University of Udine

Doctorate course in

Biomedical sciences and biotechnology

CYCLE XXVIII

Research doctorate thesis

Computational methods and pipelines for the analysis of next generation sequencing (NGS) data and pathway annotation.

Candidate:
Fabrizio Serra

Supervisors:
Professor Federico Fogolari
Professor Claudio Brancolini

ACADEMIC YEAR
2015/2016

TABLE OF CONTENTS

ABSTRACT

During the 1° year of my PhD I have worked on the evaluation of similarity measures commonly used in many bioinformatics applications. The increasing amount of data available in public database requires the development of tools for analysing them, so proper evaluation of similarity is becoming very important. Regarding the methodology that can be employed to evaluate proximity measures we pay attention to the concept of intrinsic separation ability, i.e. how well a distance discriminates objects belonging to different classes based on distance. The work I performed with Prof Fogolari was focused on finding the best similarity measure, and to compare known proximity measures versus the fraction enrichment proximity score (FES - developed by us) to assess the similarity among experiments and to identify genes that mostly contribute to similarity. During the 2° year, supervised by Prof.ssa Romualdi of Padua, I have generated a ChIP-seq data analysis pipeline exploiting the heterogeneity of different algorithms with the aim to extend graphite (bioconductor package) pathways annotation. Specifically, given a ChIP-seq result of a transcription factor, pathways annotation was expanded adding to the network the transcription factor (node) whose target genes were already annotated in the pathway. To this aim, ChIP-seq ENCODE datasets (important resource to improve pathway annotation) were used. During the 3° year, collaborating with Prof. Tell, I have worked on ChIP-seq and RIP-seq data analysis. First of all, I have generated a ChIP-seq data analysis pipeline with the aim to identify target genes directly regulated by APE1 during oxidative stress condition. The identification of target genes was performed by ChIP-seq analysis in order to identify APE1 preferential promoter binding sites. Then, by using RIP-seq data, I investigated the biological significance of the RNA bound by APE1 using several online tools. Gene Ontology analysis of biological functions was performed using DAVID online tool (https://david.ncifcrf.gov/). Interactions among proteins were identified by STRING (http://string-db.org). miRNA/mRNA targets were identified by data mining in miRGate (http://mirgate.bioinfo.cnio.es.). Another work I have been involved, aimed at studying the mef2a binding sites (common and exclusive) of GM12878 (lymphoid cell type) and K562 (myeloid cell type) cell lines by data mining process.

1 INTRODUCTION

1.1 Next generation sequencing (NGS)

The development of Next Generation Sequencing (NGS) technologies has modified the way to reason about scientific approaches in the research field [1]. NGS technologies (**Figure 1**) have permitted to perform experiments that previously were not technically possible as well as unfavorable by an economic point of view.

The human genome has been mapped in a lot of individuals; the challenge is to understand how errors in functioning lead to disease [2]. In recent years, the isolation and study of individual genes in a model system (traditional approaches), has been overcome by the generation of big data sets exploiting new high-throughput technologies. The integration of genomic, epigenomic, transcriptomic and proteomic datasets gives the possibility to answer many long standing questions [2]. In general, DNA-protein interactions may be detected with ChIP-seq, RNA-associated interaction network may be assessed by RIP-seq or CLIP-seq and the functional effects of constitutive and regulated splicing may be studied by RNA-seq [3].

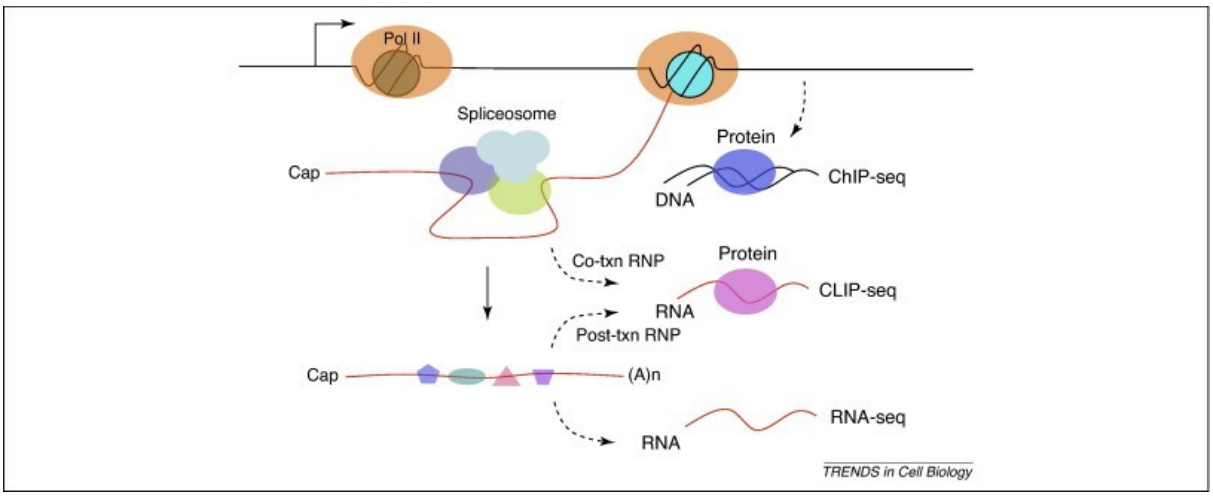


Figure 1: Next Generation Sequencing provides a wide range of applications [3].

1.2 Epigenomics and cancer

In 2003, after the completion of the Human Genome Project, it became clear that the information for life is encoded both in the DNA sequence and in the chemical modifications (**EPIGEN project –** www.supercomputing.it/project-description). These chemical modifications are deposited by enzymes on DNA and on its associated histone proteins (**EPIGEN project -**www.supercomputing.it/project-description). In eukaryotic cells chromatin is the combination of DNA and proteins, chromatin states may play a role in determining the transcription directly by altering the packing of the DNA. An eukaryotic organism contains cells having the same genome, each cell type is characterized by a certain epigenetic profile, there are hundreds of thousands of chromatin alterations that are inheritable but reversibile.

Genome-wide mapping of protein-DNA interactions, epigenetic marks and their modifications (that alter gene expression while the DNA remains unchanged), is required for a better comprehension of transcriptional regulation and cell differentiation [4]. A pattern altered by epigenetic modifications is crucial for a lot of common human diseases, including cancer During the years, the transcriptional silencing of tumour-suppressor genes by CpG-island-promoter hypermethylation has been explored in cancer cells [**5**]. Genetics and epigenetics cooperate at the stages of cancer development, this has been found thanks to the explosion of data resulting from the silencing of key regulatory genes [6].

1.3 ChIP-seq

The comprehension of transcriptional regulation mechanisms is of fundamental importance to have a clear and more complete overview of cell behaviour [7]. Protein-DNA interactions play vital roles, thus, identifying the interaction between transcription factors (TFs) and their binding DNA is essential to understand many biological processes (such as development, drug response and disease pathogenesis) [7]. Determining transcription factor binding sites (TFBSs) is arduous because the DNA segments recognized by TFs are often short and dispersed in the genome, and the target loci of a TF vary among tissues and physiological conditions. Chromatin immunoprecipitation (ChIP) followed by massively parallel sequencing (ChIP-Seq) is a new technology, the most widely used, to map protein-DNA interaction in genomes [8] and it is based on the enrichment of DNA associated with a protein of interest (**Figure 2**). Its use in studying histone modifications or nucleosomes has been essential in epigenetics research [9]. ChIP-seq data of genome-wide transcription factor binding site and chromatin modification provide precious infomation for studying gene regulation [10]. Compared with ChIP-chip assay (ChIP followed by microarray hybridization) this new technology (NGS) provides relatively high genomic coverage, high resolution, low noise and greater sensitivity and specificity while requiring a much smaller amount of starting material [11].

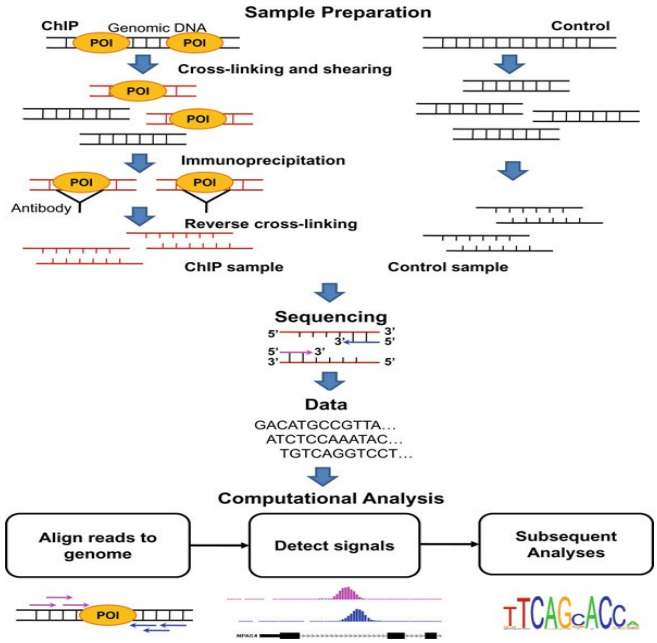


Figure 2: Chip-seq flowchart. POI is the protein of interest.

1.3.1 Immunoprecipitation with antibodies

In this technology (ChIP) cells are initially treated with a cross-linking agent (e.g. formaldehyde) that links DNA-interacting proteins to the DNA, then are lysed and the genomic DNA is isolated and sheared, usually by sonication, into a suitable fragment size distribution (200-600 bps, typically used for ChIP-seq) [12]. The protein and its bound chromatin fragments are immunoprecipitated using an antibody specific to the protein. The crosslinks are then reversed and the DNA fragment purified is assayed to determine the sequences bound by the protein (DNA sample called ChIP sample) (**Figure 3**). Usually a control sample is prepared in parallel using a similar protocol, an aliquot of sheared cell lysate is not immunoprecipitated but is processed normally (Input DNA). ChIP sample (compared to the input) is enriched in DNA fragments bound by the protein of interest.

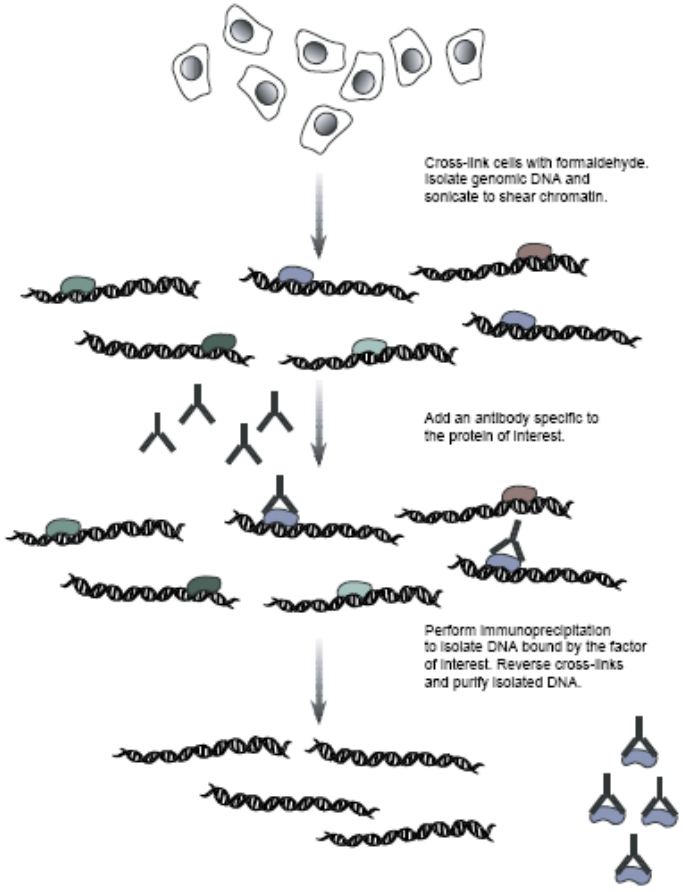


Figure 3: An overview of the chromatin immunoprecipitation (ChIP) procedure [12]

1.3.2 Amplification and Sequencing

The experimenter has to proceed through the following steps [13]:

Library preparation: Libraries may be constructed by a mixture of adaptor-flanked fragments up to a few hundred base-pairs in length. DNA fragments (from 75 to 300 bp fragments surrounding transcription factor binding sites or histone mark locations) are purified, ends repaired by adding A overhang and ligated by adapters.

Cluster generation: (Illumina/Solexa – Solid Phase Amplification – one DNA molecule per cluster). Amplified sequencing features are generated by bridge PCR, both forward and reverse PCR primers are attached to a solid substrate. All amplicons (deriving from any single template molecule during the amplification) stay immobilized and clustered to a single physical location. Each cluster obtained consist of around 1.000 clonal amplicons. It could be amplified million of clusters in different locations within each of eight independent lanes that are on a single flow-cell. The flow cell is a glass support (slide dimensions) that contains 8 lanes divided in 120 tile. The tiles are squares in which is possible to fix around 220.000 DNA molecole.

Sequencing:

The amplicons are single stranded and a primer is hybridized to a sequence flanking the region of interest. Each sequencing cycle (single-base extension with modified DNA polymerase) consists of the simultaneous addition of a mixture of four modified deoxynucleotide species, each one has one of four fluorescent labels and a reversible terminating moiety (3' hydroxyl position). Subsequently, single-base extension, acquisition of images (four channels) and chemical cleavage of both the fluorescent labels and the terminating moiety was performed. High-throughput sequencing often generates millions of 75 to 100 bp sequences (called short reads) [14] (**Figure 4**).

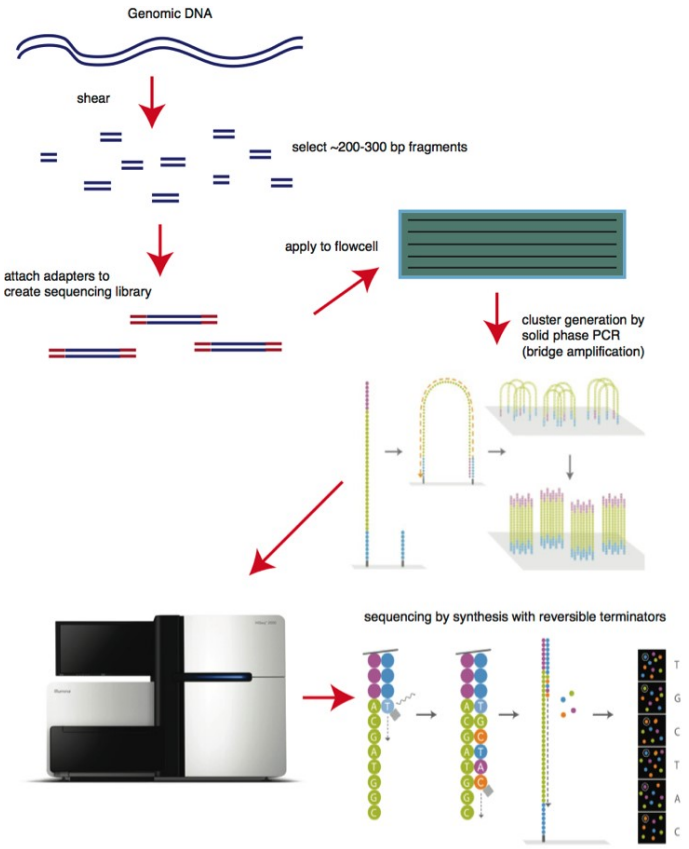


Figure 4: Library preparation, cluster generation and high throughput sequencing (**bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology**).

1.4 The analysis of ChIP-seq data

In this part, I describe the several steps involved in the computational analysis of ChIP-seq data. A flowchart of the central steps in the ChIP-seq procedure is shown in **Figure 5**.

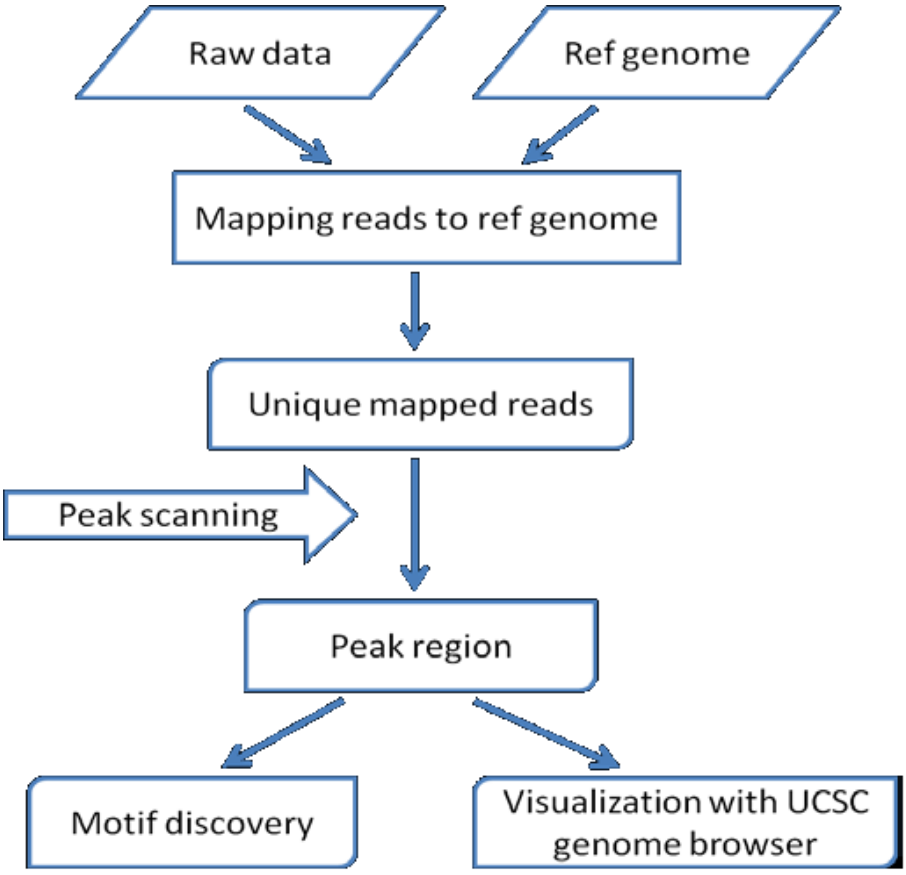


Figure 5: Flowchart of the ChIP-seq procedure. After mapping the raw sequence reads to a reference genome, the unique mapped reads are used and the significant enriched regions are detected from ChIP data compared to the control data [15].

ChIP-seq experiments generate huge amount of raw data (sequence reads) and effective computational analysis are crucial for uncovering biological mechanisms [11]. The short reads (35 bp), generated by NGS platforms, are acceptable for ChIP-seq, even if, for certain applications (e.g. de novo genome assembly), create serious difficulties [11]. These reads are short in length (around 25~30bp; the latest platform yields reads longer up to 50~100bp) and extreme high throughput (around 750MB to 1GB per lane). For mammalian transcription factors (TFs) and chromatin

modifications, which are on the order of thousands of binding sites, 20 million reads may be adequate [16]. Most histone marks (broader factors) and proteins with more binding sites (e.g. RNA Pol II) will need more reads, up to 60 million for mammalian ChIP-seq [17]. Reads should be filtered by applying a quality cutoff, trimmed to avoid lower quality bases and then mapped to the reference genome, the uniquely mapped reads are retine [18]. The distribution of reads shows separate peaks of read density on positive and negative strands [19] (**Figure 6**). The next step will be to identify regions that are significantly enriched in the ChIP sample when compared to the control, peak calling generates a list of enriched regions (peaks) that will be annotated by associating the functionally relevant genomic regions, such as gene promoters, transcription start sites, intergenic regions, etc [18].

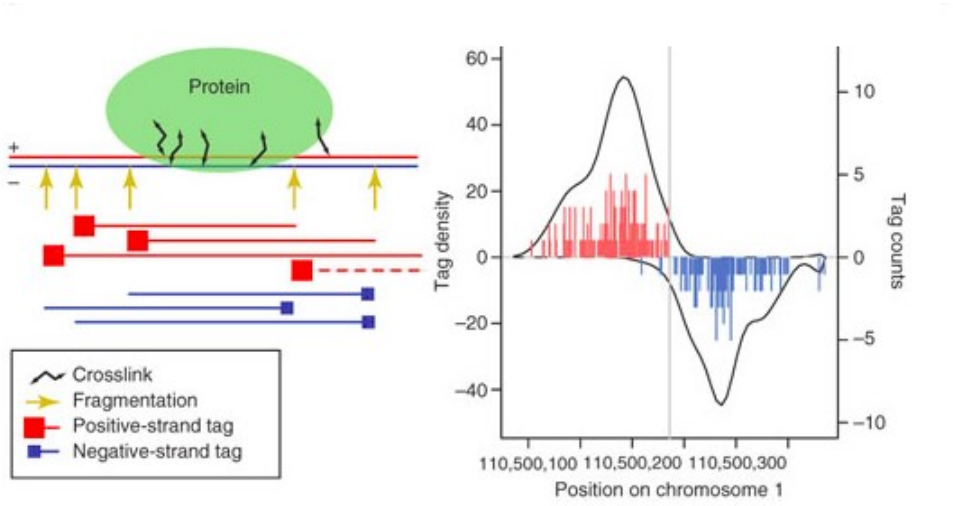


Figure 6: Fragments cross-linked to the protein of interest and read distribution around a stable binding position [19].

1.5 Human apurinic/apyrimidinic endonuclease 1 (APE1)

The human apurinic/apyrimidinic endonuclease 1 (APE1) is a crucial DNA repair protein that acts as central enzyme in BER pathway and as regulator of the intracellular redox homeostasis and redox transcriptional coactivator [20]. APE1 acquires a vital role in the maintenance of genome stability by removing the damaged base cleaving the abasic site to allow repair [21]. As master regulator of cellular response to oxidative stress, by using direct and indirect mechanisms, it acts by regulating the expression of genes involved in the inflammatory process and chemoresistance through transcriptional and post-transcriptional mechanisms. An overlooked mechanism through which APE1 can regulate gene expression is through directly binding and processing specific mRNA [22]. It has also been demonstrated that APE1 stimulates DNA binding of transcription factors involved in cancer promotion and progression [23]. APE1 is an emerging target for combination therapy of different cancers (i.e. ovarian, prostate, neurologic, hepatic, breast, non-small cell lung cancer) [24].

1.6 RNA-IP (RIP)-seq

Physical and functional interactions are important information for functional genomics and biology in general [2]. A great interest in RNA-protein interactions is booming, not just in long-standing processes such as transcription, splicing and translation, but also in fields such as gene regulation by RNA interference and non-coding RNAs (www.abcam.com). Interestingly, these aspects are currently contaminating also the field of DNA repair, in which an emerging role of RNA metabolism is establishing new paradigms. RNAs in cells are primarily associated with RNA-binding proteins (RNABPs) to constitute ribonucleoprotein complexes but RNA-protein interaction drives all the steps of RNA metabolism including: quality control, editing and degradation. The identification of sites where RNA-binding proteins interact with target RNAs using Next-generation sequencing (NGS) technologies opens new channels to understanding the vast complexity of RNA regulation [25]. Initially, to study RNA-protein complexes in their cellular environment, the immunoprecipitation combined with microarray analysis (RIP-CHIP), was employed; subsequently, to increase the specificity and positional resolution, a new strategy called RIP-seq was introduced [26]. RNA immunoprecipitation (RIP) is an antibody-based technique, the immunoprecipitation with an antibody toward a protein of interest enriches for RNA molecules actively bound to the target protein (www.epigenie.com). In the past, RIP studies used an immunoprecipitation without an agent to cross-link RNA-binding proteins to the RNA but there were a lot of problems to solve. Now, to improve the stability of the interactions, cells or tissues can be treated with formaldehyde that generates protein-RNA cross-links (www.epigenie.com). RNA combined with high-throughput sequencing (RIP-seq) is a strategy to produce transcriptome-wide maps of RNA binding with high accuracy and resolution, it is similar in principle to ChIP-seq, the main difference is that the RNA targets of the protein correspond to processed regions of the genome with different levels of expression [27]. At the present time, NGS technologies applied to IP samples do not allow to sequence the entire RNA or DNA molecules obtained from the IP, but only a short fragment at either or both 5' ends [28]. High-throughput sequencing using ILLUMINA/Solexa platform (it has

become the method of choice for this kind of experiments), produces millions of sequence reads whose length does not exceed 75-100 bps [28]. Bioinformatic expertise is required to reconstruct, from the short reads, which were the original RNA or DNA sequences bound by the protein and, also, whether the respective abundance estimated is due to a significant enrichment.

2 AIM OF THE THESIS

The work I performed during the PhD program was focused on the generation of robust computational pipelines useful for Next Generation Sequencing (NGS) data analysis in order to manage a large amount of biological data, coming from both data mining and experiments that exploit several technologies (ChIP-seq, RIP-seq, RNA-seq, and so on), aimed at obtaining reliable results. Comparing experiments from different technologies prompted refinement and testing of non-conventional similarity measures based on ranking the data. The last part of my work also dealt with data interpretation in the biological context.

3 MATERIALS AND METHODS

3.1 Downloading of ChIP-seq data from the ENCODE site

The ChIP-seq data of transcription factors I used are publicly available in the ENCyclopedia Of DNA Elements (ENCODE) site (https://genome.ucsc.edu/ENCODE/). The (ENCODE) project is an international research consortium, aimed at identifying all functional elements in the human genome sequence (https://www.encodeproject.org/). In 2007 researchers, after a pilot phase, started a second phase of technology development closed in September 2012 with the generation of high throughput data on functional elements (**Figure 1**), signaled by several publications [29]. They noticed that 80.4% of the human genome displays some functionality in at least one cell type revealed by the production of 1640 datasets focusing on 24 standard types of experiment within 147 different cell types. The consortium aims to annotate, using a correctly-annotated gene reference, all evidence-based gene features, including all protein-coding loci with alternatively transcribed variants, pseudogenes and non-coding loci with transcript evidence, in the whole human genome using a combination of manual annotation, computational analysis and experimental validation.

