# UNIVERSITY OF NAPLES FEDERICO II

**RESEARCH DOCTORATE IN**

**COMPUTATIONAL BIOLOGY AND BIOINFORMATICS**

(*XXV cycle*)

# Computational approaches to analyze next generation sequencing data


**Coordinator:**
Prof. Sergio Cocozza

**Tutor:**
Prof. Alessandro Weisz

**Co-tutor:**
Prof. Maria Luisa Chiusano

**Candidate:**
Dr Maria Rosaria De Filippo

*Il successo è sempre figlio dell'audacia*

*(Voltaire)*

# *Abstract*

Advances in next generation sequencing in the last few years have enabled an increasing number of applications in biology and medicine, from whole genome to small-RNA sequencing, with increased throughput accompanied by plunging costs. This thesis is focalized on two of the most used applications, small-RNA sequencing, to investigate the biological function of the increasing population of small non coding RNA, including micro-RNA and Exome sequencing to identify single nucleotide variations (SNV) and small insertion and deletions (InDel). In this context two different dataset were used: the first obtained from small-RNA-sequencing using human breast cancer MCF-7 cells in two different conditions and the latter obtained from exome sequencing in patients with a rare syndrome (malignant migrating partial seizures of infancy). A large amount of data were produce from each experiment, required comprehensive analysis pipelines to analyze them.

Small-RNA sequencing represents a novel technology widely used to investigate with high sensitivity and specificity small non-coding RNA populations, comprising microRNAs and other regulatory transcripts. To gather biologically relevant information, such as detection and differential expression analysis of known and novel non-coding RNAs and target prediction, the analysis requires the implementation of multiple statistical and bioinformatics tools from different sources, each focusing on a single step of the analysis pipeline. As result, a novel modular pipeline called iMir for comprehensive analysis of miRNA-Seq data, from adapter trimming, quality filter to differential expression and biological target prediction together with other useful options was designed by integrating multiple open source modules and resources in an automated workflow. The pipeline was applied to analyze simultaneously miRNA-Seq datasets from human breast cancer MCF-7 cell, resulting in a rapid and accurate identification, quantization and differential expression analysis of ~450 miRNAs, including several novel miRNAs and isomiRs, as well as identification of the putative mRNA targets of differentially expressed miRNAs.

Exome sequencing, the targeted sequencing of coding regions of the genome, is a powerful and cost-effective technique for dissecting the genetic basis of diseases and traits that have proved to be intractable with conventional gene-discovery strategies. To reduce the number of false positive variations and simplify the understanding of results a comprehensive pipeline was developed, integrating different tools. Starting from quality check and alignment, base quality score recalibration and local realignment around indels were

performed and SNV and InDel were called. Finally, different filters were applied to discard variations with low quality and coverage. The pipeline was then used to analyze data from exome sequencing in six patients with malignant migrating partial seizures in infancy, also known as MMPSI or MMPEI. After analysis and filtering common variants between 6, 5, 4 and 3 patients were studied to identify putative disease causing mutation(s). Results obtained indicate the accuracy of the pipeline to identify SNV and short InDels and the reliability to provide a global and quantitative catalogue of nucleotide variants in the exome.

# *Acknowledgements*

First of all I would like to thank my supervisor, Prof. Alessandro Weisz, for his support, encouragement and advice and also for giving me the opportunity to deepen my knowledge abroad. In this context I would like to thank all people I met in Helsinki, in particular prof. Lauri Aaltonen and Riku Katainen that helped me in the first steps of my project.

A special thanks to all the members of the lab, without whom I could not have completed my work. Special thanks to Maria, Giorgio, Roberta and Giovanni that during this period were very good friends as well as precious colleagues.

I would also like to thank my co-tutor, Prof. Maria Luisa Chiusano and the PhD program committee, in particular Prof. Cocozza, Prof. Miele and all the other Professors.

I'm very grateful to all the people that supported me in my personal life, my friend and in particular my best friend Ciro, that though far physically always advised me and supported me. Last and more important thanks go to my family (including my little dog), in particular my parents, who have always supported my studies and my carrier with lots of love and efforts, my sister and best friend for life and finally Stefano for his patience and unconditional love.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1. NEXT GENERATION SEQUENCING

The method developed by Frederick Sanger and colleagues in 1977 for determining nucleotide sequence in DNA by using 2',3'-dideoxy nucleotides for chain termination marked a milestone in the history of DNA sequencing. This concept provided a basis for the development of automated Sanger Sequencing (Smith *et al*. 1986, Ansorge *et al*. 1987) that was the core technology of the Human Genome Project, started in 1990 with the goal of determining all three billion base pairs making up the human genome. The first results of the project were produced after ten years, in 2000 (Lander *et al*. 2001, Venter *et al*. 2001), and additional three years were needed to complete it (Jasny and Roberts, 2003). The advent of the so-called Next Generation Sequencing (NGS) have completely changed the way in which we think about genetic and genomic research, in a manner akin to the introduction of Sanger Sequencing in the 1970's. NGS technologies provide opportunities for global investigation of multiple genomes and transcriptomes in at unprecedented speed in combination with low costs per base (**Figure 1**). As a consequence, the number of sequencing related data, stored in public available databases has increased significantly over the last years (**Figure 2**).

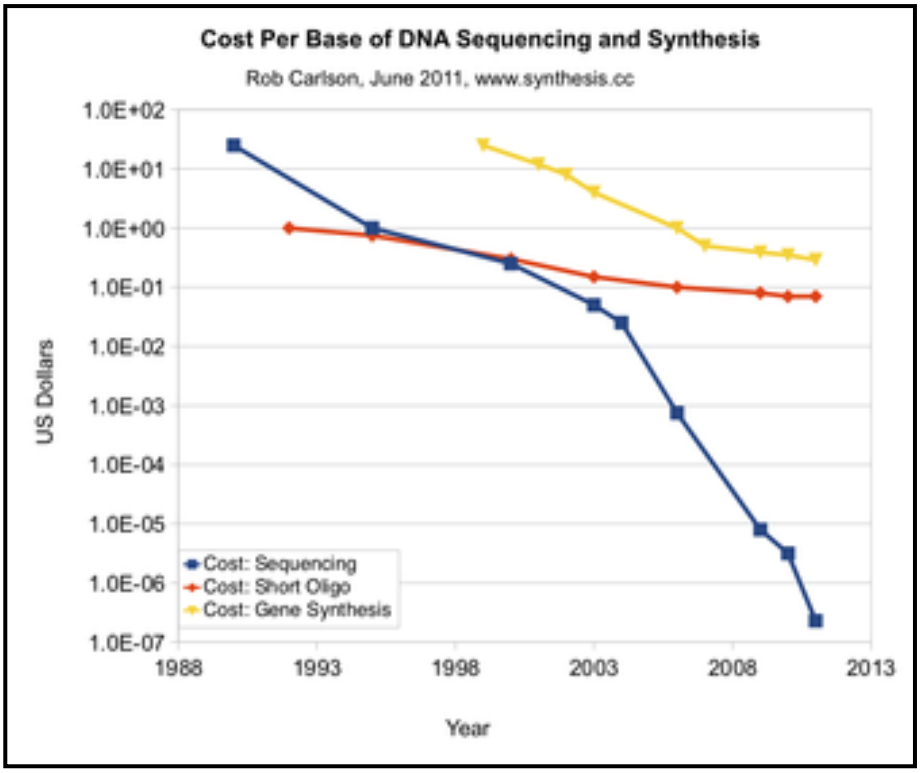



**Figure 1 Per Base Cost Development Of Dna Sequencing On A Log Scale.**
(http://www.nature.com/scitable/ blog/bio2.0/high_throughput_sequencing_and_cost)

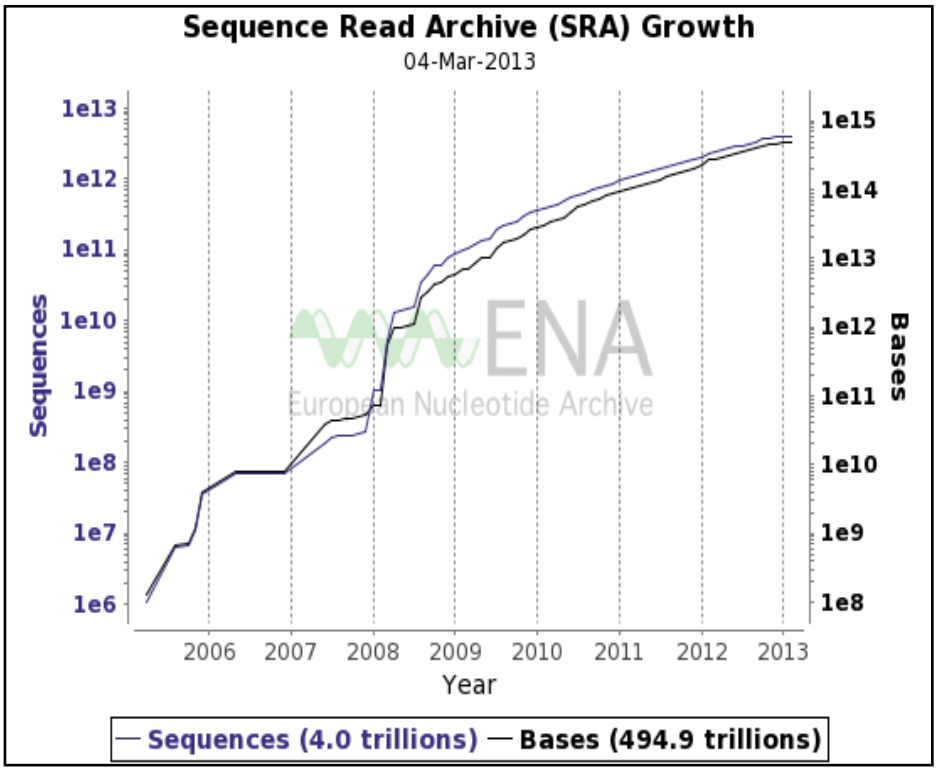



**Figure 2 Sequence Read Archive (SRA) Growth.**
Growth rates of database sequences and their corresponding number of bases in the SRA shown In log scale. (Http://Www.Ebi.Ac.Uk/Ena/About/Statistics)

High-throughput sequencing without the traditional cloning step offered by next-generation sequencing has been applied to various research areas, including but not limited to global analysis of non-coding RNAs (Morin *et al*. 2008), mutation discovery such as SNPs or indels (Barcia *et al*. 2012), identification of DNA and protein interactions (Grober *et al*. 2011), finding epigenetic modifications of histones and DNA (Eckhardt *et al*. 2006, Esteller, 2006, Callinan and Feinberg, 2006) etc.

