genetic algorithm for folding protein structures using NMR restraints. Protein science : a publication of the Protein Society **7**(2) (February 1998) 491–9 72

[222] Moult, J., Pedersen, J.T., Judson, R., Fidelis, K.: A large-scale experiment to assess protein structure prediction methods. Proteins **23**(3) (November 1995) ii–v 72

[223] Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K., Pedersen, J.T.: Critical assessment of methods of protein structure prediction (CASP): round II. Proteins **Suppl 1** (January 1997) 2–6 72

[224] Moult, J., Hubbard, T., Fidelis, K., Pedersen, J.T.: Critical assessment of methods of protein structure prediction (CASP): round III. Proteins **Suppl 3** (January 1999) 2–6 72

[225] Moult, J., Fidelis, K., Zemla, A., Hubbard, T.: Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins **Suppl 5** (January 2001) 2–7 72

[226] Moult, J., Fidelis, K., Zemla, A., Hubbard, T.: Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins **53 Suppl 6** (January 2003) 334–9 72

[227] Moult, J., Fidelis, K., Rost, B., Hubbard, T., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP)–round 6. Proteins **61 Suppl 7** (January 2005) 3–7 72

[228] Moult, J., Fidelis, K., Kryshtafovych, A., et al.: Critical assessment of methods of protein structure prediction-Round VII. Proteins **69 Suppl 8** (January 2007) 3–9 72

[229] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Tramontano, A.: Critical assessment of methods of protein structure prediction - Round VIII. Proteins **77 Suppl 9** (January 2009) 1–4 72

[230] Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP)–round IX. Proteins **79 Suppl 1** (January 2011) 1–5 72

[231] Janin, J., Henrick, K., Moult, J., et al.: CAPRI: a Critical Assessment of PRedicted Interactions. Proteins **52**(1) (July 2003) 2–9 73

[232] Reese, M.G., Hartzell, G., Harris, N.L., et al.: Genome annotation assessment in Drosophila melanogaster. Genome research **10**(4) (April 2000) 483–501 73

[233] Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC bioinformatics **6 Suppl 1** (January 2005) S1 73

[234] Levandowsky, M., Winter, D.: Distance between Sets. Nature **234** (1971) 34–35 74

[235] Bauer, B., Mirey, G., Vetter, I.R., et al.: Effector recognition by the small GTP-binding proteins Ras and Ral. The Journal of biological chemistry **274**(25) (June 1999) 17763–70 78

[236] Domanski, T.L., Halpert, J.R.: Analysis of mammalian cytochrome P450 structure and function by site-directed mutagenesis. Current drug metabolism **2**(2) (June 2001) 117–37 78

[237] Thompson, D., Lazennec, C., Plateau, P., Simonson, T.: Probing electrostatic interactions and ligand binding in aspartyl-tRNA synthetase through site-directed mutagenesis and computer simulations. Proteins **71**(3) (May 2008) 1450–60 78

[238] Fremgen, S.A., Burke, N.S., Hartzell, P.L.: Effects of site-directed mutagenesis of mglA on motility and swarming of Myxococcus xanthus. BMC microbiology **10** (January 2010) 295 78