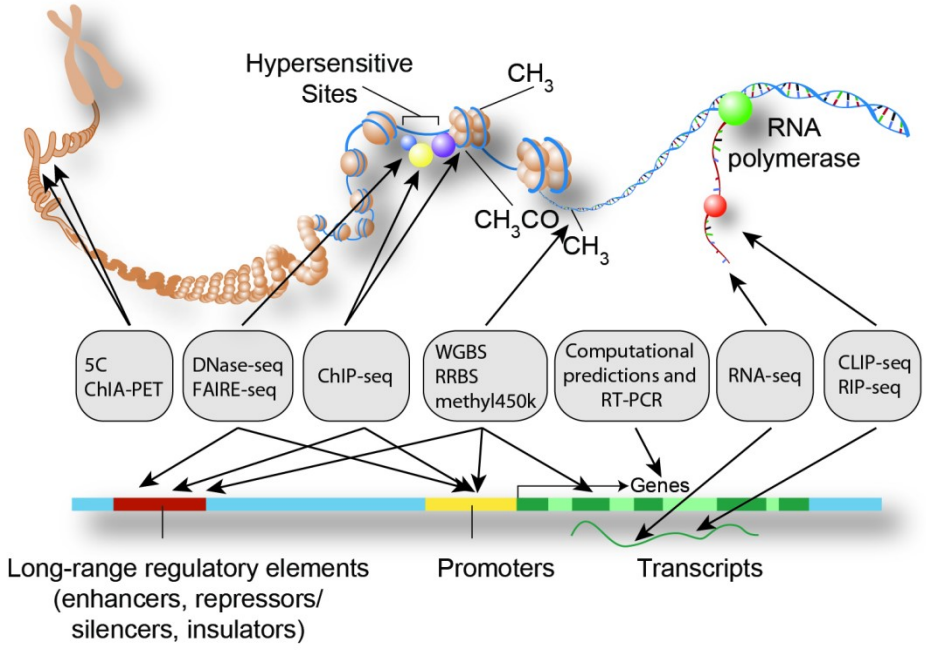


Figure 1: ENCODE provides a wide range of data types (https://www.encodeproject.org/).

3.2 Quality control using FASTQC as tool

Before starting with the computational analysis, it is important to assess the quality of the raw sequence data to identify possible sequencing errors or biases. FastQC is a freely available tool (www.bioinformatics.babraham.ac.uk/projects/fastqc) that provides a simple way to do some quality control checks on these raw sequence data through a modular set of analyses. It supports a lot of file formats (e.g. BAM, SAM). The QC report of results (summary graphs and tables) provides an overview of the areas where it's possible to find errors originated both in the sequencer and in the starting library material. In the upper part of the HTML report there is a summary of the several modules run and a brief evaluation of the results for each module that could be entirely normal (green tick), slightly abnormal (orange triangle) or unusual (red cross) (**Figure 2**) based on individual base frequency and overall sequence randomness and diversity.

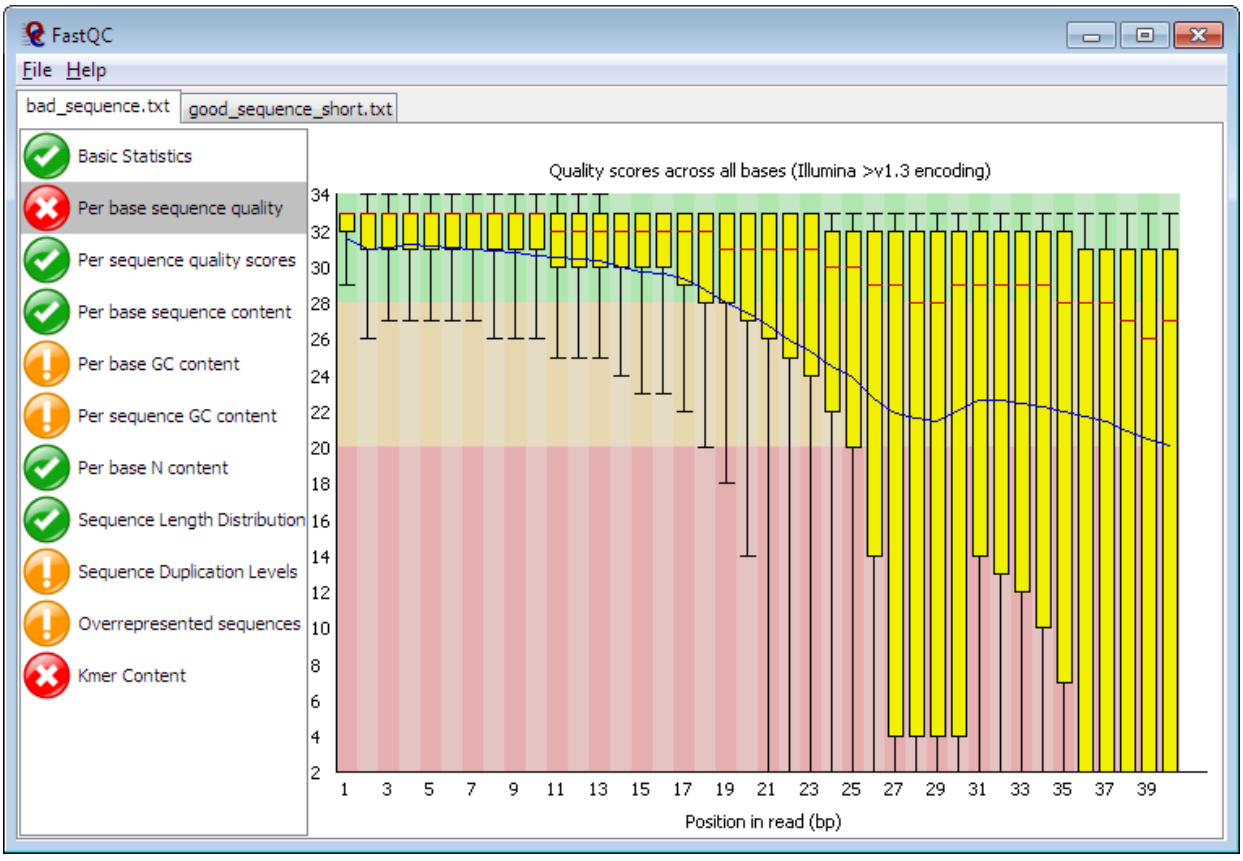


Figure 2: A quick overview to see in which areas there may be problems (www.bioinformatics.babraham.ac.uk/projects/fastqc).

3.3 Trimming using TRIMMOMATIC as tool

Technical sequences (known sequences coming from the technology such as adapters) should be eliminated during the library preparation but they usually can be found in NGS data even doing an optimal downstream analysis [30]. Trimmomatic is an efficient preprocessing tool , suitable to eliminate adaptors, that combines flexibility, correct handling of paired-end data and high performance. It is designed to work on paired-end data and it has been optimized for Illumina NGS data. This tool uses two processing steps to detect technical sequences within the reads:

1. **Simple mode** (**Figure 3**) – It works by finding a minimum overlap between the reads and the technical sequences to avoid false-positive findings. Exploiting the local alignment it aligns each read against each technical sequence.

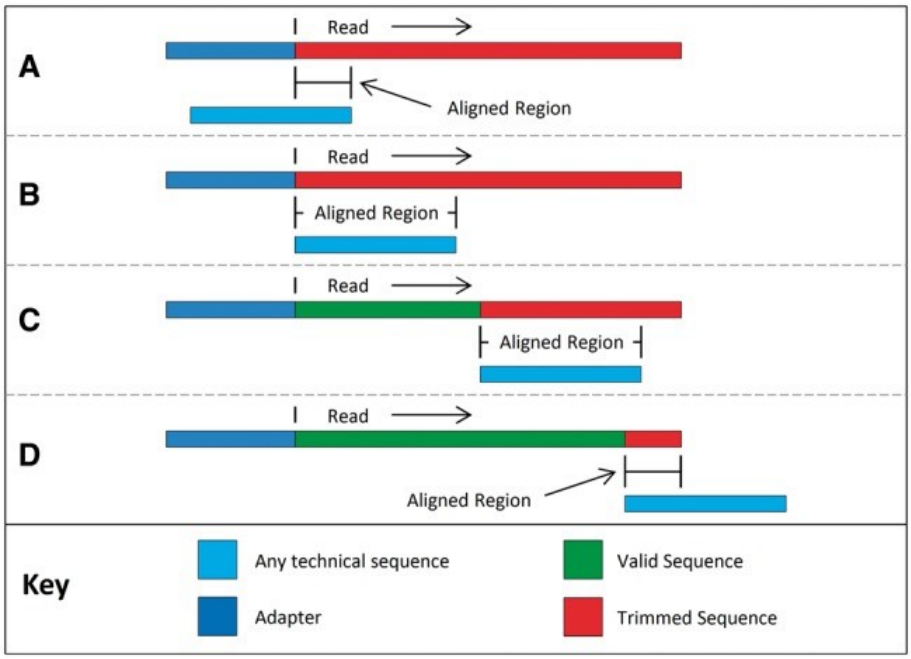


Figure 3: Simple mode putative sequence alignments [30]

2. **Palindrome mode** (**Figure 4**) – Exploiting the global alignment (total alignment score of the overlapping region), it aligns the forward and reverse reads, associated to their adapter sequences. Palindrome mode has some advantages in specificity and sensitivity respect to the simple mode but it may only be used with paired end reads.

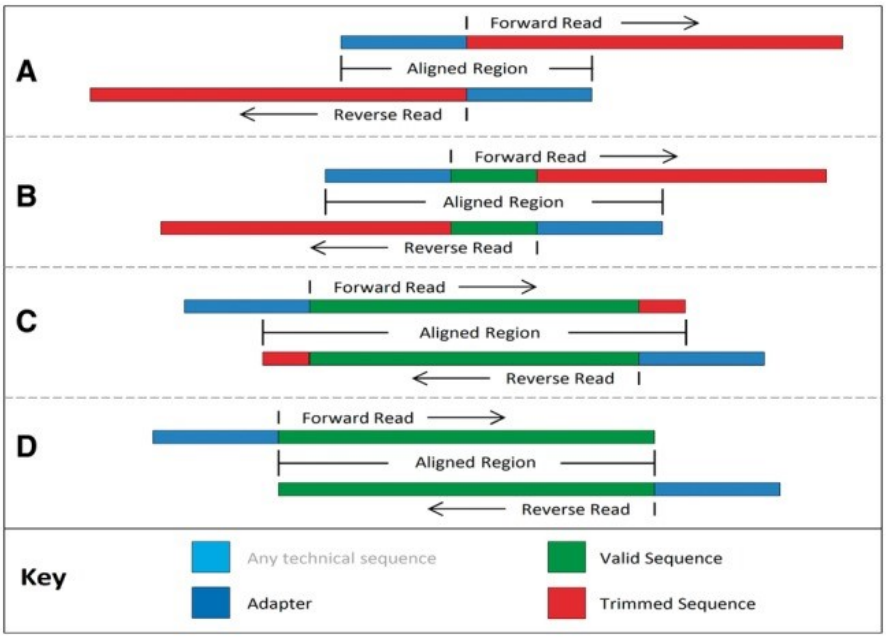


Figure 4: Palindrome mode putative sequence alignments [30].

3.4 Alignment of sequence reads

Alignment of reads to a reference genome, using one of the availaible mappers such as Bowtie [31], Bowtie 2 [32], BWA[33], should allow for 2~3 mismatches, due to sequencing errors [15]. The percentage of uniquely mapped reads obtained by the mapper varies between organisms, and for human ChIP-seq data, above 70% is normal, whereas less than 50% might be cause for concern [18]. The problem of having a low percentage of uniquely mapped reads often is due either to too much of amplification in the PCR step, not sufficient read length, or technical issues with the sequencing platform, but if the protein binds frequently in regions of repeated DNA it may be impossibile to avoid [18]. Employing paired-end sequencing, to reduce the

mapping ambiguity, can help. Multi-mapping reads will be filtered out by most peak-calling algorithms, even if they can lead the discovery of novel binding sites [34]. After mapping, the assessment (strand cross-correlation [16] or IP enrichment estimation) of the signal-to-noise ratio (SNR), will detect several possible failure modes of ChIP-seq experiment (insufficient enrichment by immunoprecipitation step, poor fragment-size selection, insufficient sequencing depth) [18]. Some peak-calling algorithms, such as SPP and MACS, perform the strand cross-correlation analysis.

3.5 Identification of enriched regions (peak shift estimation and peak detection)

The identification of regions, that are significantly enriched in the ChIP sample respect to the control, is the following step [11]. By using mapped reads (enriched near TFBSs), a model may be built to detect peaks, the tags (forward and reverse strand) are shifted 1⁄2 of fragment size, right or left [15]. The peaks can be found after modelling the shift size of reads and detecting the significant enriched DNA regions [15].

3.5.1 Peak shift estimation

The alignment of the reads (75-100 bps) to the immunoprecipitated DNA fragment (200-600 bps, typically used for ChIP-seq) generates two peaks (one on each strand) that flank the binding location of the protein of interest (this is due to the limited reads size). Using these peaks, the binding site can then be interpolated (**Figure 5**).

The algorithm needs a reference genome size (gsize), a sonication size (bandwidth) and a high-confidence fold-enrichment (mfold) that works by supposing totally N uniquely mapped reads are obtained in a ChIP sample [15].

All the reads are shifted by d/2 (“d” is the distance between the summits of the two strand peaks (forward and reverse strand)) toward the 3′ ends to the mainly expected transcription factor binding sites [15].

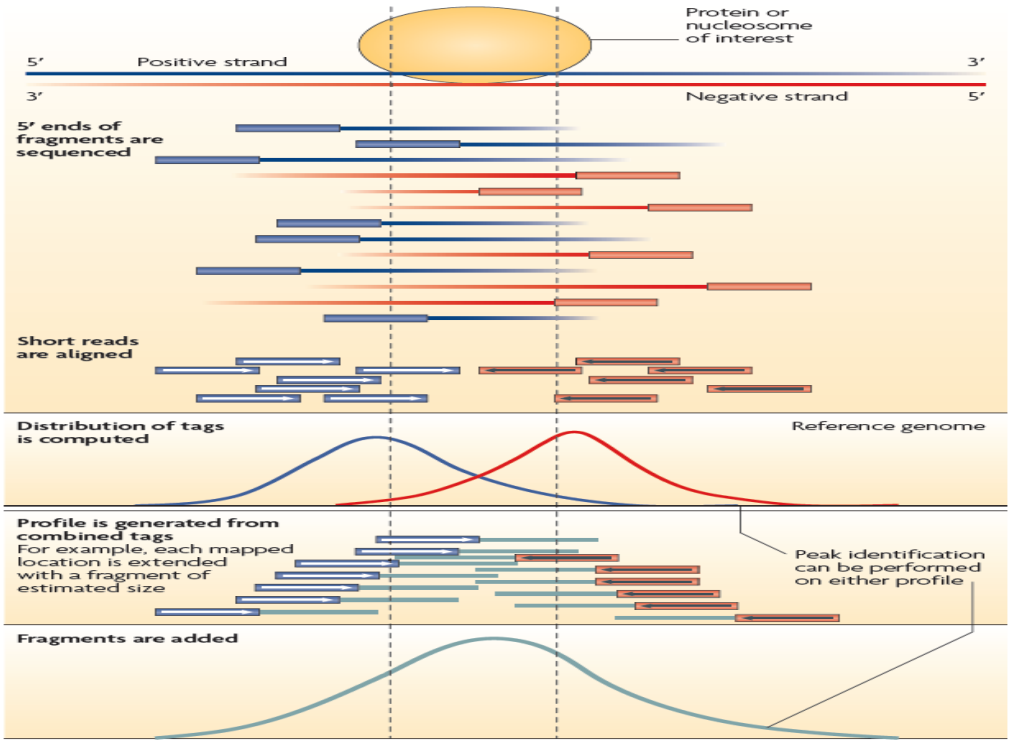


Figure 5: Sequencing of reads is possible from both ends of DNA fragments. Some sequenced short reads aligned with forward strands and other with reverse strands generate two peaks (upstream and downstream of TFBS). A new peak with potential binding sites is obtained by shifting all reads by d/2 ("d" is the distance between the summits of the forward and reverse peaks) [11].

3.5.2 Peak detection

A crucial analysis for ChIP-seq is to predict the regions of the genome where the protein is bound by finding regions significantly enriched of mapped reads (peaks) [18]. Sensitivity and specificity can greatly affect the number and quality of the peaks called, therefore they have to be considered when choosing an appropriate peak-calling algorithm. To improve specificity, duplicate reads should be removed before peak calling. The enrichment metrics (p-value or false discovery rate (FDR)), for some peak callers, might be affected by the sequencing depth and the statistical model used [18]. Therefore, employing the same p-value or False Discovery Rate (FDR) threshold does not guarantee that the number of peaks called are equal across libraries and different peak callers [35]. Some peak callers, such as MACS [36], support both single and paired-end reads. In an interesting paper, **Benjamini and**

Hochberg (1995), introduced the concept of false discovery rate (FDR) as a new point of view on the problem of multiplicity. From this new point of view, a desirable error rate to control can be the expected relative amount of errors among the rejected hypotheses.

Below I give a brief explanation of the typical peak callers used for the ChIP-seq data analysis.

3.5.2.1 SPP peak calling algorithm

The idea behind SPP method is described in the paper of Kharchenko et al, 2008 [19]**.** For the analysis of ChIP-seq data there are several specifically designed R packages tools, SPP is one of those. It exploits the cross-correlation profile to determine binding peak separation distance, and to evaluate if inclusion of tags having non-perfect alignment quality improves the cross correlation peak (**compbio.med.harvard.edu/Supplements/ChIP-seq/tutorial.html**).

SPP peak caller was specifically designed to detect protein-binding positions with high accuracy estimating the false discovery rate (FDR) **(Benjamini and Hochberg (1995))** as the number of binding positions found in the ChIP dataset, divided by that in the control set [37]. The method used, called window tag density (WTD), extends positive and negative strand tags by the required DNA fragment length to define binding positions to those tags having a great number of overlapping fragments, and score positions determined by strand-specific tags [37].

3.5.2.2 MACS2 peak calling algorithm

MACS peak caller [36] was designed to identify read-enriched regions from ChIP-seq data, it can be easily used for ChIP-seq data alone, or with control sample increasing the specificity [38]. MACS2 (update version of MACS) is specifically designed to process mixed signal types (https://github.com/taoliu/MACS). This computational method:

1. Removes redundant reads derived from the overamplification of ChIP-DNA by PCR yielding more reliable peak calls;

2. Adjusts read position based on fragment size distribution;
3. Calculates peak enrichment using local background normalization.

3.5.3 Assessment of reproducibility

At least two biological replicates of each ChIP-seq experiment are required to assess the reproducibility of experimental results [39]. It is important to check the reproducibility of both the reads, measured by computing the Pearson Correlation Coefficient (PCC) of the read counts at each genomic position [40] and identify peaks using the Irreproducible Discovery Rate (IDR) analysis [41] (threshold IDR=0.05) to assess the rank consistency between replicates. To use the Irreproducible Discovery Rate (IDR) rather than the False Discovery Rate (FDR) or the p-value (enrichment-based metric) makes the numbers of peaks more comparable across experiments [16]. For unrelated samples the range of PCC is usually 0.3-0.4, instead for replicate samples in high-quality experiments >0.9.

Low values of PCC suggest one of both replicates can be of low quality [18]. Before computing the PCC it is very important to remove the artefact regions having high ChIP signals.

3.6 Visualization of sequence reads and signals using IGV as tool

Visualization of sequence reads and signals is an essential component of genomic data analysis. Integrative Genomics Viewer (IGV) [42] is a freely available desktop tool, written in Java programming language (runs on Windows, Mac and Linux), for the visualization and real-time exploration of genomic datasets aligned to the selected reference genome **(Figure 7).** IGV is designed to be accessible to bioinformaticians and bench biologist as well. It supports simultaneous viewing of multiple datasets (same or different types of data) and several genomic file formats:

1. Nonindexed formats - GFF [43], BED [44] and WIG [45];
2. Indexed formats - BAM (sequence alignments);
3. Multiresolution formats - TDF, bigWig and BigBed formats [46].

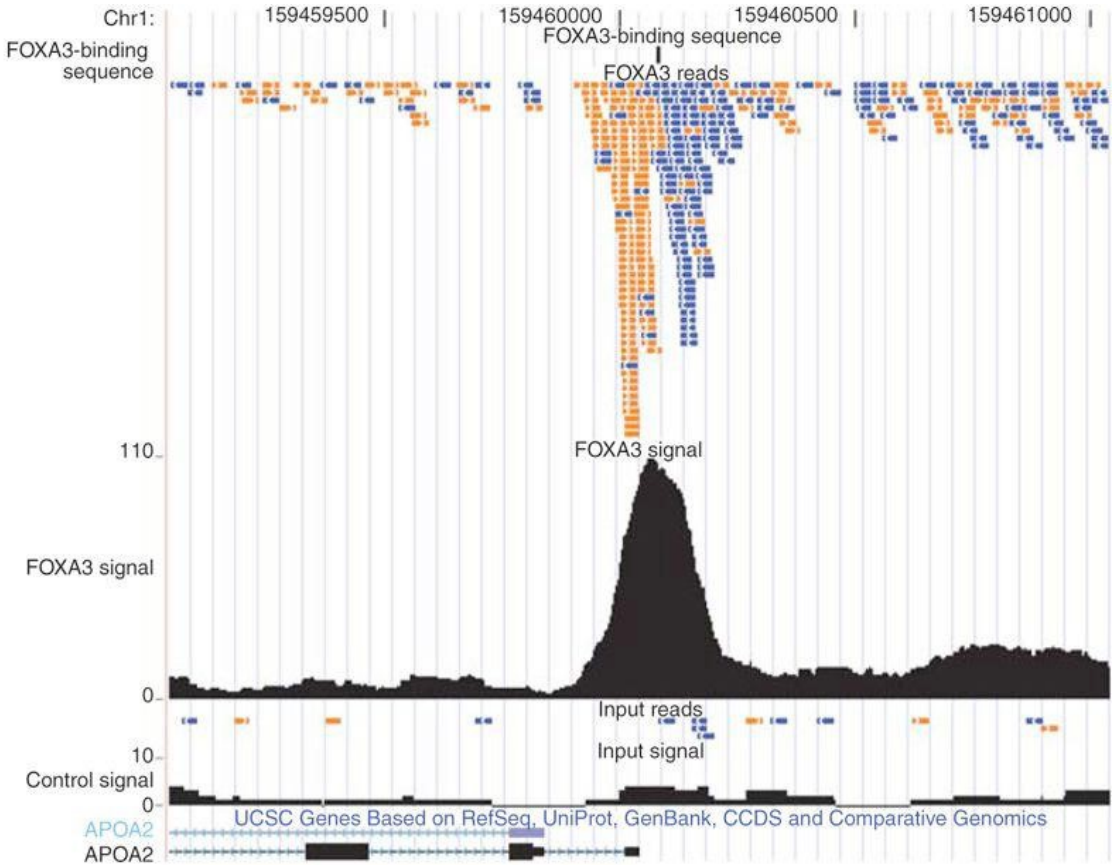


Figure 7: Example of aligned forward (orange) and reverse (blue) reads and the region of overlap (black) (www.nature.com/nmeth/journal/v6/n4/full/nmeth.f.247.html).

3.7 Peak annotation

To associate the ChIP-seq peaks with functionally relevant genomic regions (gene promoters, transcription start sites, etc), a peak annotation is needed [18]. Regions can be manually examined by looking for associations with annotated genomic features using a genome browser where to upload peaks and reads in the appropriate format requested by the browser. A systematic analysis may be carried out by computing the distance from each peak to the nearest feature (TSS).

3.8 Pathway annotation (graphite bioconductor package)

For the extension of graphite pathways annotation it is important to convert the EnsemblGeneID in EntrezGeneID since graphite uses EntrezGene IDs and Gene Symbols because of their widespread use and simplicity.

Due to their different origins and specificity, switching from an ID to another it is possible to find no correspondence between them (https://www.bioconductor.org/packages/3.3/bioc/vignettes/graphite/inst/doc/graphite.pdf). Biological interpretation is determined by the problem under study (motif discovery and analysis, Gene ontology (Gene set enrichment-style analyses), pathway analysis, etc.) and can also require to combine information from several experiments and data integration (it provides a more comprehensive complex picture of the biological system). Romualdi's research group has recently developed 'graphite', a bioconductor package useful to convert pathway topology to gene network [47]. Graphite takes the information from several databases, interprets pathway formats and reconstruct the correspondent gene-gene networks considering specific biologically driven rules. A biological pathway can be represented through a network (graph) where different types of genes and their interactions are, respectively, nodes and edges of the graph.

An appropriate representation of a pathway is important to enable efficient knowledge management and integration of data coming from several sources [48]. There are four categories of biological networks according to the nature of their nodes and their interactions: metabolic pathways, gene regulatory networks, molecular interactions and signaling pathways [49][50]. Graphite provides networks (graphs) from six public databases (KEGG [51], Reactome [52], Biocarta (www.biocarta.org), NCI/Nature Pathway Interaction Database [53], HumanCyc [54] and Panther [55]).

3.9 Fraction enrichment score (FES)

Similarity (or conversely distance) measures are becaming important due to the ever increasing amount of data available in public databases. Distance definitions, include Minkowski's distances (e.g. Euclidean, Manhattan), and clustering algorithms has been addressed by some recent works [56] [57] [58]. This project was started by the idea that sometimes global similarity measures have limitations due to the fact that only a limited set of features is responsible for the similarity and relevant signal may be hidden in noise. Thus, exploiting the fraction enrichment method used in the field of protein structure predictions [59] [60] and the method proposed by Spang and coworkers [61], based on the similarity of ranking in two ordered lists of genes, we employed a similarity measure (fraction enrichment proximity score (FES)) with the ability to recover similarities between different datasets comparing experiments in order to identify genes that best contribute to similarity. This conclusion supports the use of rank based proximity measures to gain further insight in datasets comparisons, in particular on expression data obtained by different technologies (e.g. RNA-seq and microarrays).