From each of this application, a large amount of raw data is produced, therefore the bottleneck in sequencing shifted from sequence generation to data management and analysis. Consequently, an efficient computational infrastructure is needed to process and storage data as well as comprehensive and reliable analysis pipeline that enabled the correct understanding of results.

This thesis focuses on two of the most widely used applications in biological and clinical field respectively: Small RNA Seq and Exome Sequencing.

# 2. SMALL RNA SEQUENCING

## 2.1. MICRORNAS

MicroRNAs (miRNAs) are short (~22 nucleotides) RNA molecules that control gene expression in eukaryotes by fine tuning mRNA translation (Bartel, 2004, He, Hannon, 2004, Flynt, Lai, 2008). They play pivotal roles in a variety of molecular processes, such as immune response (Tili *et al*., 2007), differentiation (Tay *et al*., 2008), development (Lagos-Quintana *et al*., 2001, Lau *et al*., 2001, Lee, Ambros, 2001), infection (Gupta *et al*., 2006, Jopling *et al*., 2005) and cancer (Huang *et al*., 2008, Silber *et al*., 2008, Paris *et al*., 2012). miRNA genes are synthesized as long precursor RNA molecules (pri-miRNAs), generally by RNA polymerase II (Lee *et al*., 2004), that are rapidly processed in the nucleus by the Drosha RNase III to release approximately 70 nucleotides long miRNA precursor stem loops (pre-miRNAs) (Lee *et al*., 2003) that, in turn, are exported to the cytoplasm by Exportin 5 (Lund *et al*., 2004). In the cytoplasm, mature miRNAs are produced through the action of Dicer RNase (Hutvagner *et al*., 2001). These small RNAs regulate gene expression by binding to target sites generally in the 3' untranslated region (3'UTR) of target mRNAs, resulting in mRNA degradation or translation inhibition (Bartel, 2004, Nilsen, 2007). Identification by miRNAs of the 3' UTR of their target mRNAs is mediated by complementary hybridization between nucleotides 2-8 at the 5'end of the small RNA (seed sequence) and the complementary sequences present in the 3'UTR of the mRNA (Bartel, 2004, Ambros, 2004, Zamore, Haley, 2005).

## 2.2. iMIR: AN INTEGRATIVE PIPELINE TO ANALYZE MIRNA-SEQ

Due to advancements in the high-throughput sequencing (HTS) technologies, small RNA-Seq represents a rapid and an effective way to gather relevant biological information from small non-coding RNAs (sncRNAs), comprising of miRNAs and other regulatory transcripts. This technology allows to investigate, at unprecedented sensitivity, both known and novel sncRNAs by combining sequence output analysis with other biological and bioinformatics knowledge sources.

Commonly, miRNA-Seq data analysis is performed by combining multiple statistical and bioinformatics tools available from different sources. Many useful programs for processing these data exist nowadays, such as RandA (Isakov *et al*., 2012), DSAP (Huang *et al*., 2010), miRTools (Zhu *et al*., 2010) and miRExpress (Wang *et al*., 2009). Two main issues hamper diffusion and implementation of such programs: (i) web-based tools have some restriction on data upload; (ii) stand-alone programs often lack one or more analysis steps, such as prediction of novel sncRNAs.

As a consequence, the analytical workflow is slowed down by the need for the continuous interventions by the operator, a critical factor when a large number of samples need to be analyzed at once. To solve this problem, a novel modular pipeline, called iMir (Giurato *et al*., 2013 *submitted*), has been developed, that allow a comprehensive analysis of miRNA-Seq data integrating multiple open source modules and resources, linked together in an automated flow. iMir workflow is designed to carry out linker removal and sequence quality controls, identification of expressed sncRNAs, differential expression analysis, detection of novel miRNAs and, in addition, it provides the possibility to predict, for expressed miRNAs -or, alternatively, only for those differentially expressed, the corresponding mRNA targets. The pipeline described here allows at present to analyze miRNA-Seq data from human, mouse, rat and Drosophila, comprising in its database all corresponding libraries, and it can

be applied to FASTQ data generated by Illumina (GAIIx, HiSeq and HiScanSQ), Roche (FLX System) and Life Technology (IonTorrent PGM) next generation sequencers.

## 2.3. WORKFLOW DESCRIPTION

The main disadvantage of the very large datasets generated with currently available Next Generation Sequencing (NGS) technologies is their management and analysis. The iMir tool was developed to help solve this problem when dealing with miRNA-Seq data analysis. It is a stand-alone tool, that integrates several open source modules and resources, set together in an automated flow which provides a user friendly and efficient solution for miRNA-Seq data analysis. iMir is applicable to single or multiple sample analyses, computes the statistical significance of differential expression when case/control small RNA libraries are analyzed (both with or without biological/technical replicates), performs miRNA target prediction and allows to predict novel miRNAs. Furthermore, this tool can be easily adapted to large-scale miRNA-Seq analyses on multiple samples, such as, for tumor typing or other oncogenomics applications, producing a raw table with corresponding read-count values for miRNAs expressed in all or selected sequenced samples that can be given as input to bioinformatics tools able to perform cluster analysis. The iMir workflow is shown in **Figure 3**, where all required and optional data analysis steps are displayed. In detail, the depicted flow takes in account two common situations that an user may have to deal with: one in which it is necessary to analyze case/control samples (with or without replicates) and an alternative, where just a sample or multiple independent samples need to be analyzed. In the first case, for a full analysis run, Steps 1 to 6 (including also optional processes) are performed, while in the second case, iMir skips Step 6, performing all remaining tasks. It is worth mentioning that iMir allows also to analyze in an unique run also a combination of these two conditions. As miRNA detection is, at present, the main focus of most miRNA-Seq analyses, let's focus on iMir performance for this application. Reads obtained from NGS are usually longer than mature miRNAs (~22 nt) and contain part of the 3' adapter that do not allow reads to align against the reference track. First essential task to carry out during analysis of such data is thus to find/identify the reads that contain adapter sequences and to remove such sequence information to allow reads to properly align against the reference track. Step 1 of iMir is the adapter removal and quality filtering, performing stringent quality filtering e.g. remove reads with low quality (representing reads affected by sequencing errors), with poly-A tail and/or without adapter sequences (both cases refer to fragments derived from RNA degradation) by using a PERL script "Adapter_trim.pl", derived from miRTools (Zhu *et al.*, 2010) website, before proceeding to Step 2.

In Step 2 of the iMir workflow, the output file from Step 1 is converted into a tab delimitated format (each line of the file is represented by the sequence and its corresponding read-count). This format is required for Step 3 and this conversion is performed by using the PYTHON script "Format_conversion.py", that in addition allows the user to set a read-count cutoff (minimum expression level) for detecting known sncRNAs. In addition, iMir provides also the possibility to investigate changes in reads length after adapter sequence removal by generating a histogram that allows to easily evaluate read length distribution after this operation.

In order to predict novel and identify known miRNAs (Step 3) iMir uses the miRanalyzer stand-alone version 0.3 that allows, to align reads against other libraries of transcriptome and sncRNAs and to perform, in this way, prediction of potential novel miRNAs in the sample.

Differential expression analysis module on known miRNAs (Step 4), exploits of the DESeq package (Anders, Huber, 2010), which computes the fold-change of the expression values and assesses its statistical significance. During this Step, iMir creates two graphs for a more comprehensive data interpretation: an heatmap showing the Euclidean distances between the samples (a dendrogram representing a hierarchical clustering), as calculated from the variance-stabilizing transformation of the count data, and a circle graph, displaying the percentage of miRNAs significantly up- or down- regulated and of those that have no statistical significant fold-change.

Steps 5 and 6 of the workflow perform mRNA target prediction. iMir allows to predict, for miRNAs detected in the sequenced samples and/or for those differentially expressed, the corresponding putative mRNA targets, using data retrieval in TargetScan (Lewis *et al.*, 2005, Grimson *et al.*, 2007, Friedman *et al.*, 2009, Garcia *et al.*, 2011) and microRNA.org (Betel *et al.*, 2010, Betel *et al.*, 2008). Starting from miRBase version 17, the miR/miR* nomenclature has been modified in favor of a -5p/-3p nomenclature. The latest miRBase version available to date (e.g. v19) uses modified nomenclature for all species. While reference tracks in miRanalyzer refer to miRBase v19, TargetScan and microRNA.org are still based on previous versions of this database, making an automatic integration impossible. To overcome the problem represented by the different miRNA nomenclatures used in the different versions of miRBase (Kozomara, Griffiths-Jones, 2011, Griffiths-Jones *et al.*, 2006, Griffiths-Jones *et al.*, 2008), iMir uses "miRNA_name_conversion.py" (Step 5), a new script that performs name conversions according to the miRBase release to be used for the subsequent analysis. Step 6 of the workflow searches then in TargetScan and

microRNA.org (local files) the putative mRNA target for each miRNA of interest, obtained with the previous analytical steps.