4 RESULTS AND DISCUSSION

Results of the different works I did in these 3 years are split in 4 sections, as you can see from the underlying paragraphs.

4.1 Similarity measures based on the overlap of ranked genes are effective for comparison and classification of microarray data **(see in section 7 – PUBLISHED PAPER)**

Similarity measures are central to many bioinformatics applications that aim at inferring novel knowledge from previous knowledge. Proper evaluation of similarity is more and more important due to the ever increasing amount of data available in public databases. We can exploit them by creating several interesting datasets. Regarding the methodology that can be employed to evaluate proximity measures we pay attention to the concept of intrinsic separation ability, i. e. how well a distance discriminates, so how well a distance is able to separate the objects belonging to different classes in a dataset .

Among the similarity measures Pearson correlation coefficient, Sperman correlation coefficient, Kendall tau correlation coefficient, Canberra distance, Mahnattan distance, Euclidean distance were the known proximity measures, whereas fraction enrichment proximity score (FES) was the one studied by us. All comparisons have been performed using a software environment for statistical computing and graphics, "R". FES is used to determine the degree of overlap between our experiment and other experiments. We tested its performance in robustness in separating replicate experiments from different unrelated experiments and different technologies (e.g. RNA-seq and microarrays).

Here we use the similarity measure proposed by Spang and coworkers [61] (shifted and scaled to bring it in the range 0 to 1, and including a linear weight decay) and:

- ✓ we compare its ability to recover similarities between different datasets with classical distances and for different choices of parameters and data pre-processing;
- ✓ we assess the relationship between such distance and the cardinality of ranked

genes with most significant overlap;

- ✓ we assess the relationship between such distance and the p-value of the overalap;
- ✓ we show that it is suited to compare data acquired with different technologies.

In the latter scenario a hybrid method like the the normalized rank-magnitude index based distance [62], which combines ranks and magnitudes of data, shows similar results, confirming its usefulness in comparing data with different scales and ranges.

Our results support the usefulness of similarity measures based on the overlap of ranked genes which perform as well or better as more traditional correlation measures for similarity recognition.

4.2 Computational analysis of ChIP-seq data and pathway annotation

During a 15-month stay in the Bioinformatics Laboratory of Professoressa Chiara Romualdi of Padua and under her supervision, I have worked on a ChIP-seq computational pipeline exploiting the heterogeneity of different algorithms with the aim to extend Graphite pathways annotation (KEGG, Reactome, Biocarta, Panther databases). Specifically, given a Chip-Seq result of a TF, pathways annotation was expanded adding to the network the TF (nodes, if not already annotated) whose target genes were already annotated in the pathway. This was an interesting work because although the information available in pathway annotation represents valuable information, only 30%-40% of coding genes are annotated in at least one pathway. This represents a strong limit to pathway-based transcriptome analysis. Thus, the presence of ChIP-seq ENCODE data (https://genome.ucsc.edu/ENCODE/) represents an important resource to improve pathway annotation. To create a robust ChIP-seq computational pipeline I downloaded from Encode site six raw datasets (fastQ) (**Table 1**):

Cell line Hela-s3	**Cell line Hela-s3**
Transcription Factor: Pol2	Transcription Factor: STAT1
2 runs: Pol2_Rep1, Pol2_Rep2	**2 runs:** STAT1_Rep1, STAT1_Rep2
1 input: SRR357521	1 input: Standard Input

Table 1: Encode six raw datasets

The choice of these two datasets (Polymerase II (POL2) and Signal Transducer And Activator Of Transcription 1 (STAT1)) was done because they are well characterized Transcription Factors. Preprocessing of ChIP-seq data is important to assess the quality of the raw reads [18]. FastQC is a tool (www.bioinformatics.babraham.ac.uk/projects/fastqc) suitable to identify possible sequencing errors or biases. Before mapping the reads onto a reference genome, it is also necessary to trim the end of reads that are of low quality (the trimming was done using trimmomatic tool [30]). The millions of short sequence reads (75-100 bps)

are mapped to the reference genome (Ensembl75/hg19), we tested three different alignment algorithms (BWA [33], Bowtie [31], Bowtie2 [32]) and at the end we decided to use the new version of Bowtie (Bowtie2). Using Bowtie2 we tried several options in order to be sure to have only uniquely aligned reads. The identical reads are a problem to solve, sometimes they are noise (experimental artefact) and not signals [18]. Peaks (regions with high read densities) are identified by peak-calling algorithms. These algorithms exploits different statistical models to obtain the p-value and the false discovery rate (FDR) **(Benjamini and Hochberg (1995))**, therefore, significance values from different peak-calling algorithms are not directly comparable [35]. For our work we decided to converge on two peak callers (SPP [19] and MACS2 (https://github.com/taoliu/MACS)) to be used in parallel for all analysis testing three different q-values cut-off (0.01, 0.05 and 0.1). The q-value [63] is a measure of statistical significance in terms of the false discovery rate (FDR) rather than the false positive rate (FPR) as the p-value. We compared our results with the results obtained by the analysis performed by Anshul Kundaje alongside the ENCODE Binding working group (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeATfbsUniform). They treated the two replicates of each transcription factor as a pool of replicates. The Encode site (https://genome.ucsc.edu/ENCODE/) is not a mere deposit of ChIP-seq raw datasets but it is also a repository of analyzed data.

Below I report the table of the number of peaks called by the three peak callers (**Table 2**), I explained above.

Transcription Factors	**Run**	**MACS2**	**SPP**	**Uniform Peak Calling** (pool of two replicates for TF)
POL2	Rep_1	23049	26504	15736
POL2	Rep_2	16920	19556	
STAT1	Rep_1	11215	19954	12629
STAT1	Rep_2	18780	28224	

Table 2: Number of peaks called by the three peak callers.

Peak calling gives a list of enriched regions where the protein of interest should be directly or indirectly associated with DNA [64]. Data visualization and exploration was done using the Integrative Genomics Viewer (IGV) desktop tool [42], converting the WIG files obtaned from SPP peak caller in TDF files (useful to visualize reads density). Our interest was to find the nearest features such as genes calculating the distance from the center of the peak to the transcription start site (TSS). The peaks were annotated using Ensembl75 [65] as database of gene annotation. A ChIP-seq peak annotation script was created suited for our purposes; target genes were found calculating the distance from the transcription start site (TSS) of the gene to the center of the peaks and selecting only peaks falling in a genomic region of 2kb from the TSS. A table with the number of genes closest to peaks called by the three peak callers is reported in the table below (MACS2 (https://github.com/taoliu/MACS), SPP [19] and Uniform Peak Calling) for the two transcription factors we decided to study (**Table 3**).

Transcription Factors	**Run**	**MACS2**	**SPP**	**Uniform Peak Calling** (genes in common between the two replicates)
POL2	Rep_1	22550	26224	14486
POL2	Rep_2	16354	10286	
STAT1	Rep_1	5292	16803	4738
STAT1	Rep_2	15452	22297	

Table 3: Number of genes closest to peaks called by three peak callers.

Considering the target genes identified we compare the results of the caller across replicates (**Table 4, 5).**

SPP (Pol2-Rep_1 vs Pol2-Rep_2)	
peaks in common between the two replicates	18620 peaks
peaks (associated to the same gene)	17530 peaks
peaks (call different genes)	1020 peaks
peaks (no associated to genes)	580 peaks
genes in common between the two replicates	12774 genes

Table 4: Studying of the sensitivity and specificity of SPP peak caller.

MACS2 (Pol2-Rep_1 vs Pol2-Rep_2)	
peaks in common between the two replicates	15282 peaks
peaks (associated to the same gene)	14081 peaks
peaks (call different genes)	881 peaks
peaks (no associated to genes)	320 peaks
genes in common between the two replicates	9578 genes

Table 5: Studying of the sensitivity and specificity of MACS2 peak caller

I have also checked the peaks width distribution of the three peak callers (MACS2 (https://github.com/taoliu/MACS), SPP [19] and Uniform Peak Calling), as one can see from the density plots below **(Figures 1, 2),** in order to see how large were the binding regions found by the three peak callers. During the first check, using POL2 as transcription factor, we noticed that the width of peaks called by SPP [19] was 5-6 times larger than that called by the other two peak callers (MACS2 (https://github.com/taoliu/MACS) and Uniform Peak Calling). In order to understand whether the result was confirmed, I performed the peak calling analysis a second time with another transcription factor (STAT1) and even in this case we obtained the same result **(Figures 3, 4).**

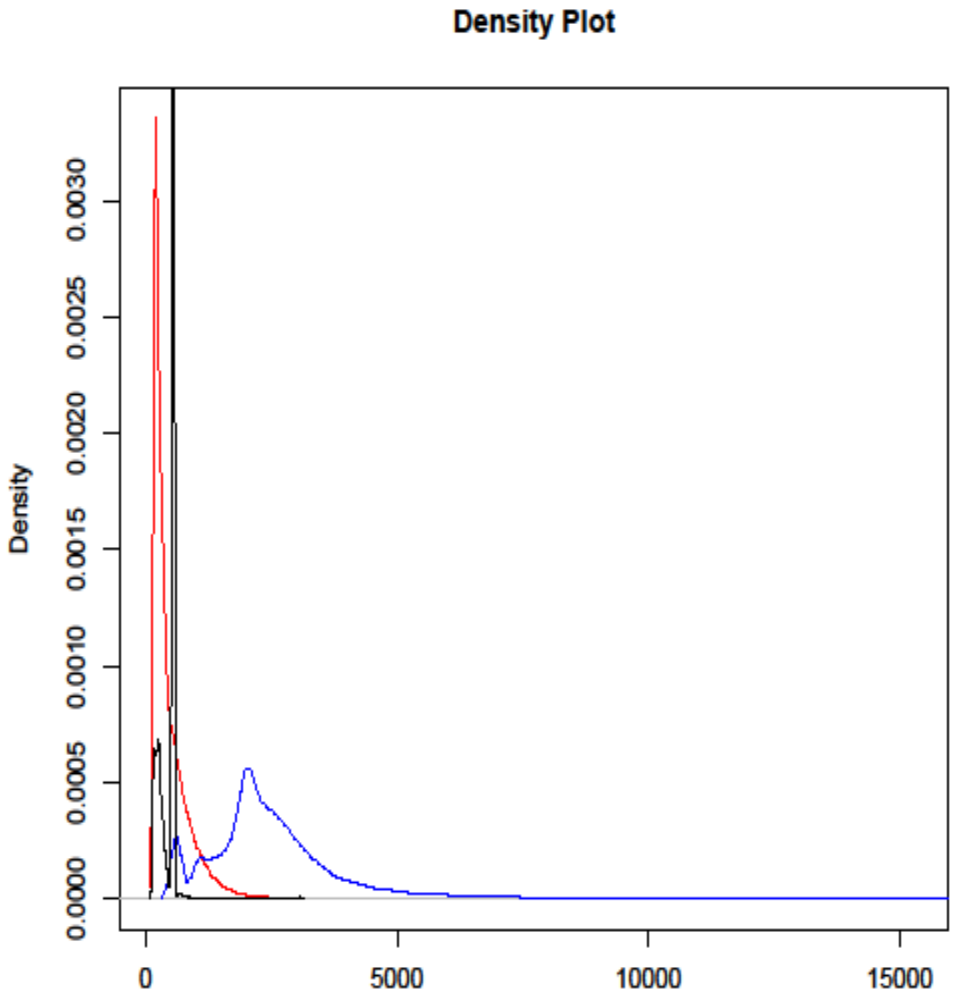


Figure 1: Width distribution of the three peak callers. Blue line SPP, Red line MACS2, Black line Uniform Peak Calling (pool of two replicates).

Mean of the distribution of peaks width (Pol2-Rep_1 - POL2)

- MACS2 = 481.905
- SPP = 2698.953
- Uniform Peak Calling = 504.1407

Median of the distribution of peaks width (Pol2-Rep_1 - POL2)

- MACS2 = 331
- SPP = 2310
- Uniform Peak Calling = 544

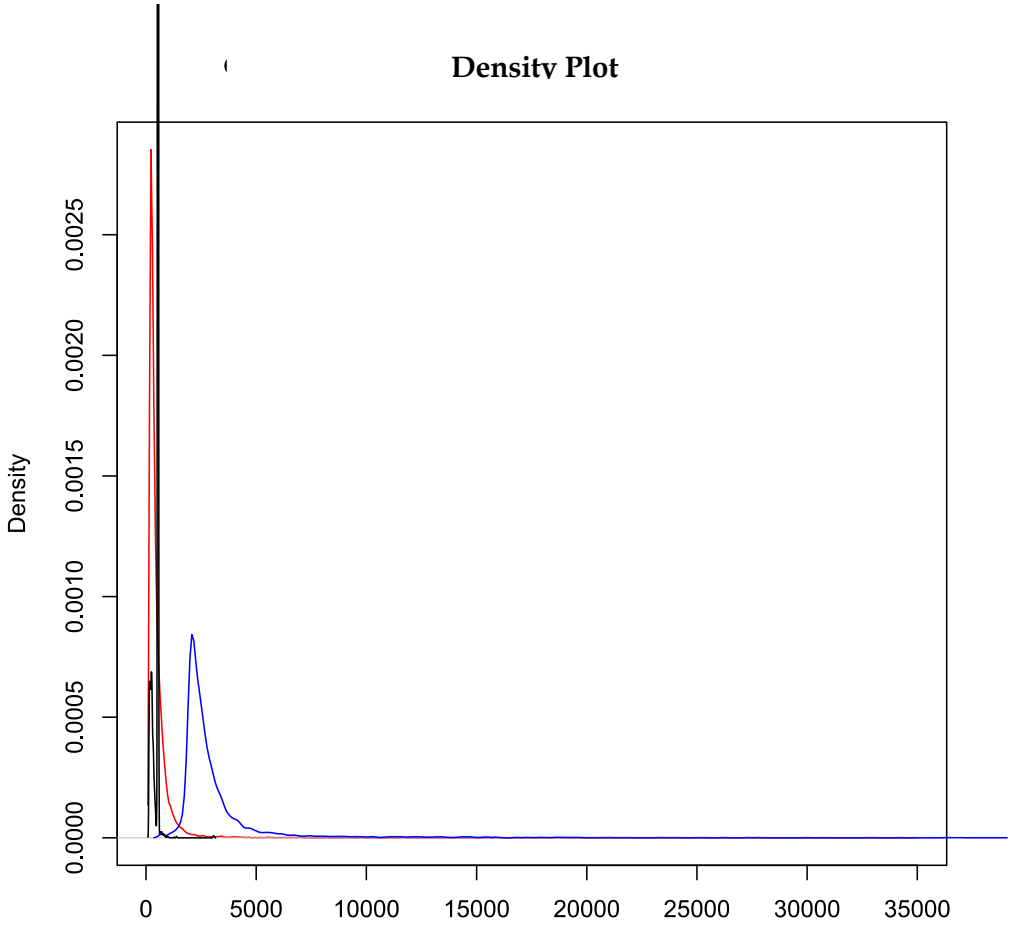


Figure 2: Width distribution of the three peak callers. Blue line SPP, Red line MACS2, Black line Uniform Peak Calling (pool of two replicates).

Mean of the distribution of peaks width (Pol2-Rep_2 - POL2)

- MACS2 = 499.6151
- SPP = 3130.317
- Uniform Peak Calling = 504.1407

Median of the distribution of peaks width (Pol2-Rep_2 - POL2)

- MACS2 = 339
- SPP = 2459
- Uniform Peak Calling = 544

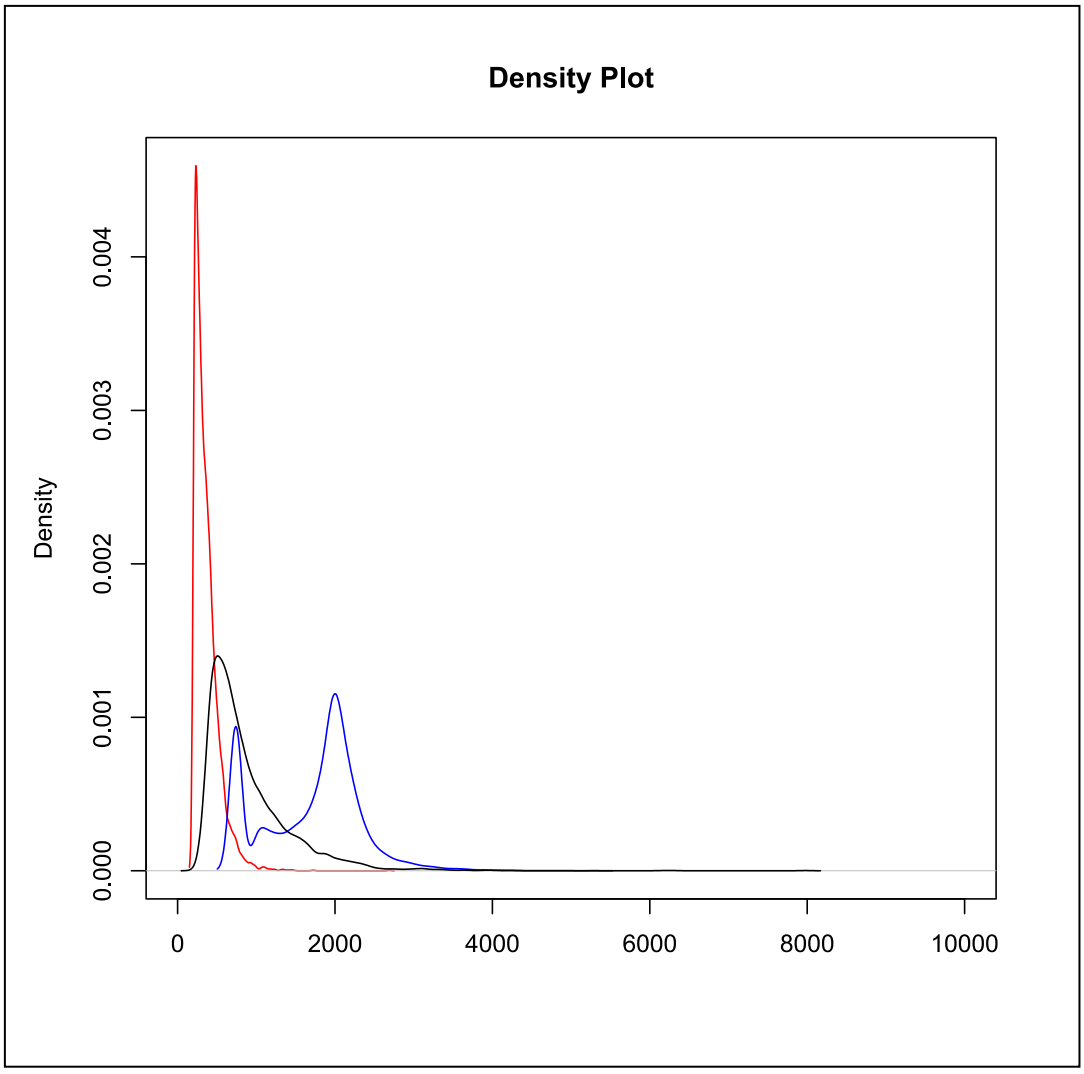


Figure 3: Width distribution of the three peak callers. Blue line SPP, Red line MACS2, Black line Uniform Peak Calling (pool of two replicates).

Mean of the distribution of peaks width (STAT1_Rep1 - STAT1)

- MACS2 = 361.6215
- SPP = 1702.674
- Uniform Peak Calling = 880.051

Median of the distribution of peaks width (STAT1_Rep1 - STAT1)

- MACS2 = 321
- SPP = 1880
- Uniform Peak Calling = 724

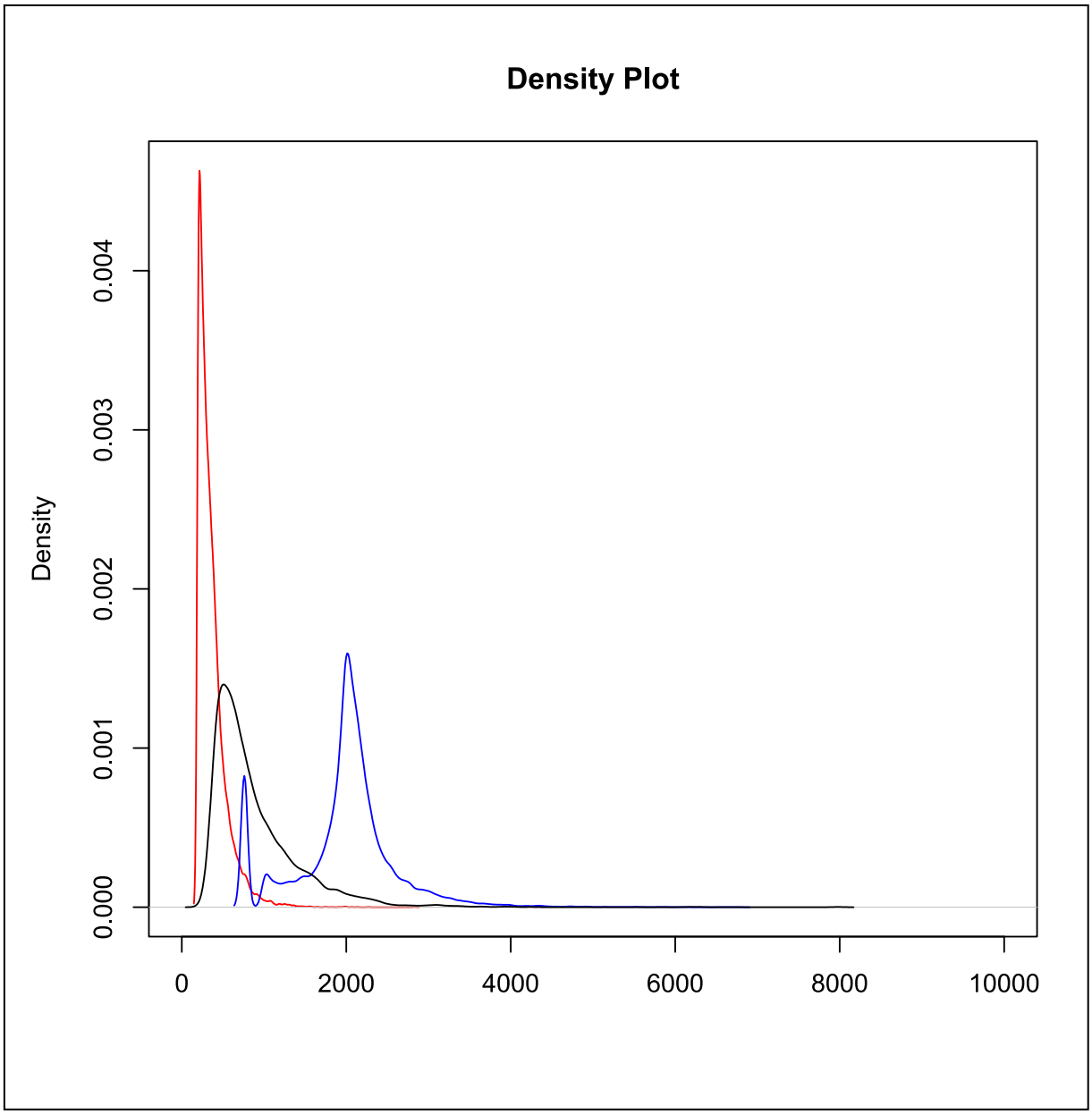


Figure 4: Width distribution of the three peak callers. Blue line SPP, Red line MACS2, Black line Uniform Peak Calling (pool of two replicates).

Mean of the distribution of peaks width (STAT1_Rep2 - STAT1)

- MACS2 = 360.8803
- SPP = 1984
- Uniform Peak Calling = 880.051

Median of the distribution of peaks width (STAT1_Rep2 - STAT1)

- MACS2 = 313
- SPP = 2027
- Uniform Peak Calling = 724

We tried different SPP [19] parameter settings but we obtained the same results, a peak width extremely large with respect to the other method. This larger peak of course influences also the target annotation. Given the odd results of SPP [19] we decided to abandon this peak caller in favour of MACS [36] that shows a more similar results with the ENCODE consortium data. Summarising, during this project I used several algorithms of alignment (BWA [33], Bowtie [31] and Bowtie 2 [32]) and peak callers to compare their performance. From the results obtained we decided to use the new version of Bowtie (Bowtie2) for alignment and MACS2 (https://github.com/taoliu/MACS) as peak caller (one of the most used peak callers) because the results were more similar to the Uniform peak calling algorithm (used by the Encode Binding working group). Testing several algorithms and tuning their functions I have generated a robust ChIP-seq computational pipeline used for analysing the transcription factors raw datasets of 22 cell lines downloaded from the Encode site (https://genome.ucsc.edu/ENCODE/). All Encode data are freely available for download and analysis. In this table I report the number of transcription factors used for each of the 22 cell lines downloaded (**Table 6**).

Cell lines	Description	Tissue	Karyotype	Number of transcription factors
HELA-S3	cervical carcinoma	Cervix	Cancer	54
GM12878	B-lymphocyte, lymphoblastoid	Blood	Normal	77
A549	epithelial cell line derived from a lung carcinoma tissue	Epithelium	Cancer	24
GM12891	B-lymphocyte, lymphoblastoid, Epstein-Barr Virus transformed	Blood		8
GM08714	lymphoblastoid cell line, Instability of heterochromatin of chromosomes 1, 9, and 16 with variable combined immunodeficiency	Blood		1
GM18526	lymphoblastoid,Epstein-BarrVirus transformed	Blood		2
HCT-116	colorectal carcinoma	Colon	Cancer	5
HEK293	embryonic kidney, cells contain Adenovirus 5 DNA	Kidney		5
H1-hESC	embryonic stem cells	Embryonic Stem Cell	Normal	49
HEPG2	hepatocellular carcinoma	Liver	Cancer	59
K562	chronic myelogenous leukemia	Blood	Cancer	100
MCF-7	mammary gland, adenocarcinoma.	Breast	Cancer	7
NB4	acute promyelocytic leukemia cell line.	Blood	Cancer	3
NT2-D1	malignant pluripotent embryonal carcinoma	Testis	Cancer	3
PANC 1	pancreatic carcinoma	Pancreas	Cancer	4
PBDE	peripheral blood-derived erythroblasts	Blood		2
PBDE FETAL	peripheral blood-derived erythroblasts from 16-19 week human fetal liver	Liver	Normal	1

RAJI	Lymphoma	Blood	Cancer	1
SH-SY5Y	metastatic neuroblastoma	Brain	Cancer	2
U2os	Osteosarcoma	Bone	Cancer	2
MCF 10A-Er-Src	mammary gland, non-tumorigenic epithelial, inducible cell line, derived from the MCF-10A parental cells and contain ER-Src, a derivative of the Src kinase oncoprotein (v-Src) that is fused to the ligand-binding domain of the estrogen receptor (ER)	Breast		5
HUVEC	umbilical vein endothelial cells	Blood Vessel	Normal	8

Table 6: Number of transcription factors used for each of the 22 cell lines.