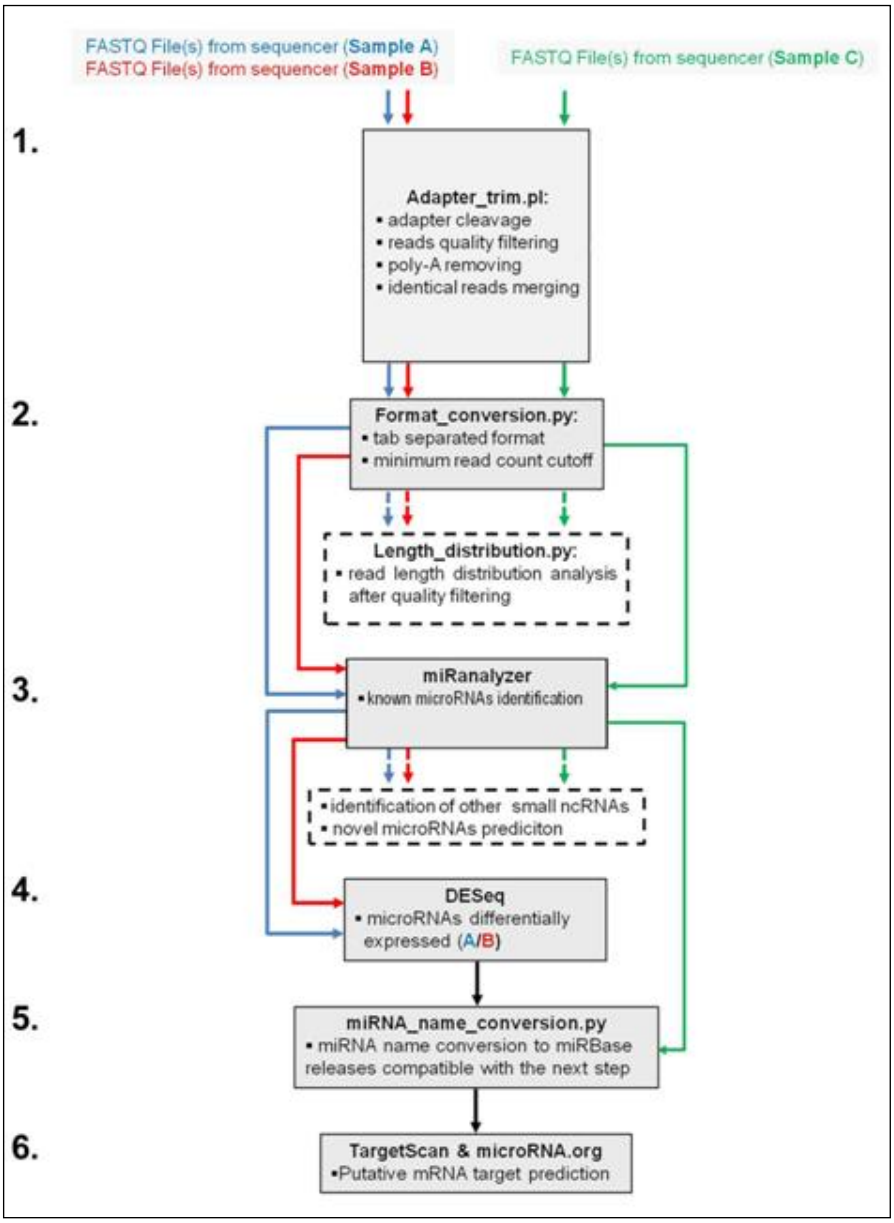



**Figure 3 iMir workflow.**
Graphic summary of iMir workflow: the pipeline accepts NGS data as input and then proceeds automatically to perform several independent analyses, most of which can be selected or excluded according to the user's needs. Dotted lines represent optional steps of the pipeline.

At present, iMir is a command line pipeline, while in a new version a graphic interface will be provide. Meanwhile a sample sheet (**Table 1**) is provided to help the user step by step.

| WorkFlow | |
|---|---|
| *Adapter_AND_Quality_Filtering*<br>*Length_Distribution*<br>*miRanalyzer*<br>*Differential Expression*<br>*Target* | Y(Yes)/N(No)<br>Y/N<br>Y/N<br>Y/N<br>Y/N |
| **Clustering** | |
| *all selected samples* | |
| **SampleData_Name** | |
| SAMPLE_NAME<br>Sample 1<br>Sample 2<br>...<br>Sample n | CLUSTERING<br>[Y/N]<br>[Y/N]<br>[Y/N]<br>[Y/N] |
| **Preprocess_Analysis** | |
| *Input_dir* | /path/to/input/directory |
| *Output_dir* | /path/to/output/directory |
| *Adapter* | 3' Adapter sequence |
| *fastq_format* | 1:Sanger format 2:Solexa/Illumina 1.0 format 3:Illumina 1.3+ format |
| *Read_count* | value (minimum read_count_cutoff) |
| **miRanalyzer_Params** | |
| *JarPath* | /path/to/miRanalyzer jar file |
| *dbPath* | /path/to/miRanalyzerDB |
| *species* | assembly / basename of bowtie index, i.e. hg19,mm9,rn4, etc. |
| *speciesShort* | short name of species (this must be the abbreviation of the species used in miRBase. For example hsa for Homo Sapiens or mmu for Mus Musculus) |
| *kingdom* | plant/animal (In order to set the models and features for the prediction of new microRNAs the program needs to know whether the species is a plant or aninal) |
| *bowtiePath* | the path with the bowtie binaries |
| *justKnown* | true/false (If set on true just known microRNAs are detected ) |
| *noMM* | value (allowed number of mismatches to known microRNAs - default 1) |
| **DiffExp_Samples** | |
| CASE_SAMPLE_NAME | CONTROL_SAMPLE_NAME |
| Sample 1<br>Sample 3 | Sample 2<br>Sample 2 |
| **DiffExp_Params** | |
| *Pvalue_adj* | value (the corrected p-value cutoff from DESeq) |
| *Fold-change* | value (fold-change cutoff from DESeq) |
| **mapper_Params** | |
| *-p* | /path/to/indexed/genome (the 'genome'string must be the prefix of the bowtie index) |
| **Target_Samples** | |
| SAMPLES | miRNAset |
| *[all]* | |
| *Sample_1,Sample2* | diffExp/all |
| *Sample_3* | |

**Table 1 Schematic overview of Sample sheet used by iMir.**
Schematic overview of the Sample sheet that help the user to set parameters step by step

# 2.4. RESULTS AND DISCUSSION

To assess the iMir functionality, it was tested on 6 miRNA-Seq datasets of 5-7 million sequence tags/run, obtained from human breast cancer MCF-7 cell lines in two different culture conditions: growth-arrest and exponential growth (Cicatiello *et al*., 2010, Ferraro *et al*., 2011). For each condition considered, three sequencing technical replicates were performed to allow a correct estimate variability during the differential expression analysis. The iMir sample sheet was set to generate a full and comprehensive miRNA-Seq data analysis. The pipeline was executed on 64 bit Linux machine, where the analytical process, from pre-processing (adapter sequence removal and quality filtering), identification of known and novel miRNAs, differential expression analysis and mRNA target prediction was completed in about one hour. After the pre-process analysis, a small percentage of reads, all <15nt-long, is discarded as the algorithm is unable at present to manage them. The read-length distribution after adapter cleavage in all samples is reported to the right of Module 1 in **Figure 4**, to show how the majority of reads obtained after this first step are~22 nt long, suggesting that they are mainly due to miRNAs. This observation is further confirmed by the number of reads that actually match known miRNAs (**Table 2**), computed to account for more than 50% of the entire dataset in each case.

| MCF-7 cells | | Raw reads | Reads after adapter cleveage | miRNA reads | tRNA reads | rRNA reads | mRNA reads | piRNA reads | Remaining reads mapping to genome | Reads not assigned |
|---|---|---|---|---|---|---|---|---|---|---|
| Exponentially growing | *Replicate 1* | 4,327,501 | 4,068,141 | 2,310,200 | 16,989 | 91,040 | 391,750 | 15,753 | 597,037 | 69,516 |
| | *Replicate 2* | 4,337,535 | 4,075,320 | 2,314,040 | 17,042 | 92,178 | 404,148 | 16,438 | 614,614 | 70,165 |
| | *Replicate 3* | 4,354,046 | 4,091,633 | 2,374,218 | 17,737 | 94,961 | 420,175 | 16,949 | 636,708 | 71,737 |
| Growth-arrested | *Replicate 1* | 6,071,484 | 5,844,875 | 4,626,170 | 13,588 | 72,460 | 181,084 | 14,831 | 234,955 | 40,941 |
| | *Replicate 2* | 6,075,950 | 5,846,690 | 4,621,008 | 12,470 | 75,251 | 185,803 | 15,122 | 242,065 | 40,667 |
| | *Replicate 3* | 6,084,784 | 5,855,090 | 4,725,975 | 12,705 | 77,842 | 192,161 | 15,582 | 249,638 | 41,494 |