As I have previously written, in Encode there are also the Unipk files (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeATfbsUniform) (derived from the peak detection done by the Encode working group using the Uniform Peak Calling algorithm) I have downloaded and used to compare the results with my data analysis. The peaks found in common were annotated using Ensembl75 [65] as database of gene annotation. To find target genes I have created a peak annotation script narrowing the research for transcription factors binding site to a region of 2 Kb near the TSS. These results were employed for working on the pathways annotation using graphite bioconductor package. To extend kegg pathways annotation using transcription factors, we need first to convert pathways into gene network (graphical structure in which a node represents a simple element such as a gene/protein) [47].
In fact, while pathway nodes could consist of multiple entities (protein complexes, gene family members and chemical compounds), we need to consider each single component of complexes and gene family separately [66]. Once converted to gene-gene network we added a transcription factor in Kegg pathway if its target genes (nodes) were already present in the pathway annotation. It is known that ChIP-seq data finds lots of false positive bindings, thus we decide to use a data-driven strategy to filter potential false positives. For this task we used the gene expression levels (RNA-seq data) derived from the Expression Atlas database (http://www.ebi.ac.uk/gxa/experiments/E-GEOD-26284) available on the same cell lines we used for the ENCODE data. This database contains a table with listed the Fragments Per Kilobase Of Exon Per Million Fragments Mapped (FPKM) of 9 out of 22 Encode cell lines used for the analysis. If the TF and its target genes are both expressed in its cell lines (we used different cut-off, greater than 5, 10, 20 FPKM) then

the edge is added to the pathway, otherwise it is filtered off. To have an idea of how many edges and transcription factors we are going to add to each pathway for the 9 cell lines considering the three FPKM cut-off, some statistics were evaluated (mean, median, min and max). The results are reported in the following 6 tables (**Tables 7, 8, 9, 10, 11, 12**).

9 Cell lines	Description	n° of pathways	n° of edges	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	197	4611	23,4061	18	1	117
A549	Lung Carcinoma	102	1007	9,87255	7	2	30
K562	Chronic Myelogenous Leukemia	199	21948	110,291	87	1	557
MCF-7	Mammary gland, adenocarcinoma	130	442	3,4	3	1	13
HCT-116	Colorectal Carcinoma	26	29	1,11538	1	1	3
HeLa-S3	Cervical Carcinoma	189	6889	36,4497	29	1	146
GM12878	B-lymphocyte	188	12377	65,8351	51	2	290
HUVEC	Umbilical vein endothelial cells	101	206	2,0396	2	1	6
HEPG2	Hepatocellular Carcinoma	189	6688	35,3862	29	1	164

Table 7: Total number of edges to add to the several pathways listed considering an FPKM cut-off of 10 and mean, median, min and max of edges to add to each pathway of each cell line.

9 Cell lines	Description	n° of pathways	n° of edges	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	226	15830	70,0442	54	1	359
A549	Lung Carcinoma	165	3047	18,4667	15	2	65
K562	Chronic Myelogenous Leukemia	227	84547	372,454	281	1	1812
MCF-7	Mammary gland, adenocarcinoma	187	1681	8,9893	7	1	40
HCT-116	Colorectal Carcinoma	58	72	1,24138	1	1	3
HeLa-S3	Cervical Carcinoma	226	27543	121,872	83,5	2	629
GM12878	B-lymphocyte	225	46916	208,516	157	2	1115
HUVEC	Umbilical vein endothelial cells	189	1278	6,7619	6	1	34
HEPG2	Hepatocellular Carcinoma	225	25911	115,16	90	1	635

Table 8: Total number of edges to add to the several pathways listed considering an FPKM cut-off of 2 and mean, median, min and max of edges to add to each pathway of each cell line.

9 Cell lines	Description	n° of pathways	n° of edges	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	229	20650	90,1747	67	1	454
A549	Lung Carcinoma	173	3860	22,3121	18	2	84
K562	Chronic Myelogenous Leukemia	229	106974	467,135	343	7	2230
MCF-7	Mammary gland, adenocarcinoma	197	2000	10,1523	8	1	50
HCT-116	Colorectal Carcinoma	66	102	1,54545	1	1	5
HeLa-S3	Cervical Carcinoma	229	35563	155,297	109	3	773
GM12878	B-lymphocyte	229	60005	262,031	196	3	1345
HUVEC	Umbilical vein endothelial cells	199	1588	7,9799	6	1	37
HEPG2	Hepatocellular Carcinoma	229	33764	147,441	112	2	777

Table 9: Total number of edges to add to the several pathways listed considering an FPKM cut-off of 0.5 and mean, median, min and max of edges to add to each pathway of each cell line.

9 Cell lines	Description	n° of pathways	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	197	1,88051	1	1	17
A549	Lung Carcinoma	102	1,5374	1	1	5
K562	Chronic Myelogenous Leukemia	199	2,61224	2	1	27
MCF-7	Mammary gland, adenocarcinoma	130	1,26648	1	1	6
HCT-116	Colorectal Carcinoma	26	1,11538	1	1	3
HeLa-S3	Cervical Carcinoma	189	2,04664	1	1	17
GM12878	B-lymphocyte	188	2,15103	1	1	23
HUVEC	Umbilical vein endothelial cells	101	1,22619	1	1	4
HEPG2	Hepatocellular Carcinoma	189	1,84751	1	1	18

Table 10: Mean, median, min and max of transcription factors to add to each pathway of each cell line considering an FPKM cut-off of 10.

9 Cell lines	Description	n° of pathways	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	226	3,42	2	1	53
A549	Lung Carcinoma	165	2,01122	1	1	13
K562	Chronic Myelogenous Leukemia	227	6,03003	3	1	96
MCF-7	Mammary gland, adenocarcinoma	187	2,34777	2	1	16
HCT-116	Colorectal Carcinoma	58	1,10769	1	1	3
HeLa-S3	Cervical Carcinoma	226	4,41111	2	1	60
GM12878	B-lymphocyte	225	4,77954	3	1	86
HUVEC	Umbilical vein endothelial cells	189	1,96918	1	1	12
HEPG2	Hepatocellular Carcinoma	225	3,63153	2	1	54

Table 11: Mean, median, min and max of transcription factors to add to each pathway of each cell line considering an FPKM cut-off of 2.

9 Cell lines	Description	n° of pathways	Mean	Median	Min	Max
H1-hESC	Embryonic stem cells	229	3,8	2	1	63
A549	Lung Carcinoma	173	2,15	1	1	14
K562	Chronic Myelogenous Leukemia	229	7,11926	3	1	118
MCF-7	Mammary gland, adenocarcinoma	197	2,53807	2	1	19
HCT-116	Colorectal Carcinoma	66	1,17241	1	1	4
HeLa-S3	Cervical Carcinoma	229	5,20536	3	1	71
GM12878	B-lymphocyte	229	5064275	3	1	104
HUVEC	Umbilical vein endothelial cells	199	2,21478	1	1	13
HEPG2	Hepatocellular Carcinoma	229	4,22737	2	1	64

Table 12: Mean, median, min and max of transcription factors to add to each pathway of each cell line considering an FPKM cut-off of 0.5.

To visualize the molecular interaction networks created I used an R package called Rcytoscape useful to export the network to Cytoscape (**www.cytoscape.org**). Cytoscape [67] is an open source software specifically built to manage biological network complexity. Several pathways have a huge number of nodes and edges, thus an efficient system of visualization is needed.

In order to evaluate the impact of this inclusion on the topological pathway analysis [66] we use a benchmark dataset (in which a true positive result is known). The topological pathway analysis was performed by Clipper (a tool freely available as an R package), a novel algorithm for pathway analysis that implements a two-step empirical approach according to the exploitation of graph decomposition into a junction tree, then the reconstruction of the most relevant signal path is performed [66]. Clipper chooses significant pathways based on statistical tests on the means and the concentration matrices of the graphs. It identifies the signal paths mostly associated to a specific phenotype. The basic idea is that despite the increase in pathway annotation the true results should continue to be identified by the statistical method.

The benchmark dataset we used for the analysis was published by Chiaretti et al. [68]. This dataset characterizes gene expression signatures in acute lymphocytic leukemia (ALL) cells. Different genetic mechanisms drive to ALL malignant transformations coming from distinct lymphoid precursor cells (committed to both T-lineage or B-lineage differentiation). The frequencies of specific molecular rearrangements and chromosome translocations are different in adults and children with B-lineage ALL. In about 25% of adult ALL cases the BCR break-point cluster region and the c-abl oncogene 1 (BCR/ABL) gene rearrangement is found, instead, in pediatric ALL cases is much less frequent (the data are available at the bioconductor site **(http://www.bioconductor.org/help/publications/2003/Chiaretti/chiaretti2/))**. Expression values, coming from Affymetrix technology, were normalized by robust multiarray analysis (rma) and quantile normalization. These expression values consist of 37 observations from patients with BCR/ABL gene rearrangement and 41 observations from patients without rearrangements.

Focusing on chronic myeloid leukaemia (CML) pathway, that includes exactly BCR/ABL fusion gene, we want that the analysis identifies a path starting from BCR/ABL toward the oncogene TP53 **(Figure 5)**.

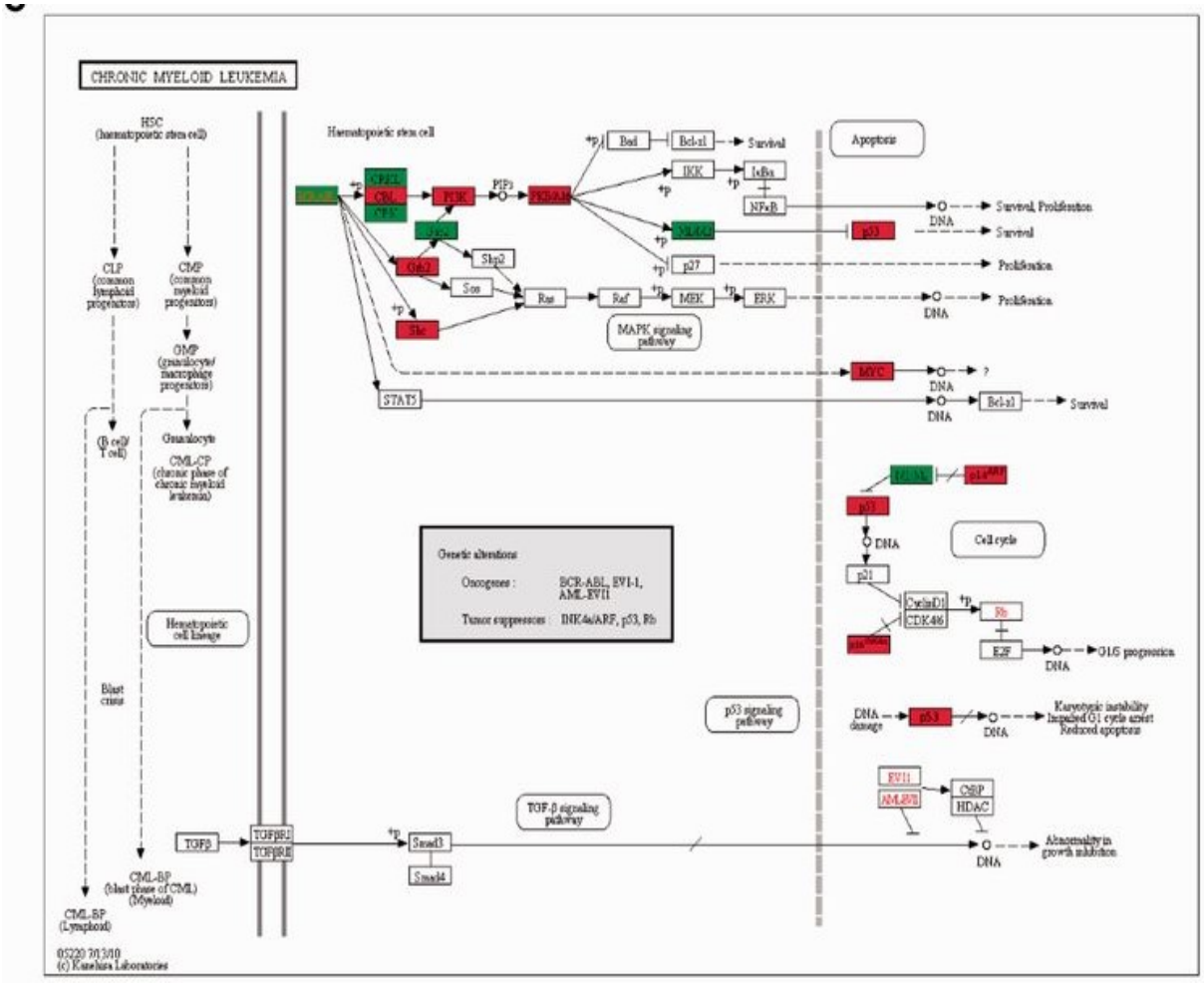


Figure 5: Clipper results on chronic myeloid leukaemia (CML) KEGG pathway obtained by Martini et al [66]. This figure shows the CML pathway with complexes belonging to the sub-path identified colored in red or green according to their expression.

The BCR/ABL fusion protein, in chronic myeloid leukaemia (CML) cells, is involved in the accumulation of p53, and neutralize the inhibitory activities of p53 by modulating the p53-MDMD2 loop [69]. By modulating this loop, c-Abl and its oncogenic forms decide the particular kind and extent of the cellular response to DNA damage.

We applied Clipper analysis on chronic myeloid leukaemia (CML) pathway in which different list of TF-target genes edges have been included (according to the three expression cut-off previously described). In all the three extended chronic myeloid leukaemia (CML) pathway annotations we identified the same path starting from the chimera gene through the TP53 gene. The results of this pilot project were encouraging, showing that despite the extension of pathway topology and then the increase in the number of parameters to be estimated, the statistical method is still robust and efficient in the identification of the true result. This could be the starting point to determine new strategy for pathway annotation extentions.

4.3 Identification of target genes directly regulated by APE1 during oxidative stress condition (ChIP-seq analysis)

The work, I performed under the supervision of Professor Tell, aimed at studying the transcriptional and post-transcriptional mechanisms responsible for the gene-regulatory functions of APE1. APE1 has transcriptional regulatory activity modulating gene-expression through redox-based co-activating function on several transcription factors involved in cancer promotion [22]. The research group of Professor Tell has recently demonstrated APE1 involvement in SIRT1 transcription through the direct binding to nCaRE elements found in SIRT1 promoter [22]. They demonstrated that APE1 is part of a multi-protein complex, which includes hOGG1, Ku70 and RNA Pol II and regulates SIRT1 promoter activation during early response to oxidative stress. Based on these findings, deep sequencing high-throughput studies might provide new insights in the comprehension of the role of APE1 in the transcriptional regulation of mammalian genes. The identification of target genes directly regulated by APE1 during oxidative stress condition was performed by deep sequencing approaches, as ChIP-seq analysis, in order to identify APE1 preferential promoter binding sites. HeLa cell clones, in which the endogenous APE1 protein expression was previously knocked down through inducible expression of stable short hairpin RNA targeting APE1 mRNA, were used and re-expressing an ectopic Flag-tagged siRNA-resistant wild type APE1 cDNA [70]. Chromatin immunoprecipitation was performed using an anti-FLAG resin in order to recover only the APE1 flag-tagged protein [22]; scramble clones, which do not express the ectopic protein, represent the negative control. Library preparation and high-throughput DNA sequencing was performed at IGA (Istituto di Genomica Applicata), using the Illumina HiSeq platform. Only the DNA coming from WT clones treated or not with H_20_2 (0.5 mM for 15 minutes) was sequenced; scramble clones were considered as internal control of negative immunoprecipitation. I have generated a ChIP-seq data analysis pipeline exploiting the heterogeneity of different algorithms. The idea is to couple ChIP-seq and RNA-seq analysis to identify potential genes directly regulated by APE1 binding to their promoters. Raw data were

subjected to quality control analysis using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). In order to avoid low quality data, adapters were removed and lower quality bases were trimmed by trimmomatic [30]. The quality-checked reads were mapped to the human reference genome Ensembl 75/hg19. Peak calling to identify APE1 preferential promoter binding sites was performed by using MACS2 (https://github.com/taoliu/MACS) algorithm. After the annotation of peaks the first control was to check APE1 binding site to SIRT1 promoter as an internal control readout. Since there was an apparent enrichment in APE1 binding in proximity of SIRT1 transcription start site (TSS), I have created a peak annotation script to narrow the research for APE1 binding site to a region of 2 Kb close to the TSS to find genes potentially regulated by APE1 under stress condition. The peaks were annotated using Ensembl75 [65] as database of gene annotation. The **figure 6** indicates the number of genes found in the two datasets (HeLa treated and non-treated with H_2O_2) and the amount of genes found in common in the two conditions.

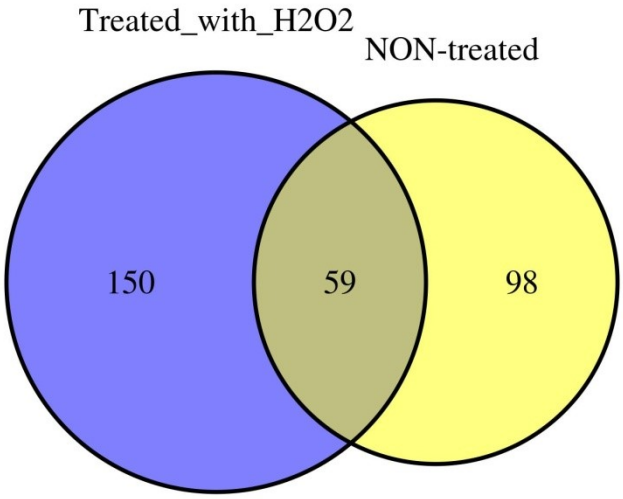


Figure 6: Number of APE1 genes target found in common between the two datasets of HeLa treated or not with H_2O_2.

In the following table the list of APE1 target genes found in common between the two datasets of HeLa treated or not with H_2O_2 is reported (**Table 13**). The corresponding chromosome, gene symbol and type are indicated (59 genes). The complete lists of APE1 target genes (HeLa treated and non-treated with H_2O_2) are present in APPENDIX - section 5.

Chromosome	Gene Symbol	Type
2	AC009299.4	Pseudo gene
5	ADAMTS12	protein_coding
1	ANKRD20A14P	unprocessed_pseudogene
5	C6	protein_coding
8	CCAT1	lincRNA
17	CDC27	protein_coding
2	CDC27P1	Pseudo gene
5	CDH12	protein_coding
5	CDH18	protein_coding
5	CDH9	protein_coding
5	CTD-2306M5.1	lincRNA
16	CTD-2522B17.8	Pseudo gene
5	CTD-2533K21.4	lincRNA
5	CTD-3007L5.1	lincRNA
5	DAB2	protein_coding
5	DNAH5	protein_coding
5	DROSHA	protein_coding
Y	DUX4L16	Pseudo gene
Y	DUX4L17	Pseudo gene
16	FAM157C	processed_pseudogene
22	FAM230B	processed_pseudogene
4	FRG1	protein_coding
14	IGHV1-68	IG_V_pseudogene
22	LA16c-83F12.6	lincRNA
5	NADK2	protein_coding
1	NBPF10	protein_coding
1	NBPF14	protein_coding
1	NBPF8	protein_coding
5	NNT	protein_coding
1	NOTCH2NL	protein_coding
8	PCMTD1	protein_coding
5	RICTOR	protein_coding
5	RNA5SP177	rRNA
4	RP11-1281K21.6	Pseudo gene
7	RP11-1324A7.2	processed_pseudogene
5	RP11-192H6.2	lincRNA
9	RP11-262H14.1	lincRNA
9	RP11-318K12.3	Pseudo gene
5	RP11-321E2.6	processed_pseudogene
1	RP11-353N4.5	lincRNA

1	RP11-417J8.1	lincRNA
1	RP11-417J8.2	lincRNA
5	RP11-42L13.2	Pseudo gene
1	RP11-435B5.5	lincRNA
4	RP11-463J17.1	lincRNA
5	RP11-53O19.1	Antisense
11	RP11-56P9.5	Pseudo gene
16	RP11-626K17.5	unprocessed_pseudogene
1	RP11-763B22.9	unprocessed_pseudogene
9	RP11-764K9.1	lincRNA
9	RP11-764K9.4	unprocessed_pseudogene
5	RP11-774D14.1	lincRNA
5	RPL36AP21	Pseudo gene
1	SEC22B	processed_pseudogene
10	SIRT1	protein_coding
5	SPEF2	protein_coding
5	TARS	protein_coding
5	UBL5P1	Pseudo gene
22	XXbac-B33L19.3	Antisense

Table 13: List of APE1-target genes found in common between the HeLa treated and non-treated with H_2O_2 datasets.

Remarkably, among these genes, long non-coding RNA (lincRNA), microRNA (miRNA) and small nuclear RNA (snRNA) are present in addition to some protein coding genes. In 2009, Prof. Tell's Lab published a work [71], in which they combined both mRNA expression profiling and proteomic analysis to determine the molecular changes associated with APE1 loss-of-expression induced by siRNA technology, using the same cellular model used in the present study. Through this approach, a role of APE1 in cell growth, apoptosis, intracellular redox state, cytoskeletal structure and mitochondrial function was suggested. To identify genes showing changes in their expression due to APE1 silencing, they compared the gene expression profiles of WT and APE1 knocked-down cell clones by using the human Affimetrix GeneChip (HG-U133 PLUS2) comprising a representation of over 20.000 genes. The data analysis (normalization and summarization) was performed by using RMAExpress algorithm [72]. Differential expression in response to siRNA treatment was calculated by exploiting the function of Cyber-T algorithm [73] and a false-discovery correction was applied to these p-values to obtain a q-value [63]. A gene was considered to be differentially expressed in APE1-deficient HeLa cells when the q-value was less than 0.05 and the fold change (FC) was greater than or less than 1.5.

By exploiting these criteria, 1126 genes were identified as differentially expressed (550 up-regulated and 576 down-regulated).

Comparing these up and down-regulated genes with the list of APE1 target genes (ChIP-seq analysis) five target genes among the down-regulated (DAB, NNT, SIRT1, TARS, OSMR) were found. Due to the poor correlation between ChIP-seq and gene expression data, it is conceivable to hypothesize that post-transcriptional mechanisms may be the main responsible for the gene-regulatory functions of APE1.

4.4 Identification of APE1-RNA interactome network through RNA-IP (RIP) analyses

APE1 regulates the expression of tumor-progression genes through transcriptional effects and post- transcriptional mechanisms [74]. Deep sequencing approach based on RIP-seq analysis was planned with the aims at implementing current translational approaches for cancer treatment which rely on APE1 (an essential protein for the maintenance of genome stability) as a target molecule, providing a complete list of target genes, mRNA, miRNA and ncRNA that are directly regulated by APE1 during cell response to genotoxic treatment in cancer cells and that could specifically mediate cancer cell resistance to chemotherapy. Based on the results obtained through the previous ChIP-seq approach, a RIP-seq strategy was used to identify the RNA molecules directly regulated by APE1 in order to test whether post-transcriptional mechanism may explain the gene-regulatory functions of APE1. RNA-bound by APE1 from HeLa cell clones, expressing an ectopic APE1 FLAG-tagged form in place of the endogenous one [70], was purified using an anti-FLAG antibody. APE1 binding to RNA was tested using three independent immunoprecipitations and to reduce potential false positives, a negative control was used. FLAG-APE1 was efficiently affinity-purified exclusively from HeLa cells immunoprecipitated with the resin carrying the FLAG antibody, this was confirmed by Western blot analysis. RNA bound to APE1 was then subject to library preparation, sequencing and bioinformatic analysis, done by Istituto di Genomica Applicata (IGA). TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat (Illumina, San Diego, CA) was used for library preparation. RNA samples and final libraries were quantified by using the Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and quality tested by Agilent 2100 Bioanalyzer RNA Nano assay (Agilent technologies, Santa Clara, CA). Libraries were then processed with Illumina cBot for cluster generation on the flowcell and sequenced on single-end mode using the HiSeq2500 (Illumina, San Diego, CA). The CASAVA 1.8.2 version of the Illumina pipeline was used for processing raw data for both format conversion and de-multiplexing. Raw sequence files were subjected to quality control analysis using

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Adapters were removed, in order to avoid low quality data, by Cutadapt [75] and lower quality bases were trimmed by ERNE [76]. The quality-checked reads were processed using the TopHat version 2.0.0 package (Bowtie 2 version 2.2.0) as FASTQ files [77][78][79]. The reads were mapped to the human reference genome GRCh37/hg19. Reads abundance was evaluated and normalized by using Cufflinks [78] for each gene. Differential enrichment analysis, between RNA-immunoprecipitated and input, and statistical significance evaluation of detected alterations were obtained using Cuffdiff as algorithm [78]. The biological significance of three lists of RNAs bound by APE1 were investigated, and analysed in different manners. By comparing 2 out of 3 lists (normalized using geometric and quartile normalization) a list of RNAs found in common was generated.

Among the 980 RNAs found in common, 58 TOP RNAs **(Table 14)** were selected by filtering for a stringent enrichment score (**-log2 Fold Change > 1.5**) between immunoprecipitated and input. These list of predicted RNAs bound by APE1 (p-value<0.01 was considered as statistically significant) will be also employed in order to validate RIP-seq result.