**Table 2. Number of reads before and after adapter cleavage and reads mapped in each sncRNA library included in iMir**

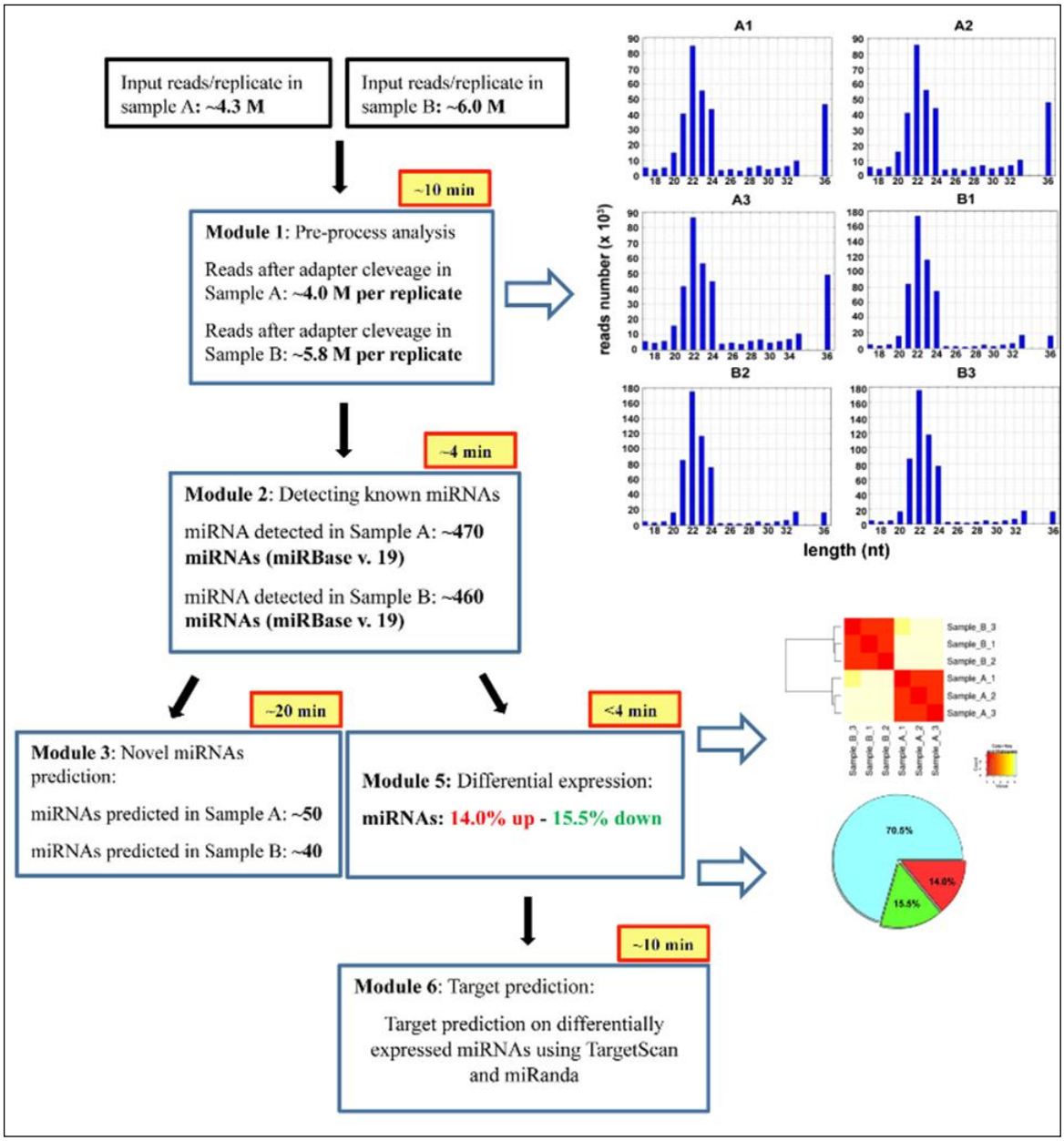



**Figure 4 Graphic representation of iMir pipeline performances.**

miRNA-Seq analysis in exponentially growing (sample A) or growth-arrested (sample B) MCF-7 cells, performed in triplicate as described in the text, were input in iMir and analyzed with the standard, complete analytical workflow of the tool. The processing time of each module are highlighted in yellow and the graphic outputs of Modules 1 (histograms showing sequence read length distribution in each replicate) and 5 (heat-map visualization of sncRNA profile differences among samples and pie-chart summarizing the results of the differential expression analysis) are shown to their right

iMir made possible to identify ~450 miRNAs for each dataset, including those differentially expressed between the two growth conditions tested and their putative mRNA targets. In addition, ~45 putative novel miRNAs were predicted among the samples together with several isomiRs (Morin *et al*., 2008).

These results, when compared with previously published data relative to miRNA modulation in the same cell line (Paris *et al*., 2012, Cicatiello *et al*., 2010, Ferraro *et al*., 2011) demonstrate how iMir, that makes possible a rational use of several widely used and reliable open source tools, is useful to rapidly and efficiently perform reliable sncRNAs data analysis.

For miRNA detection, iMir utilizes miRanalyzer (v.0.3) and implements the latest version of miRBase database v.19. This is an important feature, since this version of miRanalyzer provides the possibility to visualize alignments of the reads to precursor together with prediction of pre-miRNA secondary structure, to customize databases, to perform isomiR analysis by classifying them in different classes and to predict novel miRNAs.

Finally, for evaluation of sncRNA differential expression iMir utilizes DESeq package, the most performing and popular open source tool for this analysis that uses a negative binomial distribution model to estimate variance-mean dependence in count data from HTS experiments. Interestingly, this tool can be used also when the analysis has been performed without technical or biological replicates.

## 2.5. CONCLUSIONS

As conclusion, iMir is a pipeline that integrates multiple open source modules /resources combined in an automated flow for high-throughput miRNA-Seq data analysis. iMir performs sncRNA sequencing data analysis rapidly, accurately and efficiently, allowing to examine multiple samples at once and thereby addressing a critical factor in analysis of high-throughput sncRNA sequencing data, represented by the need for continuous interventions by the operator. The user-friendly compilation of the sample sheet is another of the advantages of iMir, allowing to customize data analysis according to different needs. iMir works on Linux and Mac operative systems with command line. In the future, based also on evolution of the NGS technologies and recommendations by users, we are determined to improve iMir features, such as for example by including GUI (Graphical User Interface) and tools for evolutionary sncRNA analysis across multiple species and specific analysis of different classes of small RNAs (pi-, si-, sn-, sno-, ti-RNA, etc).

# 3. EXOME SEQUENCING

## 3.1. VARIATIONS IN HUMAN GENOME

The total length of the human genome is over $3.1*10^9$ base pairs (bp) (Flicek *et al*., 2011), which enables an enormous potential for variation. Different kind of variations can be found in human genome from single base mutation also called SNV to structural variations that can be from 1 to 10kb in size, known as small structural variations, and from 10 kb to several Mb, known as large structural variations. A single base substitution can be defined where a single nucleotide is replaced by another nucleotide. These single base changes are also referred to as point mutations. These are the most frequent type of alterations in DNA. There are four categories of single base substitutions: missense mutations, nonsense mutations, silent mutations and splice site mutations.

**Missense mutation:** In a missense mutation, base alters the codon which results in a different amino acid being incorporated into the protein (**Figure 5**).

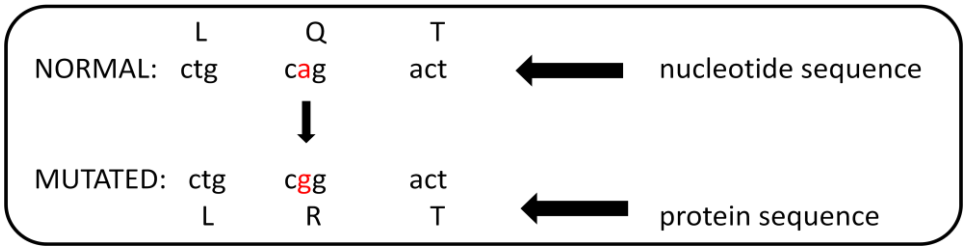



**Figure 5 Missense mutation**
A substitution of "a" (in red) in the second codon to "g" (in red), leads to an amino acid substitution of glutamine (Q) to arginine (R)

**Nonsense mutation:** In a nonsense mutation, the new base change in a codon that cause a stop codon (taa, tag, tga). This will cause translation of mRNA to stop prematurely and a truncated protein is produced (**Figure 6**).

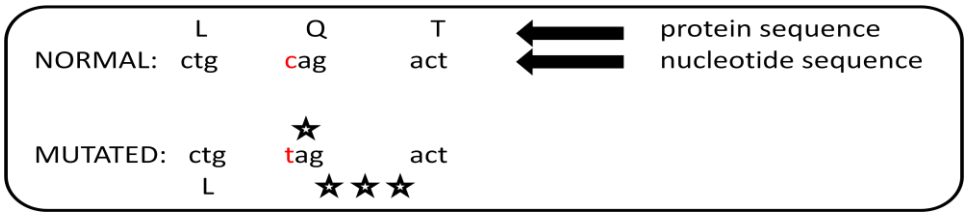



**Figure 6 Nonsense mutation**
A substitution of "c" (in red) in the second codon to "t" (in red), cause a change of glutamine (Q) to a STOP codon, leading to a premature termination of the protein.

**Silent mutation:** Silent mutations are those that don't cause any alteration in the final protein product and can only be identified by sequencing the gene. These mutations don't have any deleterious effect because they don't cause amino acid change.

**Splice site mutation:** Splice site mutation occur within genes in the noncoding regions (introns), just next the coding regions (exons). They can have deleterious effects on the resulting protein, which may lead to disease. Before mRNA leaves the nucleus, the introns are removed and the exons are joined together.

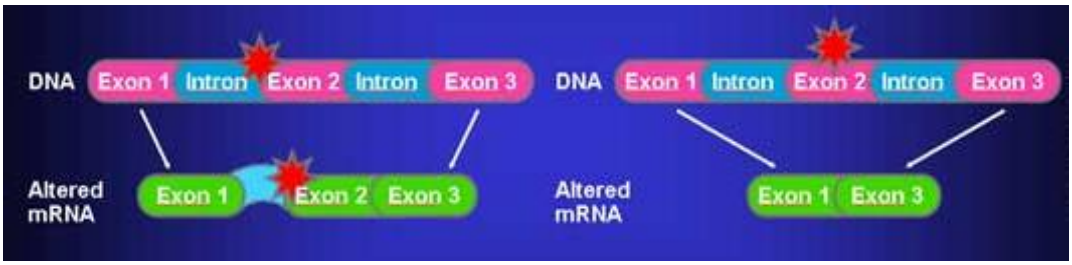



**Figure 7 Splice site mutation**
*(http://cancer.gov/cancertopics/understandingcancer).*

This process is calling splicing. Splicing is controlled by specific intron sequences, called splice-donor and splice-acceptor sequences, which flank the exons. Mutations in these sequences may lead to retention of large intronic DNA in the mRNA, or to entire exons being spliced out of the mRNA. These changes could result in production of a non functional protein (**Figure 7**).