	IP	INPUT	-log2FC
SNORD116-13	72.2457	0	-inf
MIR221	21.0948	0	-inf
LOC100506125	0,874775	0	-inf
MIR3687	10423.8	5031.77	-4.37268
MIR3648	309.496	42.7413	-2.85622
OR52D1	1.66786	0.242483	-2.78204
MIR612	850.086	150.976	-2.49329
C18orf42	0,895509	0,162174	-2,46517
ASPDH	0,911141	0,181801	-2,32532
GPBAR1	0,890021	0,201756	-2,14123
SNX32	0,921882	0,212446	-2,11748
PGLYRP1	1.27535	0.296553	-2.10454
ODF3B	1.78384	0.43536	-2.0347
PRR25	1.09965	0.273491	-2.00748
SCARNA17	47.8104	12.1262	-1.9792
LINC00482	1.33362	0.340439	-1.96988
MROH5	0,803101	0,205053	-1,96958
LRRC29	1.38985	0.357534	-1.95878
IGFALS	1.20681	0.313766	-1.94344

LOC100129722	1,01895	0,275218	-1,88844
CYP2D6	1.68907	0.460822	-1.87395
ZBTB32	1.50059	0.410849	-1.86885
LOC101928012	0,877419	0,246859	-1,82958
SYNE4	0,777649	0,218795	-1,82954
TMC4	1.44218	0.409115	-1.81767
LENEP	4.08686	1.16539	-1.81018
NRADDP	0,905055	0,260275	-1,79797
SNORD116-27	413.405	121.029	-1.7722
LOC101927310	0,724191	0,214842	-1,75309
HSPB7	0,692839	0,209781	-1,72363
TTYH1	1.18448	0.364538	-1.70011
LINC00939	0,688149	0,21183	-1,69982
GBGT1	0,778062	0,240896	-1,69147
MXRA8	3.47094	1.07996	-1.68435
ISLR	0,700077	0,219845	-1,67103
LOC100507006	1.49488	0.470257	-1.6685
ROPN1L-AS1	2.11249	0.667305	-1.66253
GPR17	1.49431	0.474197	-1.65592
IBA57-AS1	2.72828	0.87469	-1.64115
RBP5	7.78252	2.5188	-1.6275
RXFP4	2.95414	0.956859	-1.62636
SPOCK2	1.26015	0.409317	-1.6223
FOSB	5.33284	1.8214	-1.54986
TERC	1616.51	554.393	-1.5439
LOC101928865	1,23993	0,429148	-1,53071
SEMA4G	1.30355	0.452726	-1.52573
ASB16	5.23649	1.82678	-1.51929
IGFN1	10.3131	3.61827	-1.51111
BLACE	3.17273	1.11363	-1.51046
TCTEX1D4	2.62219	0.921433	-1.50882
LOC101928674	4.34574	1.52784	-1.50811
SMG1P7	25.005	8.79335	-1.50773
LINC00896	6.46334	2.2761	-1.50572
MORN1	2.46161	0.869182	-1.50187
PLIN4	0,743159	0,264117	-1,49249
CRYGS	1.44813	0.516535	-1.48726
OR13H1	5.74983	2.05198	-1.4865
IGSF5	0,812588	0,290421	-1,48438

Table 14: 58 TOP RNAs selected among the 980 RNAs by filtering for a stringent enrichment score (-log2 Fold Change > 1.5) between immunoprecipitated and input.

Differentially expressed genes obtained upon APE1 silencing [71] were compared to the list of 58 RNAs bound by APE1. This comparison showed no relationship between RNAs bound by APE1 and the expression levels of the corresponding genes (**Figures 7, 8**).

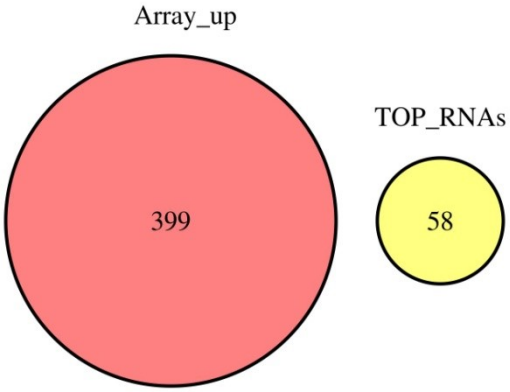


Figure 7: Comparison between the 58 TOP RNAs and the up-regulated genes (399) obtained upon APE1 silencing [71].

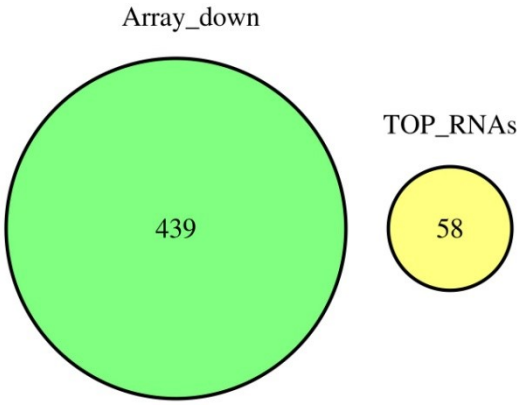
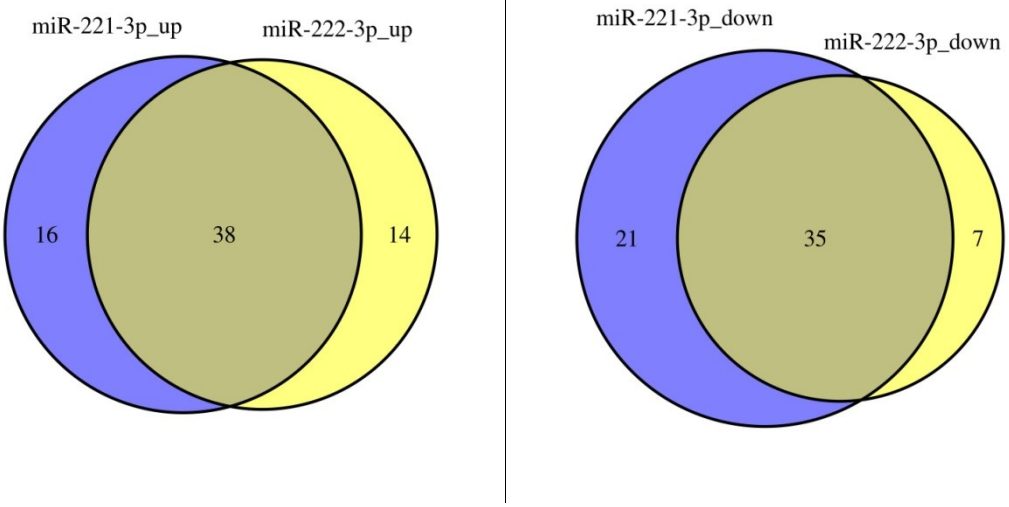


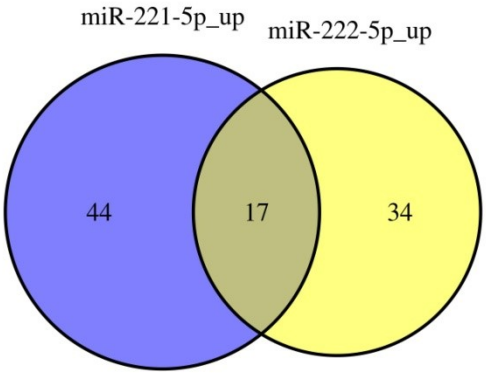
Figure 8: Comparison between the 58 TOP RNAs and the down-regulated genes (439) obtained upon APE1 silencing [71].

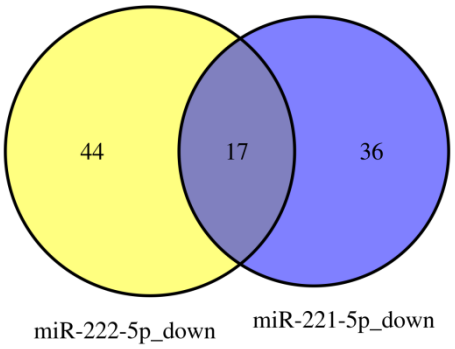
Since we were interested in APE1 post-transcriptional activity on genes involved in chemoresistance, among the 58 TOP RNAs of our list, we focused the attention on miR-221 which has been associated to tumoral progression and in particular with breast cancer chemoresistance. Furthermore, we decided to include in the study also miR-222, because it is encoded in the same genomic cluster as miR-221 and its expression is found to be often coregulated with miR-221 in different type of cancers. miRNA 221 and miRNA 222 are pre-miRNA (stem-loop sequence), processed by the DICER in the cytoplasm, thus producing 2 mature miRNA. miRNA 221 produces the miRNA-221-3p and the miRNA-221-5p. mi-RNA 222 produces the miRNA-222-3p and the miRNA-222-5p.

To find a list of mRNA targets for these two miRNAs, **miRGate**, a curated database of human, mouse and rat miRNA-mRNA targets was used [80]. We then inspected how many of the known miRNAs 221 and 222 target mRNAs are present in the list of APE1-dysregulated genes. Since miRNA 221 and miRNA 222 are paralogues, their common mRNA targets were compared with the list of up and downregulated genes obtained upon APE1 silencing [71] (**Figures 9, 10, 11, 12**).



Figures 9, 10: miRNA-221-3p e miRNA-222-3p common target genes vs up and down-regulated genes. In the first image 38 target genes are in common with the upregulated genes, in the second image 35 target genes are in common with the downregulated genes obtained upon APE1 silencing in a previous study [71].





Figures 11, 12: miRNA-221-5p e miRNA-222-5p common target genes vs up and down-regulated genes. In the first image 17 target genes are in common with the upregulated genes, in the second image 17 target genes are in common wih the downregulated genes obtained upon APE1 silencing in a previous study [71].

Therefore, a good correlation between the mRNA targets of miRNA-221 and miRNA-222 and the genes dysregulated upon APE1 silencing [71] is apparent. These data would suggest that one important mechanism through which APE1 may control gene expression is through its activity on miRNA processing.

RNA candidates to validate (qRT-PCR)

Among the list of predicted 58 TOP RNAs, four RNA candidates were selected (**miR-221, APE1 mRNA itself, FOSB (FBJ murine osteosarcoma viral oncogene homolog B) and TERC (telomerase RNA component) RNA**) and validated through qRT-PCR analysis (**Figure 13**). Primer sequences were designed accordingly to the peak region identified with the sequencing analysis. These selection was based on already established biological association between APE1 protein and the target gene of the above mentioned miRNAs (i.e. PTEN)[81][82], the role of APE1 in its own transcription [83] and on AP-1 function [84], as well as the role of APE1 in telomere maintenance [85].

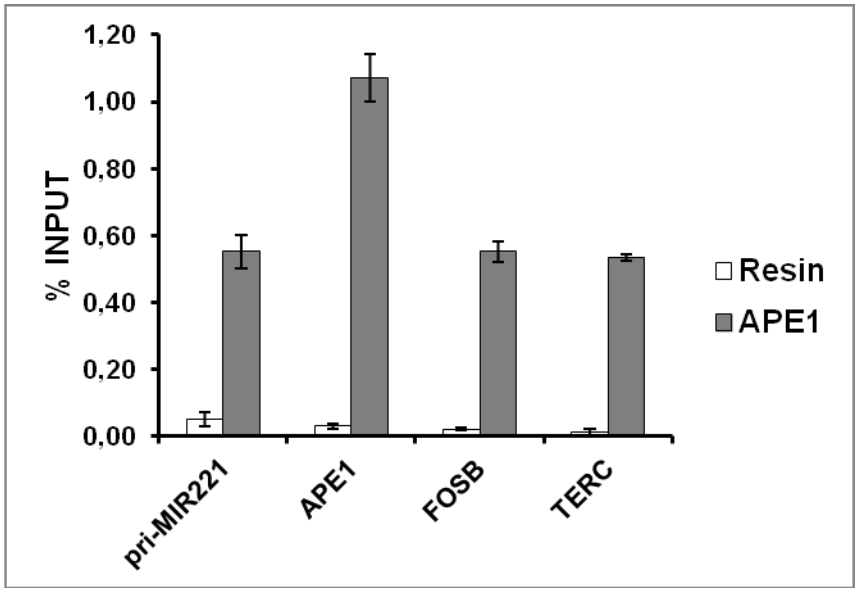


Figure 13: Realtime PCR validation of four RNA targets identified by RIP-seq. In the figure is presented the fold percentage of the amount of immunoprecipitated target RNA relative to that present in total input RNA. Resin, HeLa cell extracts immunoprecipitated with resin not having the anti-FLAG antibody; APE1, immunoprecipitated material of a pool of three replicates.

Gene ontology enrichment analysis

For a more global analysis of APE1 targets, gene ontology analysis was performed uploading on DAVID online tool [86] different lists of the 980 APE1-associated RNAs trying several IP enrichment cut-off to find the best biological enrichment **(adjusted p-value<0.05 was considered as statistically significant).** Among the several enrichment cut-off tested, the minimum FPKM (expression value) useful to find the best biological enrichment was 10. The 347 RNAs, found respecting this criterion, were compared by the analysis tool (DAVID) with RNAs present in particular functional categories in order to identify the enriched categories. APE1-associated RNAs were grouped into annotation clusters (considering the best enrichment score and adjusted p-value) for biological process (**Figure 14**), molecular function (**Figure 15**) and cellular component (**Figure 16**) to determine their functions. In the following pie-chart, the percentage of these RNAs in the TOP functional categories is reported.

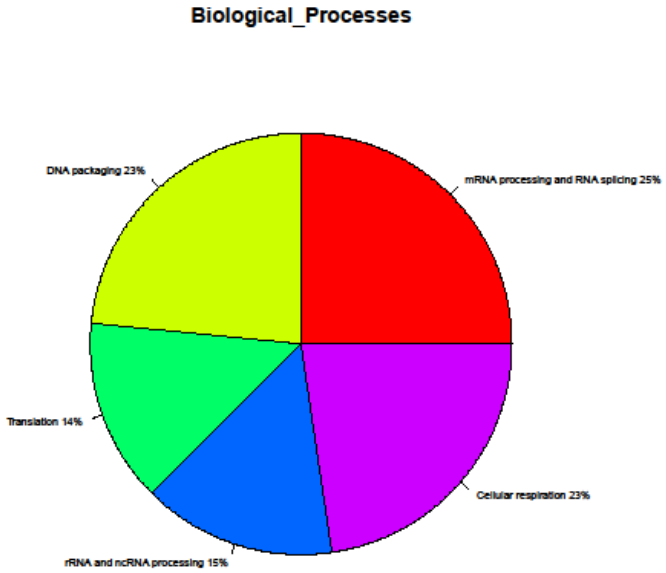


Figure 14: Distribution of APE1-RNA targets in the TOP five functional annotation clusters identified by DAVID enrichment analysis [86] based on Gene Ontology terms of biological processes. The list of RNAs for each of these clusters was curated for an IP enrichment cut-off>10 and an adjusted p-value<0.05.

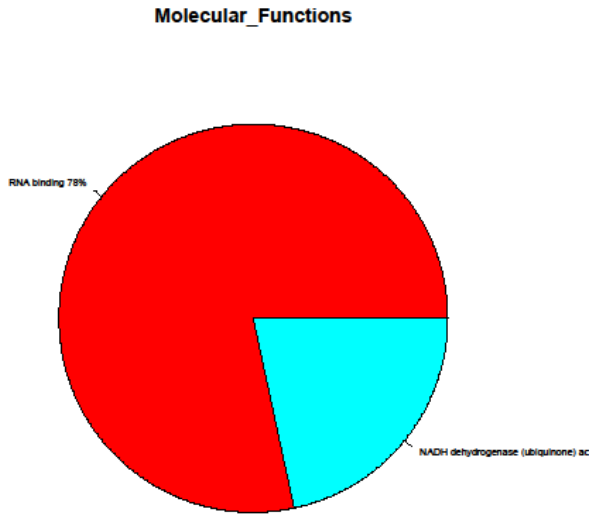


Figure 15: Distribution of APE1-RNA targets in the TOP two functional annotation clusters identified by DAVID enrichment analysis [86] based on Gene Ontology terms of molecular functions. The list of RNAs for each of these clusters was curated for an IP enrichment cut-off>10 and an adjusted p-value<0.05.

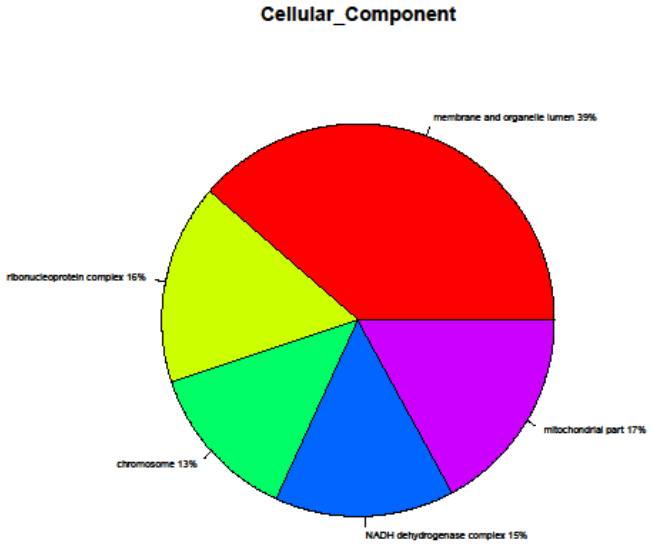


Figure 16: Distribution of APE1-RNA targets in the TOP five functional annotation clusters identified by DAVID enrichment analysis [86] based on Gene Ontology terms of cellular component. The list of RNAs for each of these clusters was curated for an IP enrichment cut-off>10 and an adjusted p-value<0.05.

These functional categories are related to APE1 and also to one another, indeed the majority of the RNAs are found across more than one category. The individual APE1-RNA targets found within the TOP categories (curated for an IP enrichment cut-off>10 and an adjusted p-value<0.05) of biological processes, molecular function and cellular component will be then curated by searching on GeneCards Human Gene Database (http://www.genecards.org/).

Protein interactome network of APE1

In the past, human protein interaction have provided an interesting platform to investigate the functional arrangement of the cells [87][88]. Therefore, data mining-based annotations of this large amount of protein-protein interactions (PPIs) have been established to bring to a more comprehensive understanding of molecular function and biological processes [89]. The aim of STRING database (http://string-db.org) is to make available a crucial evaluation and integration of protein-protein interactions, physical and functional associations are included. Exploiting this known tool, an interaction network between APE1 and its known 106 interactors was generated (proteins interaction network) (**Figure 17**).

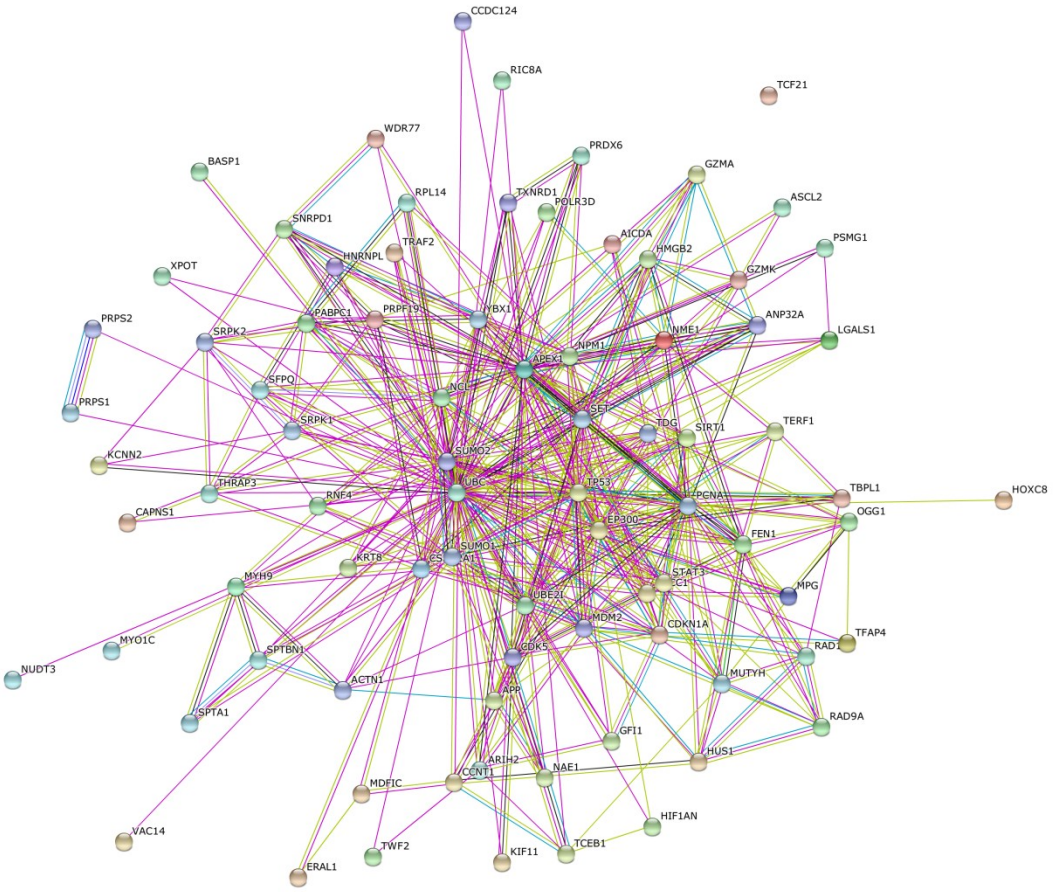


Figure 17: The string network is dedicated to functional associations between proteins [89]. A set of 106 proteins, known to be APE1 interactors, is plotted as a network , different line colors between the proteins represent the particular kinds of evidence for the functional association. Protein nodes bigger show the information of 3D protein structure.

Recent development has pointed out that protein-protein interactions are only a component of the molecular actors in cells, since an expanding list of noncoding RNAs (ncRNAs) are actively involved in multiple biological processes (e.g. fat metabolism, cell death) [90] [91]. A new regulatory RNA system has been discovered by several investigations (RNAs are able to regulate each other by competing for shared ncRNAs) [92] [93]. For example, MD1 (ncRNA) can induce miR-133 and miR135 to regulate the expression of MAML1 and MEF2C acting as a competing endogenous RNA (ceRNA) to control the kinetic of muscle differentiation in human and mouse myoblasts [94]. Therefore, particular attention should be taken on the study of RNA-associated (RNA-RNA/RNA-protein) interaction.

RNA-protein interaction

The APE1 ability to RNA could be mediated by some of the APE1 protein interacting partners. This hypothesis was also strengthened by the observation that interaction of APE1 with NPM1 was destabilized upon treatment with RNAse A [70] [95]. In literature it was reported that some of the 106 proteins bound by APE1 (e.g. NPM1, STAT3, etc) also bind RNAs; starting by this, our aim was to understand if among the 980 RNAs bound by APE1, here could be some recognized also by some of the known APE1 protein-interacting partners. By using an interesting online free tool called RAID [96], a screening of all 106 proteins bound by APE1 was done and we found that 3 proteins (SFPQ, p53, PCNA) interact with 3 out of 980 RNAs of our list (in order NEAT1, H19, TERC). The results of this preliminary work, reported in the following table **(Table 14),** open thus the question of whether, the RNAs bound by APE1 plays a biological role in modulating the APE1 protein interactome network. These data suggest that APE1 regulates the gene expression through post-transcriptional mechanisms mediated by ncRNAs.

Proteins	RNA-BINDING	RNAs	RNAs of our list (980 RNAs)	PMID
ACTN1	NO			
AID	NO			
ANP32A	NO			
ANP32C	NO			
AP-4	NO			
APP	SI	BACE1-AS	NO	18587408
ARIH2	NO			
ASCL2	NO			
BASP1	NO			
CAPNS1	NO			
CCDC124	NO			
Cdk5	NO			
CKII	NO			
DNA Lig I	NO			
ERAL1	NO			
FEN1	NO			
GAL1	NO			
GzmA	NO			
GzmK	NO			
hAPN	NO			
HDACs	NO			
HHV8GK18_gp81	NO			
HIF-1 alpha	NO			
HMG2	NO			
HMGA-1	NO			
HMGA-2	NO			
hMYH	NO			
hnRNP-F	NO			
HnRNP-H	NO			
hnRNP-K	SI	p21	NO	20673990
hnRNP-L	NO			
hnRNP-U	SI	GADD45A, HEXIM1, HOXA2, IER3, NHLH2, TNFalpha, ZFY	NO	17174306
hnRNP-UL1	NO			
HOX3	NO			
hS3	NO			
Hsp70-1	NO			
Hus1	NO			
K2C8	NO			
KIF11	NO			
KRT8	NO			
Ku Antigen p70	NO			
Ku Antigen p80	NO			

MDM2	SI	TP53	NO	19106616
MEP50	NO			
MOES	NO			
MPG	NO			
MYH9	NO			
MYO1C	NO			
NAE1	NO			
NCL	SI	RNY1, RNY3, BCL2, BCL2L1, IL2, LINE-1, REN, SLS2-A7, MBII-52	NO	
NM23-H1	NO			
NPM1	SI		NO	
NUDT3	NO			
OGG1	NO			
p21	NO			
p300	NO			
p53	SI/NO	MEG3, TP53TG1, MIR34A, MIR34C, PLAU, PLAUR, SERPINE1	H19	23222637
PABP1	NO	BCYRN1, BC1	NO	
PCNA	SI		TERC	17932748
Pol b	NO			
POLR3D	NO			
PRDX6	NO			
PRP19	NO			
PRPS1	NO			
PRPS2	NO			
PSMG1	NO			
Rad1	NO			
Rad9	NO			
Rev	SI	Rev response element (RRE)	NO	8703216
RIC8A	NO			
RL14	NO			
RL3	NO			
RL4	NO			
RLA0	NO			
RNF4	NO			
RSSA	NO			
SET	NO			
SFPQ	SI		NEAT1	19720872
SIRT1	NO			
SK2	NO			
SMD1	NO			
SPTB2	NO			

SRPK1	NO			
SRPK2	NO			
STAT3	SI		NO	
SUMO1	NO			
SUMO2	NO			
TCEB1	NO			
TCF21	NO			
TCP1-alpha	NO			
TCPA	NO			
TDG	NO			
TERF21P	NO			
THRAP3	NO			
TRAF2	NO			
TRF1	NO			
TRF2	NO			
TRX	NO			
TWF2	NO			
TXNRD1	NO			
Ubc9	NO			
Ubiquitin C	NO			
XPOT	NO			
XRCC1	NO			
YB1	NO			

Table 14: List of 106 proteins (known interactors of APE1). RNA-protein association between these proteins and the 980 RNAs of our list was investigated.

5 APPENDIX

In the following table the complete list of APE1 RNA-target genes in HeLa cells under basal conditions is reported. The corresponding chromosome, gene symbol and type are indicated (157 genes).

Chromosome	Symbol_gene	Type
2	ABC7-43041300I9.1	Pseudo gene
2	AC009299.4	Pseudo gene
5	AC010455.1	miRNA
22	AC011718.2	lincRNA
2	AC018867.1	protein_coding
7	AC027269.2	lincRNA
16	AC092291.1	Pseudo gene
4	AC118282.1	miRNA
5	ADAMTS12	protein_coding
4	AF146191.4	lincRNA
1	AL583842.2	miRNA
1	AL732363.1	miRNA
1	ANKRD20A14P	unprocessed_pseudogene
19	ARHGEF18	protein_coding
5	C5orf34	protein_coding
5	C6	protein_coding
5	C9	protein_coding
8	CCAT1	lincRNA
17	CDC27	protein_coding
2	CDC27P1	Pseudo gene
5	CDH10	protein_coding
5	CDH12	protein_coding
5	CDH18	protein_coding
5	CDH6	protein_coding
5	CDH9	protein_coding
1	CEP170	protein_coding
11	CEP57	protein_coding
9	CNTRL	protein_coding
9	CR786580.1	miRNA
5	CTD-2010I22.2	lincRNA
5	CTD-2057J6.1	Pseudo gene
5	CTD-2066L21.3	lincRNA
5	CTD-2116N24.1	lincRNA
5	CTD-2118P12.1	lincRNA
5	CTD-2134P3.1	lincRNA
5	CTD-2151L9.2	Pseudo gene
5	CTD-2201E9.1	lincRNA
5	CTD-2218G20.2	lincRNA
5	CTD-2234B20.1	lincRNA
5	CTD-2272G21.3	Pseudo gene
17	CTD-2303H24.2	retained_intron
5	CTD-2306M5.1	lincRNA

5	CTD-2353F22.1	Antisense
16	CTD-2522B17.8	Pseudo gene
5	CTD-2533K21.4	lincRNA
5	CTD-2636A23.2	Antisense
5	CTD-3007L5.1	lincRNA
5	DAB2	protein_coding
5	DNAH5	protein_coding
5	DROSHA	protein_coding
Y	DUX4L16	Pseudo gene
Y	DUX4L17	Pseudo gene
16	FAM157C	processed_pseudogene
22	FAM230B	processed_pseudogene
5	FBXO4	protein_coding
4	FRG1	protein_coding
5	FYB	protein_coding
5	GUSBP1	processed_pseudogene
1	HMCN1	protein_coding
14	IGHV1-68	IG_V_pseudogene
22	IGKV1OR22-5	IG_V_pseudogene
8	KCTD9	protein_coding
5	KIAA0947	protein_coding
12	KLRC3	protein_coding
22	LA16c-83F12.6	lincRNA
5	LIFR-AS1	Antisense
10	LINC00843	lincRNA
10	LIPA	processed_pseudogene
5	LMBRD2	protein_coding
5	MARCH11	protein_coding
5	MARCH6	processed_pseudogene
1	MIA3	protein_coding
9	MIR1299	miRNA
5	MTRR	processed_pseudogene
5	NADK2	protein_coding
1	NBPF1	processed_pseudogene
1	NBPF10	protein_coding
1	NBPF14	protein_coding
1	NBPF16	protein_coding
1	NBPF24	nonsense_mediated_decay
1	NBPF8	protein_coding
1	NBPF9	processed_pseudogene
3	NCK1	protein_coding
20	NCOR1P1	processed_pseudogene
5	NIPBL	processed_pseudogene
5	NIPBL	protein_coding
5	NNT	protein_coding
1	NOTCH2	protein_coding
1	NOTCH2NL	protein_coding
5	NPR3	protein_coding
5	NSUN2	processed_pseudogene
11	OR4C5	protein_coding
6	OR4F7P	Pseudo gene

5	OSMR	protein_coding
5	OXCT1	protein_coding
8	PCMTD1	protein_coding
1	PDE4DIP	protein_coding
3	PIK3CB	protein_coding
1	PLA2G4A	protein_coding
5	PMCHL1	processed_pseudogene
5	PRKAA1	protein_coding
8	RAB2A	protein_coding
5	RICTOR	protein_coding
8	RN7SL250P	misc_RNA
5	RNA5SP177	rRNA
18	ROCK1P1	processed_pseudogene
5	RP1-137K24.1	lincRNA
5	RP1-167G20.1	lincRNA
5	RP11-113I22.1	lincRNA
8	RP11-1195F20.7	Pseudo gene
5	RP11-122C5.3	Antisense
5	RP11-1250I15.3	lincRNA
4	RP11-1281K21.6	Pseudo gene
7	RP11-1324A7.2	processed_pseudogene
1	RP11-14N7.2	lincRNA
5	RP11-192H6.2	lincRNA
9	RP11-262H14.1	lincRNA
1	RP11-277L2.5	lincRNA
12	RP11-313F23.4	lincRNA
9	RP11-318K12.3	Pseudo gene
5	RP11-321E2.6	processed_pseudogene
1	RP11-353N4.1	lincRNA
1	RP11-353N4.5	lincRNA
1	RP11-417J8.1	lincRNA
1	RP11-417J8.2	lincRNA
5	RP11-42L13.2	Pseudo gene
1	RP11-435B5.3	lincRNA
1	RP11-435B5.5	lincRNA
5	RP11-454P21.1	lincRNA
4	RP11-463J17.1	lincRNA
5	RP11-473L15.2	Antisense
5	RP11-480D4.6	lincRNA
5	RP11-53O19.1	Antisense
5	RP11-549K20.1	lincRNA
11	RP11-56P9.5	Pseudo gene
16	RP11-626K17.5	unprocessed_pseudogene
1	RP11-763B22.9	unprocessed_pseudogene
9	RP11-764K9.1	lincRNA
9	RP11-764K9.4	unprocessed_pseudogene
5	RP11-774D14.1	lincRNA
1	RP11-782C8.2	lincRNA
5	RP11-855C21.1	sense_intronic
5	RPL32P14	Pseudo gene
5	RPL36AP21	Pseudo gene

1	SEC22B	processed_pseudogene
10	SIRT1	protein_coding
5	SPEF2	protein_coding
1	SRGAP2C	unprocessed_pseudogene
5	TARS	protein_coding
5	UBE2V1P12	Pseudo gene
5	UBL5P1	Pseudo gene
4	UGT2A1	processed_pseudogene
8	USP17L8	Pseudo gene
11	USP47	protein_coding
22	XXbac-B33L19.3	Antisense
5	ZDHHC11	protein_coding
5	ZFR	processed_pseudogene

In the table below the complete list of APE1 RNA-target genes in HeLa cells upon H_2O_2-treatment is reported. The corresponding chromosome, gene symbol and type are indicated (209 genes).

Chromosome	Symbol_gene	Type
5	AC004237.1	Antisense
7	AC006159.3	lincRNA
22	AC008079.9	Antisense
2	AC009299.4	Pseudo gene
2	AC027612.3	Pseudo gene
2	AC097374.2	processed_pseudogene
5	AC106771.1	miRNA
4	AC118282.1	miRNA
4	AC118282.2	miRNA
4	AC118282.3	miRNA
17	AC126365.1	transcribed_unprocessed_pseudogene
5	ADAMTS12	protein_coding
5	ADCY2	processed_pseudogene
8	AF228730.1	miRNA
10	AL031601.4	Pseudo gene
1	AL121985.1	Pseudo gene
9	AL353626.1	miRNA
9	AL353626.2	miRNA
9	AL353763.1	miRNA
9	AL353763.2	miRNA
20	AL441988.1	miRNA
1	AL583842.1	miRNA
1	AL583842.2	miRNA
1	AL645608.2	protein_coding
1	AL732363.1	miRNA
1	ANKRD20A12P	transcribed_unprocessed_pseudogene
1	ANKRD20A14P	unprocessed_pseudogene
2	ANKRD36	protein_coding

2	ANKRD36	retained_intron
21	BAGE2	processed_pseudogene
9	BMS1P11	unprocessed_pseudogene
5	C6	protein_coding
5	C7	protein_coding
9	CBWD7	protein_coding
8	CCAT1	lincRNA
17	CDC27	protein_coding
2	CDC27P1	Pseudo gene
Y	CDC27P2	Pseudo gene
5	CDH12	protein_coding
5	CDH18	protein_coding
5	CDH9	protein_coding
9	CDKN2B-AS1	Antisense
8	CHD7	protein_coding
21	CR381653.1	miRNA
21	CR381670.1	miRNA
21	CR392039.1	miRNA
21	CR392039.2	miRNA
9	CR786580.1	miRNA
9	CR848007.2	Pseudo gene
20	CST9	protein_coding
7	CTA-298G8.2	Pseudo gene
Y	CTBP2P1	Pseudo gene
5	CTC-305H11.1	lincRNA
5	CTD-2061E9.1	lincRNA
5	CTD-2113L7.1	Antisense
5	CTD-2127O16.2	Pseudo gene
5	CTD-2143L24.1	lincRNA
5	CTD-2161F6.2	lincRNA
5	CTD-2194L12.2	lincRNA
5	CTD-2194L12.3	lincRNA
5	CTD-2272G21.3	processed_pseudogene
5	CTD-2306M5.1	lincRNA
14	CTD-2311B13.7	lincRNA
5	CTD-2318H23.1	lincRNA
16	CTD-2522B17.8	Pseudo gene
5	CTD-2533K21.4	lincRNA
5	CTD-3007L5.1	lincRNA
5	DAB2	protein_coding
3	DNAH12	protein_coding
5	DNAH5	protein_coding
5	DROSHA	protein_coding
Y	DUX4L16	Pseudo gene
Y	DUX4L17	Pseudo gene
Y	DUX4L18	Pseudo gene
Y	DUX4L19	Pseudo gene
5	EGFLAM-AS2	Antisense
5	EGFLAM-AS4	Antisense
15	ELMO2P1	transcribed_unprocessed_pseudogene
16	FAM157C	processed_pseudogene

20	FAM182A	lincRNA
22	FAM230A	protein_coding
22	FAM230B	processed_pseudogene
10	FAM24A	protein_coding
4	FRG1	protein_coding
20	FRG1B	nonsense_mediated_decay
20	FRG1B	protein_coding
5	GHR	nonsense_mediated_decay
15	HERC2P3	processed_pseudogene
14	IGHV1-68	IG_V_pseudogene
2	IGKV1OR-1	IG_V_pseudogene
5	IL7R	processed_pseudogene
22	KB-1183D5.13	lincRNA
7	KMT2C	protein_coding
22	LA16c-83F12.6	lincRNA
3	LINC00969	lincRNA
11	MICALCL	protein_coding
9	MIR1299	miRNA
4	MLLT10P2	Pseudo gene
15	MYO1E	protein_coding
5	NADK2	protein_coding
15	NBEAP1	transcribed_unprocessed_pseudogene
1	NBPF1	protein_coding
1	NBPF10	protein_coding
1	NBPF12	protein_coding
1	NBPF14	protein_coding
1	NBPF20	protein_coding
1	NBPF8	protein_coding
1	NBPF9	protein_coding
5	NNT	protein_coding
1	NOTCH2NL	protein_coding
17	NSF	protein_coding
14	OR4N2	protein_coding
Y	PABPC1P5	Pseudo gene
8	PCMTD1	protein_coding
5	PLCXD3	protein_coding
1	PPIAL4F	Pseudo gene
8	PXDNL	protein_coding
9	RABGAP1	protein_coding
5	RANBP3L	protein_coding
5	RICTOR	protein_coding
2	RNA5SP100	rRNA
5	RNA5SP177	rRNA
X	RNA5SP503	rRNA
1	RNA5SP60	rRNA
10	RNU2-42P	snRNA
5	RNU6-738P	snRNA
18	ROCK1P1	Pseudo gene
9	RP11-111F5.3	lincRNA
5	RP11-122F24.1	lincRNA
4	RP11-1281K21.6	Pseudo gene

4	RP11-1281K21.7	processed_pseudogene
7	RP11-1324A7.2	processed_pseudogene
17	RP11-1407O15.2	protein_coding
5	RP11-152K4.2	Antisense
5	RP11-184E9.1	lincRNA
5	RP11-192H6.2	lincRNA
4	RP11-241F15.3	Pseudo gene
9	RP11-262H14.1	lincRNA
9	RP11-318K12.3	Pseudo gene
5	RP11-321E2.6	processed_pseudogene
9	RP11-350D23.4	Pseudo gene
1	RP11-353N4.5	lincRNA
5	RP11-360I2.1	lincRNA
1	RP11-417J8.1	lincRNA
1	RP11-417J8.2	lincRNA
1	RP11-417J8.3	lincRNA
1	RP11-417J8.6	lincRNA
3	RP11-423E7.1	Pseudo gene
1	RP11-423O2.5	lincRNA
1	RP11-423O2.7	lincRNA
5	RP11-42L13.2	Pseudo gene
5	RP11-432M8.13	Pseudo gene
1	RP11-435B5.5	lincRNA
1	RP11-435B5.6	lincRNA
1	RP11-435B5.7	lincRNA
5	RP11-447B18.1	lincRNA
4	RP11-463J17.1	lincRNA
5	RP11-480D4.5	Pseudo gene
5	RP11-484L7.1	Pseudo gene
5	RP11-53O19.1	Antisense
5	RP11-53O19.2	lincRNA
6	RP11-552E20.1	lincRNA
11	RP11-56P9.4	unprocessed_pseudogene
11	RP11-56P9.5	Pseudo gene
5	RP11-589F5.4	Pseudo gene
16	RP11-626K17.5	unprocessed_pseudogene
16	RP11-67H24.2	lincRNA
5	RP11-730N24.1	lincRNA
5	RP11-730N24.2	lincRNA
1	RP11-763B22.9	unprocessed_pseudogene
9	RP11-764K9.1	lincRNA
9	RP11-764K9.4	unprocessed_pseudogene
5	RP11-774D14.1	lincRNA
1	RP11-782C8.1	lincRNA
1	RP11-782C8.4	lincRNA
1	RP11-782C8.5	lincRNA
10	RP11-96F8.1	Pseudo gene
1	RP3-395P12.2	lincRNA
20	RP4-610C12.1	Antisense
1	RP4-669L17.10	lincRNA
1	RP5-968D22.3	lincRNA

5	RPL36AP21	Pseudo gene
7	SAMD9	protein_coding
1	SEC22B	processed_pseudogene
1	SELP	protein_coding
5	SEMA5A	protein_coding
5	SEPP1	protein_coding
11	SESN3	protein_coding
10	SIRT1	protein_coding
5	SLC1A3	protein_coding
22	SLC9B1P4	Pseudo gene
16	SMG1P1	transcribed_unprocessed_pseudogene
5	SNORD81	snoRNA
7	SP4	protein_coding
5	SPEF2	protein_coding
5	TARS	protein_coding
4	TMEM128	protein_coding
5	TTC23L	nonsense_mediated_decay
5	UBE2D2	protein_coding
5	UBL5P1	Pseudo gene
1	VAMP4	protein_coding
5	WDR70	protein_coding
1	WI2-3658N16.1	transcribed_unprocessed_pseudogene
17	WIPI1	processed_pseudogene
22	XXbac-B33L19.3	Antisense
1	ZBTB41	protein_coding
5	ZNF131	processed_pseudogene
5	ZNF622	protein_coding
3	ZNF717	processed_pseudogene

6 BIBLIOGRAPHY

[1] **Metzker M.L.** *Sequencing technology – the next generation*. Nature Reviews Genetics 11, 31-46.(2010)

[2] **Hawkins R.D, Hon G.C, and Ren B.** *Next generation genomics: an integrative approach.* Nature reviews Genetics 11.(2010)

[3] **Han1 J, Xiong J, Wang1 D, and Fu X.D.** *Pre-mRNA splicing: where and when in the nucleus*. Trends in Cell Biology 21, 6.(2011)

[4] **Amit Nagal.** *Role of epigenetics in Cancer Genomics*. Chapter 11, GVK Biosciences Private Limited, India.(2015)

[5] **Esteller M.** *Cancer epigenomics: DNA methylomes and hystone-modification maps*. Nature Review Genetics 8, 286-298. (2007)

[6] **Jones P A, and Baylin S B.** *The Epigenomics of Cancer.* Cell 128, 683-692.(2007)

[7] **Latchman D S**. *Eukaryotic Transcription Factors*. fifth edition. Elsevier Ltd.(2008)

[8] **Johnson D S, Mortazavi A, Myers R M, and Wold B.** *Genome-wide mapping of in vivo protein-DNA interactions*. Science 316, 1497–1502.(2007)

[9] **Ku C.S, Naidoo N, Wu M, Soong R.** *Studying the epigenome using next generation sequencing*. J Med Genet 48: 721–730.(2011)

[10] **Shones D.E. & Zhao K.** *Genome-wide approaches to studying chromatin modifications*. Nature Reviews Genetics 9, 179-191.(2008)

[11] **Park P.J.** *ChIP-Seq: advantages and challenges of a maturing technology*. Nat Rev Genet. 10, 669-680.(2009)

[12] **Hoffman B G et al.** *Genome-wide identification of DNA–protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing*. Journal of Endocrinology 201 1–13.(2009)

[13] **Shendure J, & Ji H**. *Next generation DNA sequencing*. Nature Biotechnology 26, 10.(2008)

[14] **Feng J. et al.** *Identyfing ChIP-seq enrichment using MACS*. Nat Protoc. 7, 9. (2012)

[15] **Duan J.** *Computational analysis of ChIP-Seq data*. AARHUS UNIVERSITY.(2010)

[16] **Landt S.G, Marinov G.K, Kundaje A, Kheradpour P, Pauli F, et al**. *ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia*. Genome Res 22: 1813–1831.(2012)

[17] **Chen Y, Negre N.Li Q, Mieczkowska J.O, Slattery M, et al.** *Systematic evaluation of factors influencing ChIP-seq fidelity*. Nat Methods 9: 609–614.(2012)

[18] **Bailey T. et al.** *Pratical Guidelines for the Comprehension Analysis of ChIP-Seq Data*. PLOS Computational Biology 9, 11.(2013)

[19] **Kharchenko P.V, Tolstorukov M.Y, & Park P.J.** *Design and analysis of ChIP-seq experiments for DNA-binding proteins*. Nature Biotechnology 26, 12.(2008)

[20] **Tell G. et al.** *The Many Functions of APE1/Ref1: Not Only a DNA Repair Enzyme*. Antioxid Redox Signal 11(3), 601-619. (2009)

[21] **Xanthoudakis S, Smeyne R.J, Wallace J.D, and Curran T.** *The redox/DNA repair protein, Ref-1, is essential for early embryonic development in mice*. Proc Natl Acad Sci U A 93, 8919–23.(1996)

[22] **Antoniali G. et al.** *SIRT1 gene expression upon genotoxic damage is regulated by APE1 through nCaRE-promoter elements*. Mol Biol Cell 25(4), 532-547.(2014)

[23] **Kelley M.R. et al.** *APE1/Ref1 Role in Redox Signaling: Translational Applications of Targeting the Redox function of the DNA Repair/Redox Protein APE1/Ref-1*. Curr Mol Pharmacol 5(1) 36-53.(2012)

[24] **Wilson D.M. 3rd & Simeonov A.** Small molecule inhibitors of DNA repair nuclease activities of APE1. Cell Mol Life Sci. 67(21), 3621-31. (2010)

[25] **Moore M.J, Zhang C, Gantman E.C, Mele A, Darnell J.C, Darnell R.B.** *Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis*. Nat Protoc. 9, 2, 263–293.(2014)

[26] **Konig J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner D J, Luscombe N M, Ule J.** *iCLIP -Transcriptome-wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution.* J. Vis. Exp. (50).(2011)

[27] **Murigneux V. et al.** *Transcriptome-wide identification of RNA binding sites by CLIP-seq*. Methods 63(1), 32-40. (2013)

[28] **Zambelli F. & Pavesi G.** *Rip-Seq Data Analysis to Determine RNA-Protein Associations*. RNA Bioinformatics, Volume 1269 of the series Methods in Molecular Biology, Chapter 18 , pp 293-303. (2014)

[29] **Qu H, Fang X.** *A Brief review on the Human Encyclopedia of DNA Elements (ENCODE) Project*. Genomics Proteomics Bioinformatics 11, 135-141.(2013)

[30] **Bolger A.M, Lohse M, and Usadel B.** *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics 30, 15.(2014)

[31] **Langmead B, Trapnell C, Pop M, and Salzberg S.L**. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biology 10, R25 (2009)

[32] **Langmead B, & Salzberg *S.L.*** *Fast gapped-read alignment with Bowtie 2.* Nature Methods 9, 357–359.(2012)

[33] **Li H, and Durbin R.** *Fast and accurate short read alignment with Burrows-Wheeler Transform*. Bioinformatics, 25, 1754-60.(2009)

[34] **Rui Wang, et al.** *LOcating Non-Unique marche Tags (LONUT) to improve the detection of the enriched regions for ChIP-seq data*. PLoS ONE 8, 6.(2013)

[35] **Szalkowski A.M.** *Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts*. Brief Bioinform 12, 626-633. (2011).

[36] **Zhang et al.** *Model-based Analysis of ChIP-seq (MACS).* Genome Biology vol.9, 9, pp. R137. (2008)

[37] **Kim H.** *A short survey of computational analysis methods in analysing ChIP-seq data*. Hum Genomics 5(2), 117-123.(2011)

[38] **Feng J.** *Identifying ChIP-seq enrichment using MACS*. Nat Protoc 7(9).(2012)

[39] **Rozowsky J.** *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls*. Nat Biotechnol 27, 66-75. (2009)

[40] **Bardet A.F.** *A computational pipeline for comparative ChIP-seq analyses*. Nat Protoc 7, 45-61.(2011)

[41] **Li Q.** *Measuring reproducibility of highthroughput experiments*. Ann Appl Stat 5, 1752-1779.(2011)

[42] **Thorvaldsdòttir H, Robinson J.T, and Mesirov J.P.** *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration*. Briefings in Bioinformatics 14, 178-192. (2012)

[43] **Sanger Institute.** *GFF: an Exchange Format for Feature Desciption.* **http://www.sanger.ac.uk/resources/sotware/gff/**

[44] **UCSC Genome Bioinformatics.** *BED Format.* http://genome.ucsc.edu/FAQ/FAQformat.html#format1

[45] **UCSC Genome Bioinformatics.** *Wiggle Track Format (WIG).* http://genome.ucsc.edu/goldenPath/help/wiggle

[46] **Kent W.J, Zweig A.S, Barber G, et al.** *BigWig and BigBed: enabling browsing of large distributed datasets.* Bioinformatics 26, 2204-2207.(2010)

[47] **Sales G, Calura E, Cavalieri D, and Romualdi C.** *graphite – a bioconductor package to convert pathway topology to gene network***.** BMC Bioinformatics 13, 20.(2012)

[48] **Beltrame L. et al.** *The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways*. Bioinformatics 27(15), 2127-2133.(2011)

[49] **Li E, & Davidson E.H.** *Building developmental gene regulatory networks*. Birth Defects Res. C Embryo Today 87, 123-130.(2009)

[50] **Wang E, et al.** *Cancer systems biology: exploring cancer-associated genes on cellular networks*. Cell Mol. Life Sci. 64, 1752-1762.(2007)

[51] **Kanehisa M, & Goto S.** *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Res 28, 27–30.(2000)

[52] **Matthews L. et al.** *Reactome knowledgebase of human biological pathways and processes*. Nucleic Acids Res 37, 619–622.(2009)

[53] **Schaefer C.F, et al.** *PID: The Pathway Interaction Database*. Nucleic Acids Res 37, D674-9.(2009)

[54] **Caspi R, et al.** *The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc collection of Pathway/Genome Databases*. Nucleic Acids Research 38, D473-9.(2010)

[55] **Mi H, et al.** *Panther in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of pylogenetic trees*. Nucleic Acids Res.(2012)

[56] **Giancarlo R, Lo Bosco G, and Pinello L.** *Distance functions, clustering algorithms and microarray data analysis*. Lect. Notes. Comp. Sci. 6073, 125–138.(2010)

[57] **Giancarlo R, Lo Bosco G, Pinello L, and Utro F.** *A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis*. BMC Bioinformatics 14 Suppl 1, S6.(2013)

[58] **Jaskowiak P. A, Campello R. J, and Costa I. G.** *On the selection of appropriate distances for gene expression data clustering*. BMC Bioinformatics 15 Suppl 2, S2.(2014)

[59] **Wang K, Fain B, Levitt M, and Samudrala R.** *Improved protein structure selection using decoy-dependent discriminatory functions*. BMC Struct. Biol. 4, 8.(2004)

[60] **Fogolari F, Tosatto S. C, and Colombo G.** *A decoy set for the thermostable sub domain from chicken villin headpiece, comparison of different free energy estimators*. BMC Bioinformatics 6, 301.(2005)

[61] **Yang X, Bentink S, Scheid S, and Spang R**. *Similarities of ordered gene lists*. J Bioinform Comput Biol 4, 693–708.(2006)

[62] **Campello R. J. G. B, and Hruschka E. R.** *On comparing two sequences of numbers and its applications to clustering analysis*. Inf. Sci. 179, 1025–1039.(2009)