**Structural variations:** When the variations expand in size to cover larger regions, they are referred to as structural variations (**Figure 8**); in particular, if they cover from 1 to 10 kb they are known as small structural variations, while if they cover from 10 kb to several Mb they are known as large structural variations. Structural variations can be divided in few categories. One of these, is copy number variation (CNV), which includes deletions, insertions and duplications. Another class is represented by translocations, where a part of

chromosome is integrated into another chromosome. Finally, there are the inversions, where a chromosomal section is inverted so that its start and end points switch places.

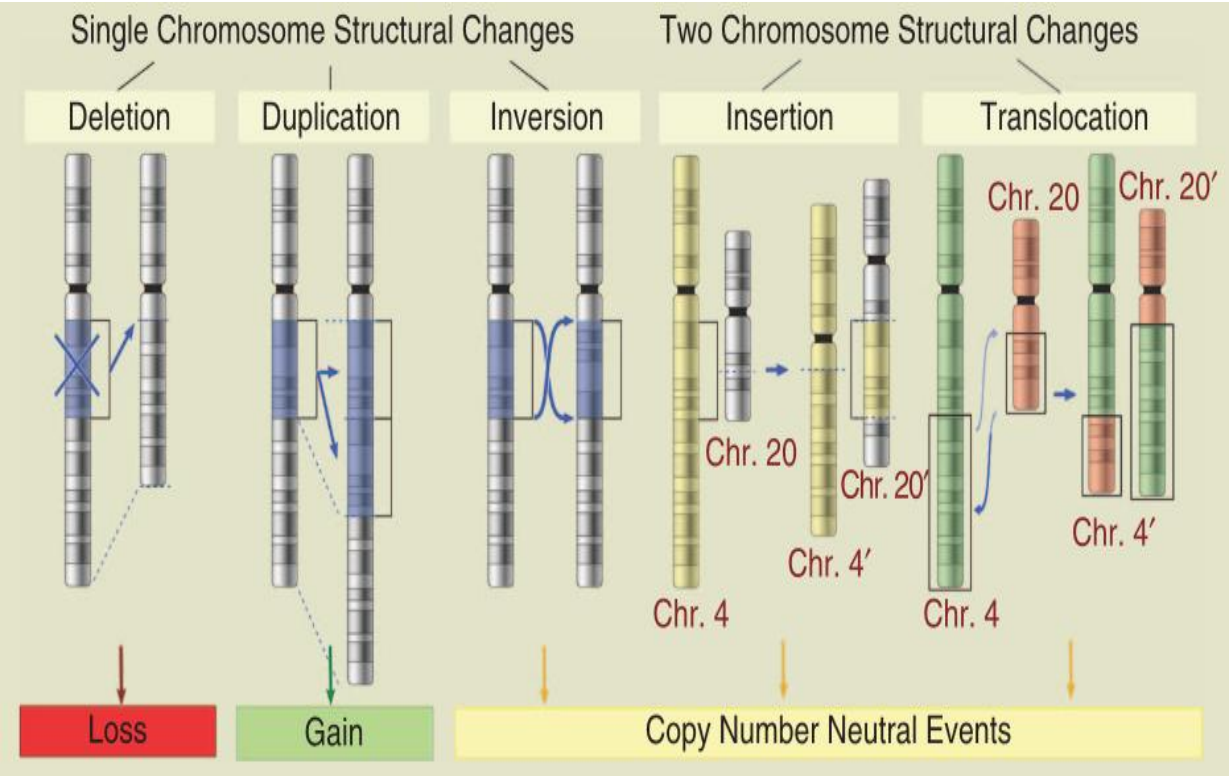



**Figure 8 Example of structural variations**
(Schwab et al., 1983)

# 3.2. TARGETED RE-SEQUENCING

Since its early days, medical research was aimed to identify the causes of disorders and establish therapeutic treatments and find cures. Because of the cost of whole genome sequencing the time required to analyze data and the more difficult interpretation of results, most of the research groups choose to use targeted re-sequencing, that isolates genomic regions of interest in a sample library, focusing on targets and mutations. Nowadays there are three major exome enrichment systems: Agilent Sure Select Human All Exon, Roche /Nimblegen SeqCap Ez Exome libraries and Illumina TruSeq Exome Enrichment. Although several methods exist, they all use a similar principle: they isolate a specific genomic fraction for subsequent NGS, ultimately resulting in an enriched pool of target sequences such that there is reduction in the genomic sequencing space, and hence greater sequence coverage for each targeted region (**Figure 9**, **Table 3**).

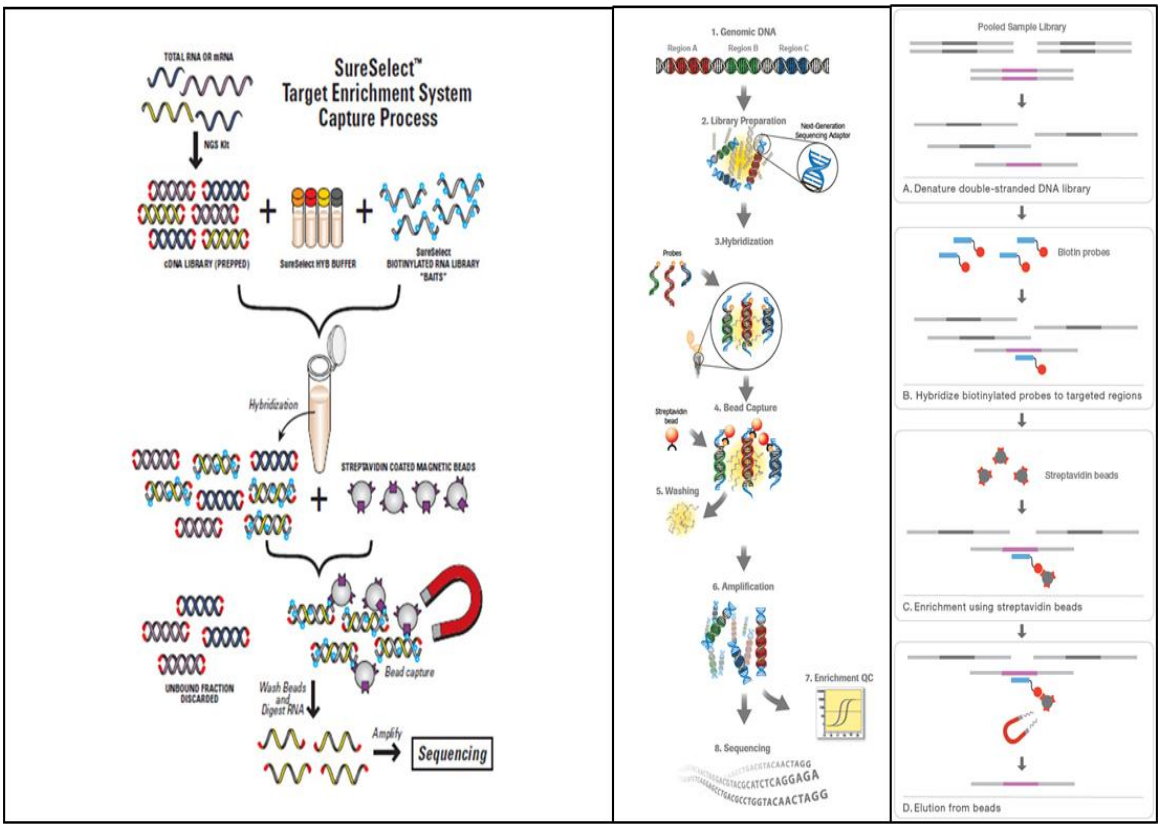



**Figure 9 Schematic representation of the three mayor enrichment protocol**
(www.genomics.agilent.com, www.nimblegen.com, www.illumina.com)

However, there are also inefficiencies in targeting process. For example, uneven capture efficiency across exons can produce DNA regions with low sequence coverage, and off-target hybridization means that at least 20% of reads come from genomic DNA outside the exome (Majewski *et al*., 2011). Furthermore, the probes in sequence capture methods are designed based on information from gene annotation databases, therefore, unknown or yet-to-be-annotated exons and other parts of the genome are not usually includes in capturing. Another important consideration is that NGS technologies have higher base calling error rates respect to Sanger sequencing, although this problem can be bypassed by increasing the depth of sequencing coverage to ensure minimal false calls (Koboldt *et al*., 2009). For this reason is important to validate variant genes using conventional sequencing techniques. All of these inefficiencies are gradually decreasing thanks to the continuous improvement of the sequencing and capture technologies. Importantly, the higher coverage of the exome that can be affordably achieved for a large number of samples makes exome sequencing highly suitable for mutation discovery and its use is becoming increasingly routine.

| | **SureSelect Human All Exon 50 Mb** | **TruSeq Human Exon** | **NimbleGen SeqCapture** |
|---|---|---|---|
| **Target Size** | **50 Mb** | 44 Mb | 62 Mb |
| **Targeted Regions** | **331,518** | ~241,693 | 201,121 |
| **CCDS (Nov.2010)** | **99.5%** | 98.4% | 97.0% |
| **RefSeqGenes (Nov.2010)** | **99.0%** | 98.2% | 96.0% |
| **GENCODE v.4** | **97.0%** | - | 93.0% |
| **Addition of Custom Exome** | **Yes** | No | No |
| **Number of Hybds. Required** | **1** | 1 | 2 |
| **Hyb. Time Required** | **24hrs** | 72hrs | 24+24hrs |
| **Insert Size** | **150-250bp** | 150-250bp | 300-400bp |
| **Amount of Seq. Required** | **5Gb** | 5Gb | 5-10Gb |

**Table 3 Three major exome enrichment platform**
Table shows (in bold) that SureSelect Human All Exon 50Mb gives you the most complete coverage of the coding content of the genome (www.genomics.agilent.com, www.nimblegen.com, www.illumina.com)

Through Exome Sequencing, mutations can be studied using two different approches, not mutually exclusive: the first, is to sequence the exome of a group of affected individuals searching for mutated gene(s) in common between all or most of them; the second is to sequence one or more parents-affected child trios to identify de novo mutations present only in the patient(s).

## 3.3. MALIGNANT MIGRATING PARTIAL SEIZURES IN INFANCY: A RARE DISEASE OF UNKNOWN ETIOLOGY

Malignant migrating partial seizures in infancy (MMPSI, MMPEI), described for the first time in 1995 by Coppola and colleagues, is a rare, severe early infantile epileptic encephalopathy, with a devastating course. Since, almost 80 cases have been reported worldwide, with males and females being equally affected. The main features of this syndrome are: (i) normal development before seizure onset, (ii) first manifestations of the pathology occur usually between 40 days and 3 months of age (range 1 day-6 months), (iii) multifocal bilateral independent seizures with ictal electroencephalogram (EEG) discharges arise from different areas of both hemispheres of the brain and "migrate" from one brain region to the other, giving the main features and name to this syndrome (**Figure 10**) (Coppola *et al*., 1995). Seizures are intractable to conventional antiepileptic drugs, causing in affected infants progressive psychomotor retardation and decline of head circumference percentile.

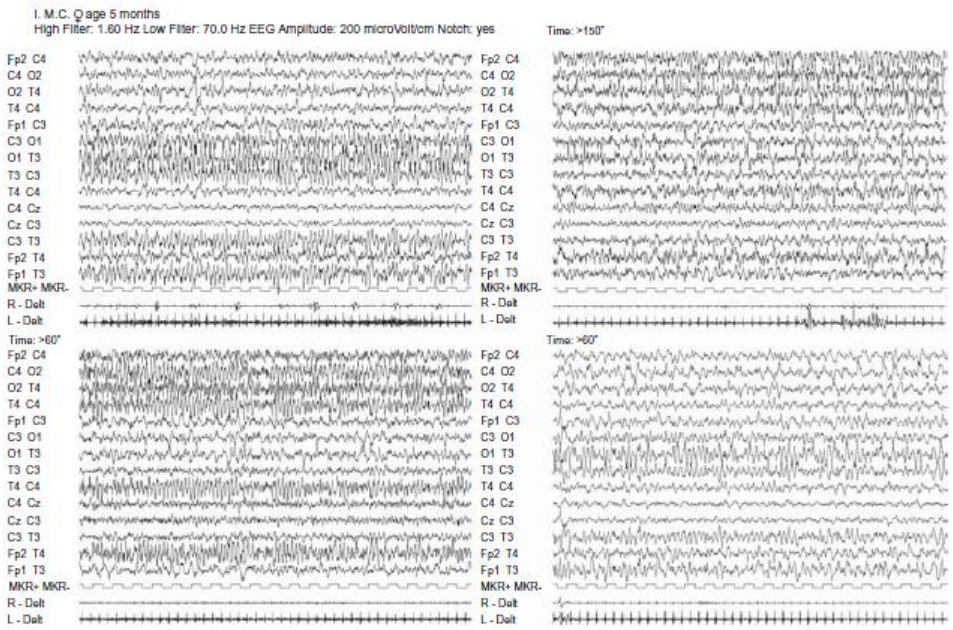



**Figure 10 Ictal Electroencephalogram in a 5-month old patient.**
A 5-month-old patient with long lasting and often difficult to identify seizures, the onset of which migrates from one cortical area to the other: left temporo occipital to right centro-temporal to right temporo-occipital to left temporal. (Coppola *et al.*, 1995).

Three distinct phases are described in the natural history of this syndrome, as described by Coppola *et al*., 1995. The first phase is heralded by the seizures beginning between the first week of life and 7 month. Seizure onset may also occur since the first day of life, but the mean age is 3 month (Hmaimess *et al*., 2006). At onset, seizures are mainly focal motor involving one limb or one side, with frequent secondary generalization. Autonomic manifestations, such as apnea, flushing or cyanosis frequently occur. This phase usually lasts a few weeks or months. In a small number of patients seizures at onset may consist of status epilepticus (Coppola *et al*., 1995; Wilmshurst *et al*., 2000; Gross-Tsur *et al*., 2004; Marsh *et al*., 2005; Zamponi *et al*., 2008); in such cases, the syndrome begins directly with the second phase.

The second phase, also defined as "stormy phase", starts at an age ranging from 3 weeks to 10 months. During this period, seizures become very frequent and polymorphous, occurring in multiple clusters in a day or being almost continuous for more days. Milder clinical manifestations may be easily overlooked by parents, and detection of frequent subclinical ictal manifestations is made possible only by long lasting video-EEG recordings. Sometimes, such stormy periods may be shortened or interrupted by a particularly favorable response to drugs. Finally, secondary tonic-clonic seizures tend to become more frequent; epileptic spasms are generally very rare in this syndrome.

The third phase begins between 1 and 5 years of age and over. It is relatively seizure-free, although spontaneous intercurrent illnesses would easily trigger clusters of seizures or occasional status epilepticus (Dulac, 2005).

The long-term outcome remains very poor in most patients. Most cases have psychomotor retardation and acquired microcephaly along with continuing seizures. Rarely, when seizures are controlled, using different drug combinations, children may acquire the ability to reach for objects and walk, but language is generally absent. A number of patients die before the end of the second year of age or later, during the follow-up, mainly because of prolonged status epilepticus and/or respiratory insufficiency.

### 3.3.1 ETIOLOGY

The etiology of this pathology is not yet known, even if the involvement of a gene mutation coding for some ion channel cannot be excluded. A first attempt of genetic search took in consideration KCNQ2, KCNQ3, SCN1A, SCN2A, CLCN2 and MECP2 genes (Coppola *et al*., 2005). Mutational analysis of these genes was performed on three patients with clinical symptoms of MMPSI, but no genetic abnormality in the coding regions of sodium, potassium and chloride ion channel genes was found. Subsequently, because of some features suggest a genetic basis for this syndrome, several research groups looked for causative gene mutations. Recently, Freilich *et al*. (2011) and Carranza Rojo *et al*. (2011) identified missense mutation and deletion of the sodium channel gene SCN1A, while Bedoyan *et al*. (2010) reported a patient with duplication in the region 16p11.2 of chromosome 16. Furthermore, in the last year, Poduri *et al*. (2012) described an homozygous PLCB1 deletion in a child with MMPSI while Barcia *et al*. (2012) identified distinct mutations in the sodium-activated potassium channel KCNT1, in 6/12 affected individuals (**Table 4**).

| Gene/Locus | Chr | Type of variation | Mutation | Deletion size | Protein alteration | Reference |
|---|---|---|---|---|---|---|
| **SCN1A** | 2 | de novo missense mutation | c.2584C>G | | p.R862G | *Freilich et al, 2011; Carranza Rojo et al, 2011* |
| **q24.2q31.1** | 2 | Deletion | | 11.6Mb | | |
| **SCN1A** | 2 | heterozygous missense mutation | c.5006C>A | | p.A1669E | |
| **16p11.2** | 16 | Duplication | | 598kb | | *Bedoyan et al, 2010* |
| **PLCB1/20p12.3** | 20 | Deletion | | ~476kb | | *Poduri et al, 2012* |
| **KCNT1** | 9 | heterozygous missense mutation | c.2800G>A | | p.A934T | *Barcia et al, 2012* |
| **KCNT1** | 9 | heterozygous missense mutation | c.1283G>A | | p.R428Q | |
| **KCNT1** | 9 | de novo mutation | c.1421G>A | | p.R474H | |
| **KCNT1** | 9 | de novo mutation | c.2280C>G | | p.I760M | |

**Table 4 Genes Associated with MMPSI**

# 3.4. BIOINFORMATIC ANALYSIS FOR EXOME SEQUENCING DATA

## 3.4.1 PHRED QUALITY SCORE

The Phred quality score was used to automatically identify the DNA sequence and assign quality score for each base from DNA sequencing trace file. The Phred quality score has become a highly acceptable standard and is defined by logarithmical transformation of probability of sequencing error (**Table 5**).

| Phred Quality Score | Probability of incorrect Base Call | Base Call Accuracy |
|---|---|---|
| **10** | 1 in 10 | 90% |
| **20** | 1 in 100 | 99% |
| **30** | 1 in 1000 | 99.9% |
| **40** | 1 in 10000 | 99.99% |
| **50** | 1 in 100000 | 99.999% |

**Table 5. The Phred quality scores and their corresponding error rates.**
(Ewing and Green, 1998)

## 3.4.2 FILE FORMAT

### 3.4.2.1 FASTQ FORMAT

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. For better storage and interpretation, both the sequence letter and quality score are encoded with a single ASCII character. FASTQ format also contains a single line before the content of sequence beginning with "@" followed by the sequencing identifier and an optional description. Other three lines contain the raw sequence letters, a "+" character and the quality values for the sequence in Line 2 respectively. The last line must contain the same number of symbols as letters in the sequence. This is an example of FASTQ format:

*@SEQ_ID*
*GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT*
*+*
*!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65*

### 3.4.2.2 SANGER FASTQ FORMAT

Welcome Trust Sanger Institute first introduced the Sanger FASTQ format to combine sequence data and its corresponding quality. Early FASTQ file were used for Sanger capillary sequencing, and it was natural to use PHRED quality scores. Storing PHRED scores as characters was very easy to understand but required too much space. In order that the file be human readable and easily edited, this restricted the choices to the ASCII printable characters 32–126 (decimal), and since ASCII 32 is the space character, Sanger FASTQ files use ASCII 33–126 to encode PHRED qualities from 0 to 93. The wide range of error probability from $10^{-9.3}$ to 1 makes it easy to be adopted in raw sequence data storage and post processing for relative high base quality (Cock *et al*., 2010).