[63] **Storey J.D, Tibshirani R**. *Statistical significance for genomewide studies*. Proc Natl Acad Sci USA 100, 9440-9445.(2003)

[64] **Mercier E. et al.** *An Integrated Pipeline for the Genome-Wide Analysis of Transcription Factor Binding Sites from ChIP-seq*. PLoS ONE 6(2).(2011)

[65] **Kersey P.J, et al.** *Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species*. Nucleic Acids Res. 40, D91-D97.(2012)

[66] **Martini P, Sales G, Massa S M, Chiogna M, and Romualdi C.** *Along signal paths: an empirical gene set approach exploiting pathway topology.* Nucleic Acids Research 41, 1.(2012)

[67] **Smoot M.E, Ono K, Ruscheinski J, Wang P.L, Ideker T.** *Cytoscape 2.8: new features for data integration and network visualization.* Bioinformatics *27*, 3, 431–432.(2011)

[68] **Chiaretti S, Li X, Gentleman R, Vitale A, Wang K.S, Mandelli F, Fo R, and Ritz J**. *Gene Expression Profiles of B-lineage adult acute lymphocytic leukemia reveal genetic patterns that identify lineage derivation and distinct mechanisms of transformation*. Clin. Cancer Res, 11, 7209-7219. (2005)

[69] **Levav-Cohen Y, Goldberg Z, Zuckerman V, Grossman T, Haupt S, and Haupt Y.** *C-Abl as a modulator of p53*. Biochem. Biophys. Res. Commun 331, 737-749. (2005)

[70] **Vascotto C, Fantini D, Romanello M, Cesaratto L, Deganuto M, Leonardi A, Radicella J P, Kelley M R, D'Ambrosio C, Scaloni A, *et al.*** *APE1/Ref-1 interacts with NPM1 within nucleoli and plays a role in the rRNA quality control process*. Mol Cell Biol 29, 1834–54. (2009)

[71] **Vascotto C, Cesaratto L, Zeef L A, Deganuto M, D'Ambrosio C, Scaloni A, Romanello M, Damante G, Taglialatela G, Delneri D, et al.** *Genome-wide analysis and proteomic studies reveal APE1/Ref-1 multifunctional role in mammalian cells*. Proteomics, 9, 1058–74. (2009)

[72] **Bolstad B.M, Irizarry R.A, Astrand M, Speed T.P.** *A comparison of normalization methods for high density oligonucleotide array data based on bias and variance*. Bioinformatics 19, 185-193. (2003)

[73] **Baldi P, Long A.D.** *A bayesian frame work for the analysis of microarray expression data: regularized t-Test and statistical inferences of gene changes*. Bioinformatics 17, 509-519.(2001)

[74] **Tell G. & Demple B.** *Base excision DNA repair and cancer.* Oncotarget 6(2), 584-5.(2015)

[75] **Vezzi F, Del Fabbro C, Tomescu A I, and Policriti A.** *rNA: a fast and accurate short reads numerical aligner.* Bioinforma. Oxf. Engl. 28, 123–124.(2012)

[76] **Martin M.** *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal 17, pp. 10–12.(2011)

[77] **Trapnell C, Pachter L, and Salzberg S L.** *TopHat: discovering splice junctions with RNA-Seq.* Bioinforma. Oxf. Engl. 25, 1105–1111.(2009)

[78] **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D R, Pimentel H, Salzberg S L, Rinn J L, and Pachter L.** *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat. Protoc. 7, 562–578.(2012)

[79] **Trapnell C, Hendrickson D G, Sauvageau M, Goff L, Rinn J L, and Pachter L.** *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nat. Biotechnol. 31, 46–53.(2013)

[80] **Andrès-Leòn E, et al.** *miRgate: a curated database of human, mouse and rat miRNA-mRNA targets.* Database (Oxford). (2015)

[81] **Chun-Zhi Z, Lei H, An-Ling Z, Yan-Chao F, Xiao Y, Guang-Xiu W, Zhi-Fan J, Pei-Yu P, Qing-Yu Z, and Chun-Sheng K.** *MicroRNA-221 and microRNA-222 regulate gastric carcinoma cell proliferation and radioresistance by targeting PTEN*. BMC Cancer, 10, 367. (2010)

[82] **Zhao G, Cai C, Yang T, Qiu X, Liao B, Li W, Ji Z, Zhao J, Zhao H, Guo M, *et al.*** *MicroRNA-221 induces cell survival and cisplatin resistance through PI3K/Akt pathway in human osteosarcoma*. PloS One, 8, e53906. (2013)

[83] **Fantini D, Vascotto C, Deganuto M, Bivi N, Gustincich S, Marcon G, Quadrifoglio F, Damante G, Bhakat K.K, Mitra S, et al.** APE1/Ref-1 regulates PTEN expression mediated by Egr-1. Free Radic Res, 42, 20–9. (2008)

[84] **Ando K, Hirao S, Kabe Y, Ogura Y, Sato I, Yamaguchi Y, Wada T, and Handa H.** *A new APE1/Ref-1-dependent pathway leading to reduction of NF-kappaB and AP-1, and activation of their DNA-binding activity.* Nucleic Acids Res, 36, 4327–36. (2008)

[85] **Madlener S, Ströbel T, Vose S, Saydam O, Price B.D, Demple B, and Saydam N**. *Essential role for mammalian apurinic/apyrimidinic (AP) endonuclease Ape1/Ref-1 in telomere maintenance.* Proc. Natl. Acad. Sci. U. S. A., 110, 17844–17849. (2013)

[86] **Dennis G. Jr, Sherman B.T, Hosack D.A, Yang J, et al.** *DAVID: Database for Annotation, Visualization, and Integrated Discovery*. Genome Biol. 4, 3.(2003)

[87] **Bossi A, Lehner B.** *Tissue specificity and the human protein interaction network*. Mol Syst Biol 5, 260.(2009)

[88] **Vidal M, Cusick M.E, Barabasi A.L.** *Interactome networks and human disease*. Cell 144, 986-998.(2011)

[89] **Szklarczyk D, et al.** *STRING v10:protein-protein interaction networks, integrated over the tree of life*. Nucleic Acids Research 43, D447-D452.(2014)

[90] **Guttman M, Rinn JL.** *Modular regulatory principles of large non-coding RNAs.* Nature 482: 339–346.(2012)

[91] **Li Y, Zhuang L, Wang Y, Hu Y, Wu Y, Wang D, Xu J.** *Connect the dots: a systems level approach for analyzing the miRNA-mediated cell death network.* Autophagy 9, 436–439.(2013)

[92] **Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi P.P.** *A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?* Cell 146, 353–358.(2011)

[93] **Sumazin P, Yang X, Chiu H.S, Chung W.J, Iyer A, Llobet-Navas D, Rajbhandari P, Bansal M, Guarnieri P, Silva J, et al.** *An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma.* Cell 147: 370–381.(2011)

[94] **Cesana M, et al.** *A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA*. Cell 147,358-369.(2011)

[95] **Fantini D, Vascotto C, Marasco D, D'Ambrosio C, Romanello M, Vitagliano L, Pedone C, Poletto M, Cesaratto L, Quadrifoglio F, Scaloni A, Radicella JP, Tell G.** *Critical lysine residues within the overlooked N-terminal domain of human APE1 regulate its biological functions.* Nucleic Acids Res. 38, 22, 8239-56.(2010)

[96] **Zhang X, et al.** *RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein interaction*. Bioinformatics 20, 989-993.(2014)

Journal of Computational Biology: http://mc.manuscriptcentral.com/liebert/jcb

Similarity measures based on the overlap of ranked genes are effective for comparison and classification of microarray data

Journal:	*Journal of Computational Biology*
Manuscript ID	JCB-2015-0057.R2
Manuscript Type:	Original Paper
Keyword:	FUNCTIONAL GENOMICS, GENE EXPRESSION
Abstract:	Similarity (or conversely distance) measures are at the heart of most bioinformatic applications. When the similarity involves only a small subset of features out of many, global similarity measures may be significantly affected by noise. Selecting only a subset of (putatively relevant) features for comparison is a widespread solution to the problem albeit affected by arbitariness and manual intervention. The problem is becoming more and more important due to the increasing amount of experimental data available. In recent years measures based on ranking similarities between two datasets have been proposed. Here, we use one of the proposed rank similarity measures, sharing some aspects with the fraction enrichment score used for protein structure prediction and the Gene Set Enrichment Analysis and test its performance in classifying experiments. The discrimination ability of the similarity measures based on the overlap of ranked genes tested here compares well or better with standard measures of similarity. This conclusion supports the use of rank based proximity measures to gain further insight in datasets comparisons, in particular on expression data obtained by different techonologies (e.g. RNA-seq and microarrays).

Similarity measures based on the overlap of ranked genes are effective for comparison and classification of microarray data

Fabrizio Serra[1], Chiara Romualdi[2], Federico Fogolari[1,3]

[1] Department of Biomedical Sciences and Technologies,
University of Udine
Piazzale Kolbe, 4
33100 Udine - Italy
fabrizio.serra,federico.fogolari@uniud.it
[2] Department of Biology,
University of Padova,
via U. Bassi 58/B,
35121 Padova, Italy
chiara.romualdi@unipd.it
[3] Istituto Nazionale Biostrutture e Biosistemi
Viale Medaglie d'Oro 305, 00136 Roma, Italy

Abstract. Similarity (or conversely distance) measures are at the heart of most bioinformatic applications. When the similarity involves only a small subset of features out of many, global similarity measures may be significantly affected by noise. Selecting only a subset of (putatively relevant) features for comparison is a widespread solution to the problem albeit affected by arbitariness and manual intervention. The problem is becoming more and more important due to the increasing amount of experimental data available. In recent years measures based on ranking similarities between two datasets have been proposed. Here, we use one of the proposed rank similarity measures, sharing some aspects with the fraction enrichment score used for protein structure prediction and the Gene Set Enrichment Analysis and test its performance in classifying experiments.
The discrimination ability of the similarity measures based on the overlap of ranked genes tested here compares well or better with standard measures of similarity. This conclusion supports the use of rank based proximity measures to gain further insight in datasets comparisons, in particular on expression data obtained by different techonologies (e.g. RNA-seq and microarrays).

1 Introduction

Similarity measures are central to many bioinformatic applications that aim at inferring novel knowledge from previous knowledge. Proper evaluation of similarity is more and more important due to the ever increasing amount of data available in public databases. For entities that can be represented by a vector of numerical features, some similarity measures have emerged as a *de facto* standard. The interplay of distance definition and clustering algorithms has been thoroughly addressed by some recent works (Giancarlo *et al.*, 2010, 2013; Jaskowiak *et al.*, 2014). Distance definitions include Minkowski's distances (e.g. Euclidean, Manhattan) and dissimilarity measures based on correlation (e.g. Pearson, Spearman correlation).
Sometimes however global similarity measures have limitations due to the fact that only a limited

set of features (out of many) is responsible for the similarity and relevant signal may be hidden in noise.

In the field of protein structure predictions, for instance, the true structure is not known and typically many predictive molecular models are proposed which must be screened and evaluated according to some scoring function. It is important to assess how one's own quality score of predictive models relates to the true quality of the models (e.g. measured by coordinates root mean square deviation (RMSD) with respect to the true native structure). A straightforward comparison of quality scores with RMSDs, however, is not particularly significant because the interest is in close to native predictive models, whereas for wrong predictions it is not important how much distant from the target structure they are. In other words it is not important if the RMSD is 8.0 Å or 15 Å so long as in both cases the prediction is considered wrong.

The test of scoring functions is typically performed on decoy sets which deviate from the true native structure. If non-native models in the set are many more than native-like models, global similarity measures will fail to detect the best scoring function because the signal will be hidden by noise.

Similarly if only a small subset of genes is differentially expressed in two microrarray experiments the non-differentially expressed gene set will introduce noise on global similarity measures.

Non-global similarity measures based only on a subset of genes (features) are to be preferred in this respect and indeed, at least in the field of microarray data analysis, have reached widespread use. On the other hand these methods require often the use of cutoffs or a threshold value to select e.g. significant fold-changes.

In recent years to overcome the limitations of global measures of similarity a number of approaches based on rank-rank correlation have been proposed (Yang *et al.*, 2006; Plaisier *et al.*, 2010; Antosh *et al.*, 2011, 2013). A simple way to apply this principle is to select a gene set from the experiment (typically the most over/under-expressed genes) and check for over-representation of terms belonging to some biologically relevant set (e.g. terms with some gene Ontology annotation, or belonging to the same biological pathway). This over-representation analysis amounts to a comparison between one's own experiment (the experiment under analysis, idealized by a set of differentially expressed genes) and an ideal experiment where only genes belonging to a certain class are differentially expressed.

Two important solutions to the problem have been proposed in the past. In the field of protein structure predictions the problem has been addressed by considering the so called "fraction enrichment" (Wang *et al.*, 2004; Fogolari *et al.*, 2005). In practice the 10% best scoring predictions are considered and the overlap with the 10% best models is evaluated. The choice of 10% is arbitrary and should be tuned to the set of models with some internal method.

In the field of microarray data experiments the Gene Set Enrichment Analysis (Mootha *et al.*, 2003; Subramanian *et al.*, 2005) measures the enrichment of a set of genes (derived from and thus ideally representing a microarray experiment) in the experimentally most over/under-expressed genes. The idealization process is only on the experiment at the source of the reference gene set, whereas all the results of the experiment under analysis are considered. Similar ideas are used in standard overrepresentation analyses, e. g. as implemented in the server David (Huang *et al.*, 2009).

Albeit the two above mentioned approaches are different they share common principles and effective

ways to combine the ideas underlying the two methods in a single method have been proposed. In particular Spang and coworkers (Yang *et al.*, 2006) proposed a global similarity measure based on the similarity of ranking in two ordered lists of genes.
Here we use the similarity measure proposed by Spang and coworkers (Yang *et al.*, 2006) (shifted and scaled to bring it in the range 0 to 1, and including a linear weight decay) and:
1) we compare its ability to recover similarities between different datasets with classical distances and for different choices of parameters and data pre-processing;
2) we assess the relationship between such distance and the cardinality of ranked genes with most significant overlap;
3) we assess the relationship between such distance and the p-value of the overalap;
4) we show that it is suited to compare data acquired with different technologies.
In the latter scenario a hybrid method like the the normalized Rank-Magnitude index based distance (Campello and Hruschka, 2009), which combines ranks and magnitudes of data, shows similar results, confirming its usefulness in comparing data with different scales and ranges.
Our results support the usefulness of similarity measures based on the overlap of ranked genes which perform as well or better as more traditional correlation measures for similarity recognition.

2 Methods

2.1 Fraction Enrichment (FE)

Given a set of n_e experiments testing n_g common genes we refer to $d[k,i]$ as the $k-th$ gene sampled in the $i-th$ experiment. Data have been taken: i) as they are; ii) processed by centering; iii) processed by centering and scaling by their standard deviation. Centering and scaling is performed by averaging across the experiments and computing the root mean square deviation of genes across experiments, as detailed below. Results are reported for the three kinds of processing.
After suitable processing, data corresponding to each experiment are sorted. Then for each value $1 \leq m \leq n_g$ we count how many genes are found in the first m sorted genes that are common to two sorted subsets. This quantity (typically, but not here, divided by m) is defined as the fraction enrichment corresponding to level m, n_g. Thus, if $S_{i,m}$ is the set of first m sorted genes of the set of genes S sampled by experiment i, the fraction enrichment (FE) at level m, n_g of set S_i with respect to the set S_j is:

$$FE_{m,n_g}(i,j) = |S_{i,m} \cap S_{j,m}| \tag{1}$$

Here we use this definition without further normalization.

2.2 Fraction Enrichment p-value (FEP)

In principle the behaviour of the FE or of its sum up to the m^{th} level could be used to select the top scoring genes and to assess the p-value of the enrichment. In other words a p-value is defined in

the present case as the probability that, given n_g genes, taking randomly two sets of m genes each, they share FE_{m,n_g} genes or more in common. The p-value could be computed in principle for each m and the minimum FE p-value (FEP) could be used as a measure of similarity. The computation of the p-values at each level based on the hypergeometric distribution is obviously not practical, but the burden of computation can be scaled down by a factor n evaluating the p-value only at multiples of n. In practice at intervals of $n = 50$ the p-value of having FE_{m,n_g} common genes in two randomly chosen sets of m genes out of a set of n_g genes is computed and the most signficant level (discretized by this procedure at n, $2n$, $3n$...) is chosen corresponding at the minimum p-value. For the minimum p-value the number of common genes is also recorded. Note that p-values are used here only for this purpose and therefore they are not corrected here for multiple testing.

2.3 Fraction Enrichment sum (FES)

In an approach inspired by that used in the Gene Set Enrichment Analysis (GSEA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), the fraction enrichments are summed over m and the result is normalized relative to the minimum and maximum possible values of the sum. Since no p-value computation is performed the calculation is performed at each level.
The maximum FE is obtained when the two sets have exactly the same order. In this case the FE is:

$$FE_{m,n_g} = m$$

and the sum of all FE_{m,n_g} up to the k^{th} term is:

$$\sum_{m=1,k} FE_{m,n_g} = \frac{k(k+1)}{2}$$

The minimum FE is obtained when the two sets have exactly the opposite order. In this case if $\lfloor \frac{n_g}{2} \rfloor$ is the largest integer lesser or equal to $\frac{n_g}{2}$ and $\lceil \frac{n_g}{2} \rceil$ is the smallest integer larger or equal to $\frac{n_g}{2}$:

$$FE_{m,n_g} = 0 \qquad m \leq \lfloor \frac{n_g}{2} \rfloor$$

$$FE_{m,n_g} = 2(m - \lceil \frac{n_g}{2} \rceil) + 2(\frac{n_g}{2} - \lfloor \frac{n_g}{2} \rfloor) \qquad m > \lfloor \frac{n_g}{2} \rfloor$$

and the sum up to the k^{th} value is:

$$\sum_{m=1,k} FE_{m,n_g} = 0 \qquad k \leq \lfloor \frac{n_g}{2} \rfloor$$

$$\sum_{m=1,k} FE_{m,n_g} = (k - \lceil \frac{n_g}{2} \rceil)(k - \lceil \frac{n_g}{2} \rceil + 1) + 2(k - \lfloor \frac{n_g}{2} \rfloor)(\frac{n_g}{2} - \lfloor \frac{n_g}{2} \rfloor))$$

$$k > \lfloor \frac{n_g}{2} \rfloor$$

The sum of FE's will be larger if common genes are found in the first positions of the ordering. In practice it turns out that summing the FE's is equivalent to weighting the common genes according to the list scanning level at which they are found present in both sets.
In order to have a single value representing the similarity between two experiments, the Fraction Enrichment Sum (FES) is defined as:

$$FES = \frac{\sum_{m=1,n_g} FE_{m,n_g} - min}{max - min} \tag{2}$$

with

$$max = \frac{n_g(n_g+1)}{2}$$

and

$$min = (\frac{n_g + n_g \bmod 2}{2})(\frac{n_g - n_g \bmod 2}{2} + 1)$$

2.4 Distance based on Fraction Enrichment sum (DIS_{FES})

The summation scheme discussed in the previous section is a particular form of the weighting scheme proposed by Spang and coworkers (Yang *et al.*, 2006). Instead of simple summation they introduced a weighting scheme decreasing with increasing m in the above equation 2. The weight proposed is exponentially decreasing with a parameter α equal to the decay rate, i. e.:

$$FES^{\alpha} = \frac{\sum_{m=1,n_g} FE_{m,n_g} e^{-\alpha m} - min}{max - min} \tag{3}$$

When the parameter α is set to 0 we recover equation 2. The only modification adopted here is that the summation is shifted and scaled in such a way that its range is 0 to 1.
We tested also a linear weighting scheme where the weight decreases linearly to zero in k terms.

$$FES^{\alpha} = \frac{\sum_{m=1,k} FE_{m,n_g} \frac{m-k+1}{k} - min}{max - min} \tag{4}$$

The corresponding dissimilarity measure, defined as DIS_{FES}, is obtained as:

$$DIS_{FES} = 1 - \frac{(FES + FES_r)}{2} \tag{5}$$

where FES_r is the FES for the reverse sorting of both gene expression sets. This choice amounts to giving an equal importance to up-regulated (at the top of the list) and down-regulated (at the bottom of the list) genes (Yang *et al.*, 2006).
Different choices, not followed here, involve taking only FES or FES_r in the formula for distance.

2.5 Choice of parameters

Both exponential (Eq. 3) or linear (Eq. 4) weighting employ a parameter, α or k, respectively.
For the exponential weighting Spang and coworkers suggest an adaptive procedure which however cannot be adopted only on the ranked gene lists, but rather maximizes the area under curve in ROC analysis, which assumes knowledge of the true classification of data.
We suggest here to perform the analysis with $\alpha = 0$ and to use FEP to assess the cardinality (m in Eq. 1) of the sets leading to the most significant overlap (i.e. the lowest FEP). If we refer to the latter value of m as m_{opt}, the parameter α should decay significantly after the first m_{opt} positions. A reasonable choice for α is thus $\alpha = \frac{1}{m_{opt}}$. A similar rationale has been adopted by Graeber and coworkers (Plaisier *et al.*, 2010) and Neretti and coworkers (Antosh *et al.*, 2013), whose optimal choice of overlapping sets is based on maximization of probability values.
The linear weighting scheme is somewhat more rigid, but similarly we may take $k = 2m_{opt}$.

2.6 Comparing sorted lists

Genes are sorted by their expression in time $\mathrm{O}(n_g \log n_g)$. Once two sorted list of genes are obtained (say list1 and list2), the fraction enrichment for all m levels and its sum is computed in time $\mathrm{O}(n)$ by keping track of the found genes in two boolean vectors, say found1 and found2. The two boolean vectors are initialized at 0. The lists are scanned in parallel for the first element, then the second,..., the k^{th}, ... the last element. Everytime a gene is found in list1 (list2) the corresponding element in found1 (found2) is set to 1 and the same element in found2 (found1) is checked. If the same element was already set to 1, it means that the same gene had been already found and the number of common genes up to that level is incremented by 1. This operation is performed in time $\mathrm{O}(n)$.
If n_e experiments are to be compared all-against-all, sorting must be performed n_e times, whereas the comparison of sorted genes must be performed $n_e(n_e + 1)/2$ times. In this respect it is convenient to choose a cut-off on the weight for the comparison when using weighting of contributions. For instance if the weight is scaled linearly down to zero in k terms, only k terms will be considered in the comparison of sorted genes. Similarly for the exponentially decaying weight terms beyond two or three decay lengths could be ignored, making sometimes the computation significantly faster.

2.7 Data processing

When explicitly mentioned data were centered and normalized across experiments (notation as in the above paragraph, $d[k, i]$ is the expression level of the $k - th$ gene sampled in the $i - th$ experiment) so that:

$$d[k, i] \leftarrow \frac{\left(d[k, i] - \overline{d[k]}\right)}{\sigma_d[k]}$$

with $\overline{d[k]}$ and $\sigma_d[k]$ the mean and the standard deviation of $d[k, i]$ over all experiments, respectively. The rationale for centering the data about the average expression was to remove obvious

similarities in ranking of genes due to overall expression levels, and to compare deviations from average expression levels. Normalization was chosen to bring all variations with respect to average to a common scale.

2.8 Test sets

We considered two different test sets for which classification is available and that have been used also by other authors to assess similarity measures, and a test set including data obtained by both RNA-seq and microarray technologies. In particular:
i) the first set is a collection of 35 experiments compiled from the literature (de Souto *et al.*, 2008; Jaskowiak *et al.*, 2013). The data are provided in tabular form by the same authors and allow straightforward comparison with the similarity measures studied by the same authors (Jaskowiak *et al.*, 2013). These data sets were obtained using single-channel Affymetrix chips (21 sets) and double-channel cDNA (14 sets) microarrays. Each of the 35 sets includes a variable number of experiments (22 to 248, with average 90), each experiment within a set is classified according to tumor type or status. The number of classes for each set ranges from 2 to 14 with average 3.5. All references are reported in the original papers (de Souto *et al.*, 2008; Jaskowiak *et al.*, 2013).
ii) the second set is the set of 950 expression data for 19204 genes made available by Bioinformatics Grid^{TM} (caBIG) of the National Cancer Institute (downloaded from https://cabig.nci.nih.gov/caArray_GSKdata/ and presently available from ftp://caftpd.nci.nih.gov/ pub/ caARRAY/ in the directory transcript_profiling/). The platform used for all experiments was Affymetrix GeneChip HG-U133 Plus2, and the experiments were normalized using the MAS5 algorithm (Lim *et al.*, 2007). The comparisons for the latter set are more challenging because no selection of genes is performed. For each experiment annotation indicates replicates and different tissue and cancer or normal types. We consider here three classifications:
1) each set including one or more replicate experiments constitutes a single class. The test on this set of experiments is able to measure how well very similar experiments are separated from other experiments;
2) all sets coming from the same tissue constitute a single class.
3) all sets coming from the same tissue and same cancer type constitute a single class.
The latter tests are able to measure how well less similar experiments are recognized as similar;
We will refer to these first two datasets as de Souto and GSKdata, respectively.
iii) the third set of data includes both RNA-seq and microarray absolute expression levels. The latter set has been chosen to show the usefulness of rank based similarity measures compared to other commonly employed measures, when different techniques are used for the assessment of expression levels. The set consists of data collected on Human CCR6^{+} CD4 memory T cells at two hours intervals following stimulation with Anti-CD3/CD28 (Zhao *et al.*, 2014). For each technique (RNA-seq and microarray) and each time interval data have been acquired in duplicate. Overall correspondence between data could be found for 9603 genes. RNA-seq count data have been log2-transformed after the addition of 1 pseudo-count, before use in the analysis. We will refer to this

set as Zhao.