### 3.4.2.3 SAM AND BAM FORMAT

As the wide availability of next generation sequencing, many alignment tools have been developed for locating raw reads to their origins in the reference genome. The Sequence Alignment/Map (SAM) format describes the alignment of query sequences or sequencing reads to a reference sequence or assembly. It is a tab-delimited text based file and consists of one header section and one alignment section (Heng *et al*., 2009). Head should be in front of alignment part and start with "@". The alignment part contains 11 mandatory fields, for incorporating essential alignment information, and a variable number of optional fields. **Table 6** shows the mandatory fields in the SAM format.

| No. | Name | Description |
|---|---|---|
| **1** | QNAME | Query NAME of the read or the read pair |
| **2** | FLAG | Bitwise FLAG (pairing, strand, mate strand, etc.) |
| **3** | RNAME | Reference sequence NAME |
| **4** | POS | 1-Based leftmost POSition of clipped alignment |
| **5** | MAPQ | MAPping Quality (Phred-scaled) |
| **6** | CIGAR | Extended CIGAR string (operations: MIDNSHP) |
| **7** | MRNM | Mate Reference NaMe ('=' if same as RNAME) |
| **8** | MPOS | 1-Based leftmost Mate POSition |
| **9** | ISIZE | Inferred Insert SIZE |
| **10** | SEQ | Query SEQuence on the same strand as the reference |
| **11** | QUAL | Query QUALity (ASCII-33=Phred base quality) |

**Table 6 Mandatory fields in the SAM format**
**(***Heng et al., 2009***)**

The conventional CIGAR format allows for three types of operations: M for match or mismatch, I for insertion and D for deletion, while the extended CIGAR format futher allows four more operations (**Table 7**), to describe clipping, padding and splicing: SAM files and its corresponding binary format, binary Alignment/Map (BAM) format, contain the same information.

| Op. | Description |
|---|---|
| **M** | Alignment match (can be a sequence match or mismatch |
| **I** | Insertion to the reference |
| **D** | Deletion from the reference |
| **N** | Skipped region from the reference |
| **S** | Soft clip on the read (clipped sequence present in <seq>) |
| **H** | Hard clip on the read (clipped sequence NOT present in <seq>) |
| **P** | Padding (silent deletion from the padded reference sequence) |

**Table 7 Operations in CIGAR format**

### 3.4.2.4 VCF FORMAT

The Variant Call Format (VCF, **Figure 11**) is a standardized text file format for storing the most prevalent type of variations SNP, indel, and structural variation calls.

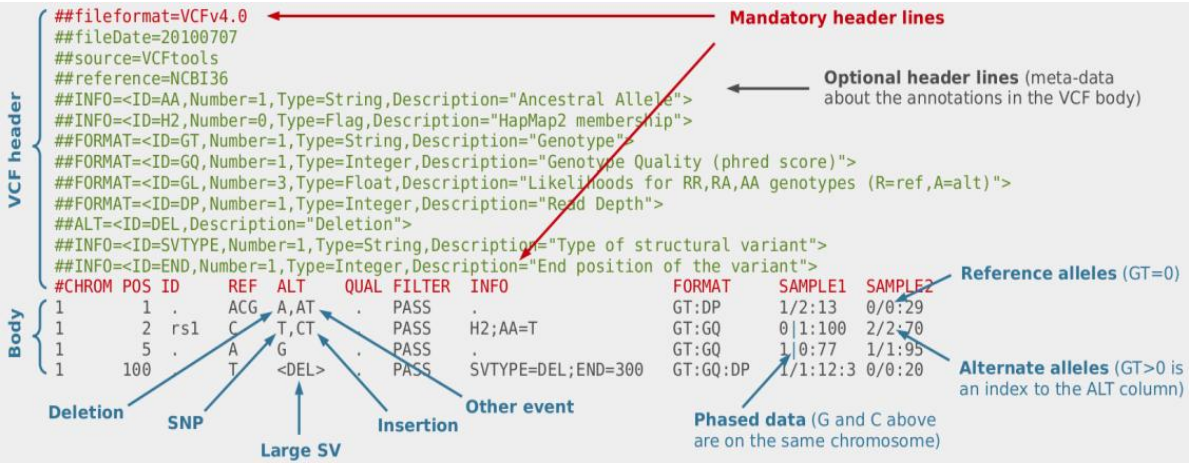



**Figure 11 Vcf Format**
(http://bioinf.comav.upv.es/courses/sequence_analysis/snp_calling.html)

It is the primary (and only well-supported) format used by the GATK for variant calls and consists of a header section and a data section. The header contains an arbitrary number of meta-information lines, each starting with characters '##', and a TAB delimited field definition line, starting with a single '#' character (Danecek *et al*., 2011).

# 3.5. PATIENT SELECTION AND METHODS

## 3.5.1 STUDY DESIGN

Six individuals presenting clinical and EEG features of MMPSI were selected. The patients, 5 male and one female, were all Italian except one and fulfilled the criteria of the syndrome (**Table 8).** All children were born full-term and were normal except one that showed a mild perinatal cyanosis. Familial history was positive for epilepsy in two patients and for febrile seizures in one. Psychomotor development was normal in all patients before seizure onset that occurred during the first month of life, with an average age of 4 months. Seizures were refractory to various antiepileptic drugs that were used mostly in combination. All children developed a severe neurological impairment after seizure onset. Computed tomography (CT) and Brain MRI was normal in four patients, while two showed a mild enlargement of lateral ventricles. EEG for all individuals showed the typical features of malignant migrating partial seizures, after an average period of 2.5 months (range 2 weeks to 4 months). All children showed profound delay both psychomotor and mental.

| Patient sex | Family history positive for epilepsy/febrile | Pregnancy and delivery | Age at seizure onset | Psychomotor development before seizure | CT/MRI | Drug resistance | Clinical outcome at last follow-up |
|---|---|---|---|---|---|---|---|
| 1 M | No | At term, normal | 4 m | Normal | Normal | present | Severe psychomotor delay/mental |
| 2 M | No | Mild perinatal cyanosis | 2 m | Normal | Mild enlargement of lateral | present | Severe psychomotor delay/mental |
| 3 M | Yes (maternal uncle) | At term, normal | 5 m | Normal | Normal | present | Severe psychomotor delay/mental |
| 4 M | No | At term, normal | 1 m | Normal | Mild enlargement of lateral | present | Severe psychomotor delay/mental |
| 5 F | Yes (paternal uncle) | At term, normal | 9 m | Normal | Normal | present | Severe psychomotor delay/mental |
| 6 M | Yes (mother with febrile seizures) | At term, normal | 3 m | Normal | Normal | present | Severe psychomotor delay/mental |

**Table 8 Clinical overview of patients with MMPSI.**
"M" is male; "F" is female and "m" is month.

## 3.5.2 DNA ISOLATION

For DNA isolation 5 ml whole blood was collected in the EDTA vacutainer from the patients participating in the study and salting-out protocol was used. For the osmotic lysis of the erythrocytes, blood was transferred into 50 ml Falcon tubes, about 25 ml (20-30 ml) NaCl 0.2% was added, shaked gently and incubated at 4°C for 15 minutes. After that, samples were centrifugated at 2500 rpm for 15 minutes and then the supernatant was carefully tipped off to waste leaving the pellet stuck to the bottom of the tube. Steps of centrifugation and the subsequent one were repeated to get a cleaner/whiter/purer pellet.
Next, to each Falcon, 2 ml of WBC lysing solution was added and cell lysates were digested overnight at 37°C using for each ml of lysing buffer 65 µl of SDS 20% and 10 µl of protease K solution. After digestion was complete, 670 µl of saturated NaCl for each ml of lysing buffer was used, shaked vigorously for few seconds and then centrifugated at 2500 rpm for 15 minutes at 4°C. The precipitated protein pellet was left at the bottom of the tube and the supernatant containing the DNA was transferred to another 15 ml tube. Exactly 2 volumes of room temperature absolute ethanol were added and the tubes inverted several times until the DNA precipitated. The precipitated DNA strands were removed with a sterile Pasteur, washed with ethanol 70% and dried. Finally the pellet was dissolved in an appropriate amount of sterile water and preserved at $4^{o}$C.

## 3.5.3 EXOME SEQUENCING

### 3.5.3.1 TARGET ENRICHMENT AND HIGH THROUGHPUT SEQUENCING

3µg DNA from each of the six patients was used to perform massively parallel sequencing. After shearing using Bioruptor sonicator (95 cycles), the DNA fragments were blunt end repaired and adenylated at the 3' end of fragments. Illumina adaptors were then ligated and fragments size selected for 350-400 bp products before amplification and validation using the Agilent Bioanalyzer. Exome capture was performed using the SureSelect Human All Exon 50Mb kit. DNA libraries were hybridized to streptavidin exome capture beads, washed, hybridized a second time, eluted and amplified. Quality was assessed using Bioanalyzer, Qubit and Real Time PCR and finally 72 paired-end enriched libraries were

sequenced-by-synthesis using an Illumina Genome Analyzer IIx according to Illumina Protocols.

### 3.5.3.2 BASIC DATA PROCESSING

Primary data analysis includes image analysis, base calling and demultiplexing to obtain fastq raw sequence files.

The first two steps are automatically done by the GAIIx sequencer using the GApipeline, which implements image analysis (Firecrest) and base-calling (Bustard).

There are four dominant source of noise affecting the intensities generated by the illumine sequencer:

- *Crosstalk*: the intensity channel are not independent. This is due to the fact that fluorescent markers for A,C and G,T emit photons with similar wavelengths and get excited by the same laser. Moreover fluorescent markers from one cycle can only be chemically partially removed before the cycles for the next nucleotides.
- *Fading*: with successive cycles, the absolute intensity of light emitted from the cluster of DNA strands decreases because fluorescent markers can only bind to fewer and fewer strands within the clusters.
- *Phasing*: in any given cycle, a few DNA molecules within a cluster may fail to extend. This is referred to as "phasing", as the molecules which fail to extend are now out of phase with the rest of the cluster. Phasing can occur when a base is not incorporated so that the resultant strand will be a base behind.
- *Pre-phasing*: occurs when an unblocked base is incorporated in the growing DNA molecule, allowing a second nucleotide to be incorporated. The resultant strand will be then a base ahead (**Figure 12**).