2.9 Comparison and classification of gene expression data

The power of the DIS_{FES} to identify related experiments was assessed by the Intrinsic Separation Ability (ISA) as defined by Giancarlo et al. (Giancarlo *et al.*, 2010, 2013) and used by Jaskowiak et al. (Jaskowiak *et al.*, 2013). This test has the advantage of singling out the effect of the distance definition from the effect of the clustering algorithm, although the two are not independent (Giancarlo *et al.*, 2010, 2013; Jaskowiak *et al.*, 2014).
The ISA of each similarity measure is assessed following the cited authors by building the receiver operating characteristic curve (ROC) and measuring the area under the curve (AUC). In practice ROCs are built by considering for all pairs of experiments (i, j) their distance $D(i, j)$ and a binary vector $I(i, j)$ which is 1 if class(i) == class(j) and 0 otherwise. Then the distance threshold for predicting whether two experiments belong to the same class is varied and for each threshold the true positive classification rate is plotted vs. the false positive classification rate. The area under this curve is the AUC. A better than random classifier has AUC over 0.5. The AUC has the advantage of condensing in a single figure the performance of the various distance based classificators. Further details are given in the cited references (Giancarlo *et al.*, 2010, 2013; Jaskowiak *et al.*, 2014).
The separation from the mean (expressed in standard deviation units, i. e. the so-called z-score) of the FEP or FES within a group sharing the same annotation versus other experiments was also tested, but it was less significant, because strongly dependent on the test set and was not further considered. In order to compare the performance of DIS_{FES} with widely used and accepted distance measures (D) (Giancarlo *et al.*, 2010) we considered for two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$:
i) the 1-norm of the difference vector (Manhattan distance)

$$D_{a,b} = \sum_i |a_i - b_i| \tag{6}$$

ii) the 2-norm of the difference vector (Euclidean distance)

$$D_{a,b} = \sqrt{\sum_i (a_i - b_i)^2} \tag{7}$$

iii) Pearson correlation (r)

$$D_{a,b} = 1 - r = 1 - \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{n \sigma_a \sigma_b} \tag{8}$$

where n is the number of components, $\bar{a}$ and $\bar{b}$ are the average of a and b components, respectively, and σ_a and σ_b are the standard deviations of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively.
iv) Spearman correlation (i.e. Pearson correlation on ranks, ρ)

$$D_{a,b} = 1 - \rho = 1 - \frac{6\sum_i \Delta_i^2}{n(n^2-1)} \tag{9}$$

where ρ is Spearman's correlation, Δ_i is the difference in ranks of a_i and b_i.
Note that Spearman correlation uses sorted list of genes, as FES does, but it uses the differences in rank (instead of the proximity as in FES) and each item contributes only once to the statistics, whereas each item contributes many times, with different weights in FES.
v) Kendall rank correlation coefficient (Kendall's τ)
This is obtained by checking any possible pairs of components (i, j) on the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The observations are defined as concordant if
$(a_i > a_j$ and $b_i > b_j)$ or $(a_i < a_j$ and $b_i < b_j)$
and discordant if
$(a_i > a_j$ and $b_i < b_j)$ or $(a_i < a_j$ and $b_i > b_j)$
We thus define the distance:

$$D_{a,b} = 1 - \tau = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{10}$$

where n_c and n_d are the numbers of concordant and discordant pairs of observations.
Kendall tau could not be used on all sets because the complexity of its implementation in the function cor() in R is $O(n_g^2)$, although algorithms with complexity $n \log n$ have been reported (Knight, 1966; Christensen, 2005; Campello and Hruschka, 2009). Using R the running time on the large GSK dataset ($n_g = 19204$,$n_e = 950$) was unpractical. However we tested its behaviour on the test set used by Jaskowiak et al. (de Souto *et al.*, 2008; Jaskowiak *et al.*, 2013) described above and the results where almost overlapping with the results obtained using Spearman correlation with correlation coefficient 0.998. We reasonably expect that its application to the larger set of data would result in similar performance as Spearman correlation.
vi) Normalized Rank-Magnitude index
This measure was introduced and tested by Campello and coworkers' (Campello and Hruschka, 2009) as a sensitive measure of correlation for data acquired with different technologies and scales. One of the vectors to compare is ranked (low indices are assigned to the lowest magnitude components) and the normalized Rank-Magnitude index (RM) uses the scalar product of the ranks of the first vector with the second vector:

$$D_{a,b} = \frac{2\sum_i \mathrm{Rank}(a_i)b_i - RM_{max} - RM_{min}}{RM_{max} - RM_{min}} \tag{11}$$

where RM_{max} and RM_{min} are the maximum and the minimum of the scalar product of the ranks of the first vector with the second vector, assuming the ordering of the first vector components in the most favorable or most unfavorable way, respectively. As can be seen the distance is not symmetric for an exchange of $\boldsymbol{a}$ and $\boldsymbol{b}$ and it has been symmetryzed by: $D_{a,b} = D_{b,a} \leftarrow \frac{D_{a,b}+D_{b,a}}{2}$.
All distances (D) have been transformed in similarity measures (s) in the range 0 to 1 by equation:

$$s = \frac{D - \min(D)}{\max(D) - \min(D)} \tag{12}$$

for use as predictors of two experiments belonging to the same class.

3 Results and discussion

3.1 Comparison of similarity measures

The results obtained on the first two datasets considered here, GSK and de Souto datasets, are summarized in Table 1 where the ten (or more for equal values) best results are highlighted in boldface.We report in the table the area under curve (AUC) of the receiver operating characteristic (ROC) curve.

It is apparent from the table that among the standard distances those based on Pearson and Spearman correlations are in all cases the best performing. Scaling the expression levels by the average root mean square deviation from the average (over the set of experiments considered) has effect but it is dependent on the analysed set as could be expected.

For the DIS_{FES} dissimilarity measure centering the data on the average expression levels of the data set considered improves always the performance, whereas scaling has effects which depend on the set.

The effect of the parameter α is in turn dependent on centering and scaling of the dataset.

For all choices of α, however, a performance comparable to the well established Pearson and Spearman correlations is observed.

The classification of experiments in the GSK dataset allows to assess the performance of the dissimilarity measures in discriminating very similar (repicates) and less similar (cell types and cell and tumor types) expression sets.

For the best choice of α (0.001) the DIS_{FES} dissimilarity measure classification is comparable or better than the Pearson and Spearman correlations based classification.

The test performed with the normalized Rank-Magnitude index based distance (Campello and Hruschka, 2009), which uses both ranks and values, show results similar to DIS_{FES} using the best parameter choices, confirming the power of such distance.

The AUC constructed with the best performing similarity measures on GSK dataset have been compared using the DeLong statistical test (DeLong *et al.*, 1988) as implemented in the R package pROC (Robin *et al.*, 2011).

The results are reported in Table 2. It is seen that most pairs of AUCs are significantly different. Exceptions are for DIS_{FES} with exponential decay constant 0.001 on centered data and Pearson correlation on scaled data, Rank-Magnitude index on centered data, DIS_{FES} with exponential decay constant 0.001 on centered and scaled data and DIS_{FES} with linear decay in 1000 steps on centered and scaled data. The latter is in turn equivalent to DIS_{FES} with exponential decay constant 0.001 on centered and scaled data.

3.2 $\mathbf{DIS}_{FES}$ correlates with the statistical significance of fraction enrichment

The correlation of the DIS_{FES} dissimilarity measure with the minimum p-value corresponding to the overlap of genes sampled at discrete intervals was checked. The results are reported in Figure 1. It is seen that a strong similarity is always corresponding to a minimum p-value close to zero. The tests are reported for the centered and scaled GSK datasets for the three parameters $\alpha = 0, 0.001, 0.01$. Both the minimum p-value and its logarithm are reported in Figure 1. Although minimum p-values are not corrected here for multiple testing, even with the most conservative correction the picture would not change significantly.

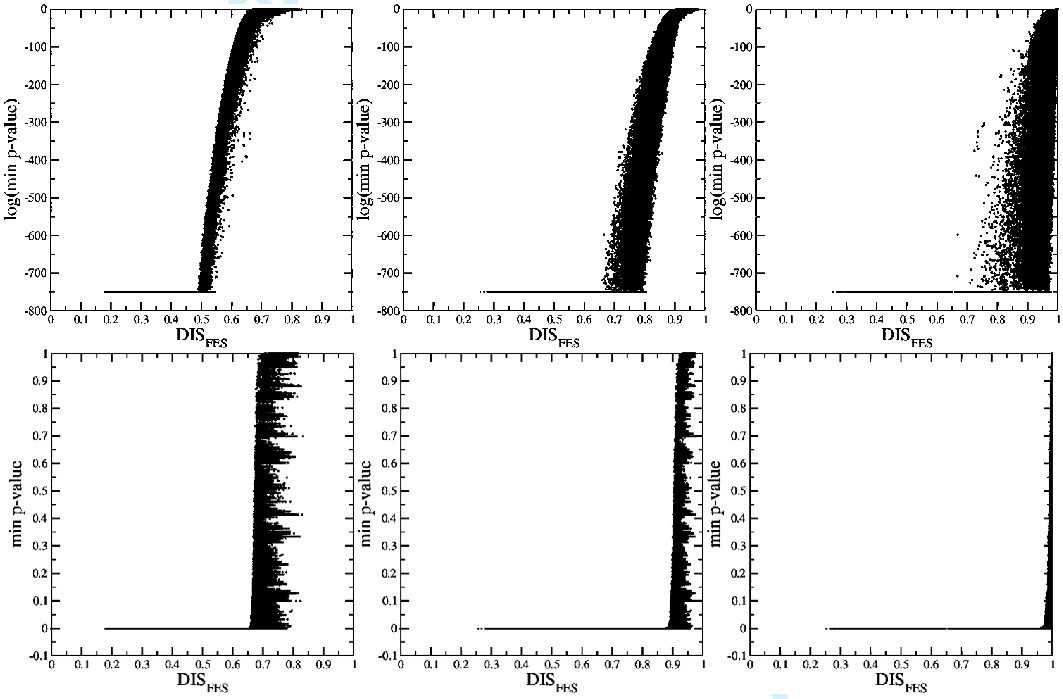


Fig. 1. Logarithm of the minimum p-value (upper panels) and minimum p-value (lower panels) versus DIS_{FES} dissimilarity measure for the GSK centered and scaled datasets. Logarithm values lesser than -750 were plotted at -750. The parameter α is 0, 0.001, 0.01 for the left, center, and right panels, respectively. Each point in the figure corresponds to the similarity between two experiments.

3.3 The number of genes contributing to similarity

Although the DIS_{FES} is a global measure of dissimilarity it is interesting to check the number of common genes corresponding to the minimum p-value. This gives an idea of the number of genes

most contributing to the similarity between two sets.
The results are reported in Figure 2. The figure reports also the detail at the closest similarity. It is seen that in all cases only a few hundred of genes are shared by the sets for which the p-value is minimum, which is in line with the the typical number of genes considered over- or under-expressed in microarray experiments. These numbers could be however linked with the specific GSK dataset.

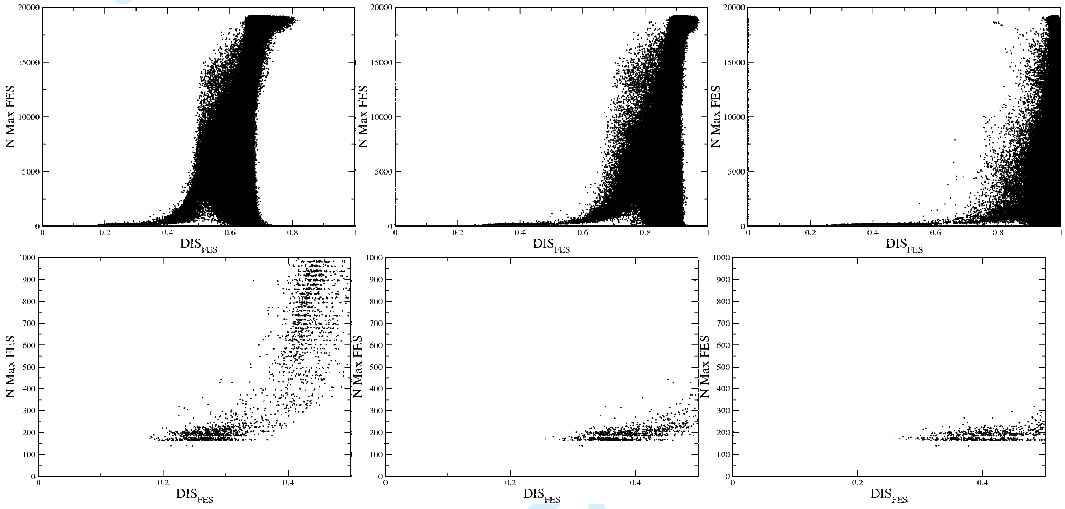


Fig. 2. Number of common genes corresponding to the minimum p-value versus DIS_{FES} dissimilarity measure for the GSK centered and scaled datasets. The parameter α is 0, 0.001, 0.01 for the left, center, and right panel, respectively. In the lower panels the detail at closest similarities is reported. Each point in the figure corresponds to the similarity between two experiments.

3.4 Application to data obtained with mixed technologies

Finally, we challenged DIS_{FES} with a dataset where expression levels are assessed by RNA-seq and microarray on the same samples. The set consists of 24 experiments (12 microarray and 12 RNA-seq) for each sample and there are two replicates for each technique. The design of the experiments allows one to test detection of similarities when data obtained with different techniques are considered and to check also the effect of technique versus sample similarity. The data considered here are not ratios of expression levels, but absolute values obtained from the supplementary tables accompanying the paper of (Zhao *et al.*, 2014).
The results are reported in Table 1. Two kinds of analyses are performed, first data obtained with microarrays are compared with data obtained with RNA-seq (first value in Table 1), then all data are analysed at the same time (values in parentheses in Table 1).
As expected measures based on differences like Euclidean or Manhattan distances perform poorly

and scaling definitely improves the performances. The computed AUCs are smaller when all data are considered together, because similarity in absolute values due to the same experimental technique overcome the similarity in relative expression levels.
Standard correlation measures perform better. Both Spearman and Pearson correlation can classify exactly scaled data. Pooling all data together decreases the performance as discussed above for both Pearson and Spearman correlation.
The performance of DIS_{FES} depends on the decay parameter chosen, but can classify exactly data (for the choice $k = 1000$ on centered and scaled data, also when pooled). Similar results are obtained using the normalized Rank-Magnitude similarity, also using ranks, on scaled and scaled and centered data.

4 Conclusions

The results reported in this paper show that similarity measures based on the overlap of ranked genes are as effective (or better) as the well established Pearson and Spearman correlation measures in identifying similarities between expression gene sets.
The scheme proposed by Spang and coworkers (Yang *et al.*, 2006), slightly adapted here, shows that for all close similarities the corresponding p-values for the overlap of ranked genes are close to zero, thus supporting the soundness of the approach. The number of genes contributing to the similarity, for the GSK dataset studied here, is in the range of 200-300 genes, consistently with the typically adopted choice of over- or under-expressed gene sets, according to different criteria. The latter observation suggests that the parameter that enters FES should be chosen as to weight significantly more the first few hundreds ranked genes (i.e. $\alpha \approx 0.01$–0.001).
Pearson and Spearman coefficients are able to detect linear relationships while FES measure could be able to detect similarities in more complex scenarios. Expression values derived from the same technologies (such as microarrays) are characterized by a clear linear relationship; similar samples have similar expression trend on the same range of values. Thus, in the context of microarray data classification the similar performance of FES with Pearson and Spearman coefficient is expected. However we would expect that FES would guarantee a better performance whenever samples are measured with different technologies such as a combination of RNA-seq and microarray. The results on a dataset containing RNA-seq and microarray data for the same samples are confirming this expectation showing that FES based measures could be useful for comparing different types of experiments. Similar results are reached for another rank-based distance tested here, i.e. the normalized Rank-Magnitude index.
Meta-analysis, that is the combination of different expression studies on the same biological problem, with the aim to increase sample size and then statistical power, would certainly benefit of the use of these types of similarity measure.

Acknowledgement

This work was partly supported by Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN 2012A7LMS3_001). We thank Dr. Raffaella Picco for providing the GSK dataset.

Bibliography

Antosh, M., Fox, D., Cooper, L. N., and Neretti, N., 2013. CORaL: comparison of ranked lists for analysis of gene expression data. *J. Comput. Biol.* 20, 433–443.

Antosh, M., Fox, D., Helfand, S. L., Cooper, L. N., and Neretti, N., 2011. New comparative genomics approach reveals a conserved health span signature across species. *Aging (Albany NY)* 3, 576–583.

Campello, R. J. G. B. and Hruschka, E. R., 2009. On comparing two sequences of numbers and its applications to clustering analysis. *Inf. Sci.* 179, 1025–1039.

Christensen, D., 2005. Fast algorithms for the calculation of kendall's τ. *Comp. Stat.* 20, 51–62.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., and Schliep, A., 2008. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* 9, 497.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845.

Fogolari, F., Tosatto, S. C., and Colombo, G., 2005. A decoy set for the thermostable subdomain from chicken villin headpiece, comparison of different free energy estimators. *BMC Bioinformatics* 6, 301.

Giancarlo, R., Lo Bosco, G., and Pinello, L., 2010. Distance functions, clustering algorithms and microarray data analysis. *Lect. Notes. Comp. Sci.* 6073, 125–138.

Giancarlo, R., Lo Bosco, G., Pinello, L., and Utro, F., 2013. A methodology to assess the intrinsic discriminative ability of a distance function and its interplay with clustering algorithms for microarray data analysis. *BMC Bioinformatics* 14 Suppl 1, S6.

Huang, d. a. 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.

Jaskowiak, P. A., Campello, R. J., and Costa, I. G., 2013. Proximity measures for clustering gene expression microarray data: a validation methodology and a comparative analysis. *IEEE/ACM Trans Comput Biol Bioinform* 10, 845–857.

Jaskowiak, P. A., Campello, R. J., and Costa, I. G., 2014. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics* 15 Suppl 2, S2.

Knight, W., 1966. A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.* 61, 436–439.

Lim, W. K., Wang, K., Lefebvre, C., and A., C., 2007. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* 23, 282–288.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 102, 267–273.

Plaisier, S. B., Taschereau, R., Wong, J. A., and Graeber, G., 2010. Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucl. Acids Res.* 38, e169.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Mueller, M., 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550.

Wang, K., Fain, B., Levitt, M., and Samudrala, R., 2004. Improved protein structure selection using decoy-dependent discriminatory functions. *BMC Struct. Biol.* 4, 8.

Yang, X., Bentink, S., Scheid, S., and Spang, R., 2006. Similarities of ordered gene lists. *J Bioinform Comput Biol* 4, 693–708.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X., 2014. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE* 9, e78644.

	GSKdata AUC			de Souto AUC	Zhao AUC
Distance	replicate	type and tissue	tissue		
Euclidean	0.985	0.758	0.701	0.678 ± 0.125	0.736 (0.429)
Manhattan	0.983	0.737	0.688	0.680 ± 0.123	0.678 (0.447)
Pearson	0.985	0.767	0.734	**0.718 ± 0.139**	0.913 (0.362)
Spearman	0.984	0.767	0.731	**0.701 ± 0.133**	0.965 (0.345)
Euclidean scaled	0.978	0.709	0.657	0.659 ± 0.115	0.991 (0.979)
Manhattan scaled	0.980	0.712	0.666	0.667 ± 0.118	**1.000** (**0.983**)
Pearson scaled	**0.997**	**0.786**	**0.786**	**0.701 ± 0.123**	**1.000** (**0.997**)
Spearman scaled	**0.996**	**0.783**	**0.785**	0.695 ± 0.121	**1.000** (**0.995**)
Rank-Magnitude	0.984	0.768	0.732	**0.710** ± 0.134	0.946 (0.351)
Rank-Magnitude centered	**0.996**	**0.793**	**0.799**	**0.716** ± 0.127	**1.000** (**1.000**)
Rank-Magnitude centered and scaled	**0.997**	**0.785**	**0.786**	0.699 ± 0.121	**1.000** (**0.995**)
DIS_{FES} ($\alpha = 0$)	0.982	0.751	0.720	0.690 ± 0.131	0.965 (0.345)
DIS_{FES} centered ($\alpha = 0$)	0.993	**0.780**	**0.782**	0.697 ± 0.120	**1.000** (**0.992**)
DIS_{FES} centered and scaled ($\alpha = 0$)	0.995	**0.780**	**0.781**	0.691 ± 0.116	**1.000** (**0.995**)
DIS_{FES} ($\alpha = 0.001$)	0.943	0.733	0.620	0.694 ± 0.128	0.972 (0.343)
DIS_{FES} centered ($\alpha = 0.001$)	0.993	**0.787**	**0.797**	0.698 ± 0.120	**1.000** (0.982)
DIS_{FES} centered and scaled ($\alpha = 0.001$)	**0.996**	**0.788**	**0.785**	0.691 ± 0.116	**1.000** (**0.995**)
DIS_{FES} ($\alpha = 0.01$)	0.943	0.657	0.620	**0.709 ± 0.128**	0.869 (0.741)
DIS_{FES} centered ($\alpha = 0.01$)	**0.996**	**0.786**	**0.800**	**0.711 ± 0.122**	**1.000** (0.850)
DIS_{FES} centered and scaled ($\alpha = 0.01$)	**0.996**	**0.781**	0.762	0.692 ± 0.116	**1.000** (**0.998**)
DIS_{FES} linear ($k = 10$)	0.886	0.587	0.595	0.634 ± 0.100	0.500 (0.650)
DIS_{FES} linear centered ($k = 10$)	**0.997**	0.713	0.661	0.630 ± 0.095	0.500 (0.500)
DIS_{FES} linear centered and scaled ($k = 10$)	0.995	0.590	0.448	0.584 ± 0.083	0.500 (0.500)
DIS_{FES} linear ($k = 100$)	0.972	0.607	0.636	**0.706 ± 0.125**	0.852 (0.725)
DIS_{FES} linear centered ($k = 100$)	**0.997**	0.777	**0.778**	**0.702 ± 0.120**	0.595 (0.572)
DIS_{FES} linear centered and scaled ($k = 100$)	0.995	0.742	0.699	0.671 ± 0.110	0.877 (0.886)
DIS_{FES} linear ($k = 1000$)	0.988	0.704	0.691	**0.700 ± 0.128**	0.999 (0.680)
DIS_{FES} linear centered ($k = 1000$)	**0.997**	0.777	**0.778**	**0.700 ± 0.120**	**1.000** (0.969)
DIS_{FES} linear centered and scaled ($k = 1000$)	**0.997**	**0.788**	**0.778**	0.691 ± 0.116	**1.000** (**1.000**)

Table 1. Summary of the results (see text for details). First column: distance used for classification; second to fourth columns: AUCs corresponding to classification based on experimental replicate, same type of tumor and tissue, same tissue for GSK dataset; fifth column: average and standard deviation of the AUCs obtained on the 35 sets included in the de Souto dataset; sixth column: AUC for the Zhao dataset when only data obtained with different techniques (RNAseq and microarray) are compared. In parentheses AUCs are reported when all data are pooled together.

	P	PS	S	SS	RM	RMC	RMCS	$F_{0.001}$	$F_{0.001}C$	$F_{0.001}CS$	F_{1000}	$F_{1000}C$	$F_{1000}CS$
P	-		=					2.2E-16			2.2E-16		
PS	2.2E-16	-	2.2E-16	2.2E-16	2.2E-16		2.2E-16	2.2E-16	=		2.2E-16	1.1E-08	
S	=		-					2.2E-16			2.2E-16		
SS	2.2E-16		2.2E-16	-	2.2E-16			2.2E-16			2.2E-16	2.9E-04	
RM	2.2E-16		2.2E-16		-			2.2E-16			2.2E-16		
RMC	2.2E-16	2.2E-16	2.2E-16	2.2E-16	2.2E-16	-	2.2E-16	2.2E-16	2.2E-16	2.2E-16	2.2E-16	2.2E-16	1.2E-08
RMCS	2.2E-16		2.2E-16	2.2E-16	2.2E-16		-	2.2E-16	=		2.2E-16	1.4E-06	
$F_{0.001}$								-			2.2E-16		
$F_{0.001}C$	2.2E-16	=	2.2E-16	2.8E-05	2.2E-16		=	2.2E-16	-	=	2.2E-16	7.4E-12	=
$F_{0.001}CS$	2.2E-16	1.2E-06	2.2E-16	2.2E-16	2.2E-16		2.2E-16	2.2E-16	=	-	2.2E-16	9.3E-13	=
F_{1000}											-		
$F_{1000}C$	1.4E-08		3.1E-08		1.6E-06			2.2E-16			2.2E-16	-	
$F_{1000}CS$	2.2E-16	4.4E-03	2.2E-16	2.0E-10	2.2E-16		1.4E-05	2.2E-16	=	=	2.2E-16	1.7E-15	-

Table 2. Comparison of classifiers based on different distances: P - Pearson correlation, S - Spearman correlation, PS - Pearson correlation on scaled data, SS - Spearman correlation on scaled data, $F^{0.001}$ - DIS_{FES} with exponential decay constant 0.001 , $F^{0.001}C$ - DIS_{FES} with exponential decay constant 0.001 on centered data, $F^{0.001}CS$ - DIS_{FES} with exponential decay constant 0.001 on centered and scaled data, F_{1000} - DIS_{FES} with linear decay in 1000 values, $F_{1000}C$ - DIS_{FES} , with linear decay in 1000 values on centered data, $F_{1000}CS$ - DIS_{FES} with linear decay in 1000 values on centered and scaled data. The p-value (if smaller than 0.05) is reported. The presence of a p-value in a cell of the matrix means that a classifier based on the distance in the row performs better and is significantly different from the classifier based on the distance in the column. Conversely a void cell indicates that the classifier based on the distance in the column performs better and is significantly different from the classifier based on the distance in the row. If the two classifiers are not significatly different an equal sign is reported.

8 ACKNOWLEDGMENTS

I would like to thank all my supervisors of these three years (Prof. Federico Fogolari, Prof. Claudio Brancolini, Prof.ssa Chiara Romualdi and Prof. Gianluca Tell) from each of whom I have acquired some new knowledge and skills.

Thanks also to the research group of Prof.ssa Chiara Romualdi (University of Padua), in particular Paolo Martini, and to my collegue Raffaella.

A special acknowledgment goes to my family and Cristina for their support.