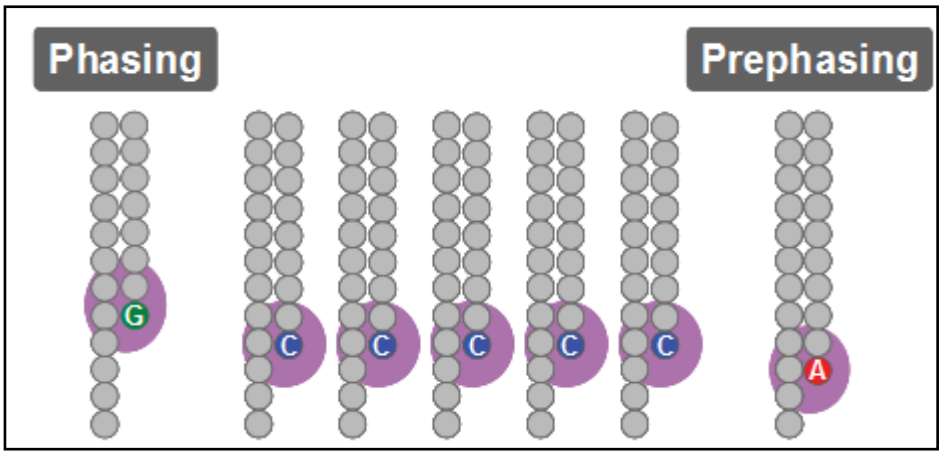



**Figure 12 Phasing and Prephasing Example**
( http://www.illumina.com)

After the corrections have been done, the base with the highest intensity is chosen (**Figure 13**). For quality control, a sample of the bacteriophage genome is included in one of the eight lanes of the flow cell (Ledergerber et al., 2011, Menges et al., 2011)

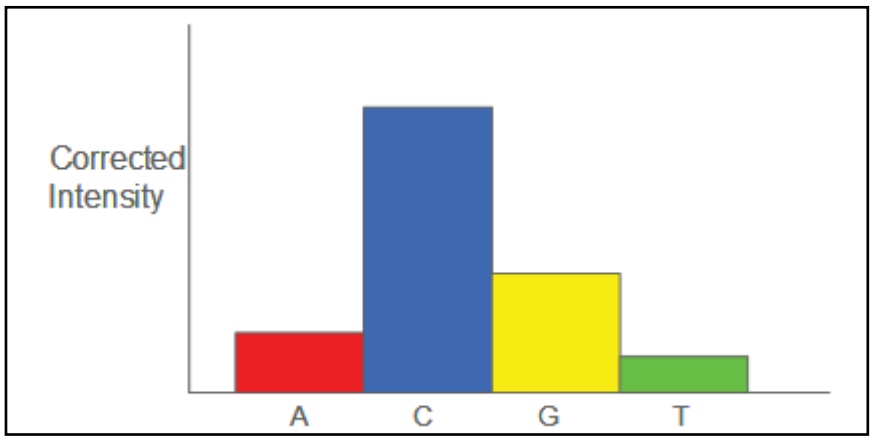



**Figure 13 An example of a corrected intensity plot**
(http://www.illumina.com)

Demultiplexing step is performed using the Perl script *configureBclToFastq.pl*. It takes as input .bcl files produced by Bustard and a text file, called sample sheet, that contains all information about sequenced samples. This procedure create different folders, one for each sample, in which put the fastq files.

***/path/to/CASAVA/bin/configureBclToFastq.pl --input-dir <BaseCalls_DIR> --output-dir <Unaligned> --sample-sheet <Input DIR>/SampleSheet.csv***

cd /path/to/RunFolder/Unaligned

***nohup make –j <n>***

### 3.5.3.3 FASTQ QUALITY CONTROL

To assess the quality of raw sequences, FASTQC tool was used that provides basic quality statistics and creates a comprehensive report by looking at the composition and quality of high throughput sequence library (**Figure 14**) (Andrews, 2010).


**Figure 14 Fastq quality report**
An example of statistics for one of our sample shows for all fields and in particular for "per base sequence quality " quality is very good.

### 3.5.3.4 ALIGNMENT

For alignment step BWA was used. It is a software package for mapping sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. For this experiment the first algorithm was used that is invoked with different sub-commands: aln/sampe. The first command finds the suffix array (SA) coordinates of good hits of each individual read while sampe command converts SA coordinates to chromosomal coordinates and pairs reads (Li and Durbin, 2009).

***bwa aln -n 0.06 -t 4 -q 5 reference_genome.fa fastq_file1 > outputfile1.sai***

***bwa aln -n 0.06 -t 4 -q 5 reference_genome.fa fastq_file2 > outputfile2.sai***

where:

- "-n" is the false negative rate you are willing to accept. When given a false negative rate, bwa then select appropriate limits on substitutions. To do this, it needs to know how likely a substitution is, and it simply assumes 2%. So this parameter is nothing more than "the fraction of alignments missing assuming an error rate of 0.02".
- "-t" is the number of thread.
- "-q" is the parameter used to trim read below that Q score according to.

***bwa sampe -a 800 reference_genome.fa outputfile1.sai outputfile2.sai fastq-file1 fastq-file2 > output.sam***

where the parameter

- "a" is the maximum insert size for a read pair to be considered being mapped properly.

### 3.5.3.5 BAM CONVERSION AND SORTING

The resulting SAM file was first converted in the binary format BAM, that contains the same information of the SAM, but reducing storage usage. Because after alignment sequence reads are not sorted in any order, it is requires sorting them on coordinate and

chromosome order of the reference. To this aim we used the java script *SortSam.jar* from Picard Tool (http://picard.sourceforge.net).

***java -Xmx2G -jar PICARD_DIR/SortSam.jar I=output.sam O=sample.bam SO=coordinate VALIDATION_STRINGENCY=SILENT***

where:

- "SO" sort order of output file for coordinate.
- "VALIDATION_STRINGENCY" when setting to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.

#### 3.5.3.6 RAW VARIANT CALL

There is a wide range of available variant callers. Depristo *et al*. at the BROAD Institute have suggested a best practice protocol tailored for use with 1000 genomes, that relies on the Genome Analysis Toolkit (GATK), currently one of the best tools for this purpose (De Pristo *et al*., 2011, McKenna *et al*., 2010).
The first step in our pipeline is the raw variant call that was performed with *Unified Genotyper* tool. It takes the BAM file produced by BWA as input and the reference FASTA file and produce as output a .vcf file containing the raw variant called.

***java -jar GenomeAnalysisTK.jar***
***-T UnifiedGenotyper***
***-I sample.bam***
***-o sample.raw.vcf***
***-R REF_GENOME***
***-glm BOTH***
***-stand_call_conf 1.0***
***-stand_emit_conf 1.0***
***-A AlleleBalance***
***-A DepthOfCoverage***
***-deletions 1.00***

***-nt 4***

***-rf BadCigar***

where:

- "glm" calls SNP, INDEL or in our case BOTH.
- "stand_call_conf" is the call threshold, that means what Phred confidence is required to consider a site confidently called. In the case of raw variant it was set to 1.0.
- "stand_emit_conf" determines what GATK will write as a site, even if the confidence is below the call confidence. In the case of raw variant it was set to 1.0.
- "A" is the annotations to apply to variant call. This annotations are added to the output VCF file. In particular *AlleleBalance* is the fraction of ref bases over ref + alt bases, while *DepthOfCoverage* is Total (unfiltered) depth over all samples.
- "deletion" is the maximum number of reads with deletions spanning a locus for it to be callable.
- "nt" is the number of thread.
- "rf" is a read filter for reads with a bad cigar strings.

#### 3.5.3.7 LOCAL REALIGNMENT AROUND INDEL

The local realignment tool is designed to consume one or more BAM files and to locally realign reads such that the number of mismatching bases is minimized across all the reads. In general, a lot of regions requiring local realignment are due to the presence of an insertion or deletion (indel) in the individual's genome with respect to the reference genome. Such alignment artifacts result in many bases mismatching the reference near the misalignment, which are easily mistaken as SNPs. Moreover, since read mapping algorithms operate on each read independently, it is impossible to place reads on the reference genome such at mismatches are minimized across all reads. Consequently, even when some reads are correctly mapped with indels, reads covering the indel near just the start or end of the read are often incorrectly mapped with respect the true indel, also requiring realignment. Local realignment serves to transform regions with misalignments due to indels into clean reads containing a consensus indel suitable for standard variant discovery approaches.

There are 2 steps to the realignment process:

- Determining (small) suspicious intervals which are likely in need of realignment using RealignerTargetCreator tool.
- Running the realigner over those intervals using IndelRealigner tool.

**Determining Intervals:**

***java -jar GenomeAnalysisTK.jar***
***-T RealignerTargetCreator***
***-R REF_GENOME***
***-o sample.gatk_target.intervals***
***-I sample.bam***
***-rf BadCigar***
***-nt 4***
***--known sample.raw.vcf***

**Local Realignment:**

***java -jar GenomeAnalysisTK.jar***
***-T IndelRealigner***
***-I sample.bam***
***-R REF_GENOME***
***-targetIntervals sample.gatk_target.intervals***
***-o sample.real.bam***

### 3.5.3.8 DUPLICATE REMOVAL USING PICARD TOOL

Examines aligned records in the BAM file to locate duplicate molecules. All records are then written to the output file with the duplicate records flagged.

***java -Xmx2G -jar PICARD_DIR/MarkDuplicates.jar***
***I=sample.real.bam***
***O=sample.real.nodup.bam***
***M=sample.dup_metrics***
***REMOVE_DUPLICATES=true***
***READ_NAME_REGEX=\"[^:]+:[0-9]:([0-9]+):([0-9]+):([0-9]+).*\"***

***VALIDATION_STRINGENCY=SILENT***

where:

- "M" in the name of file in which to write duplication metrics.
- "READ_NAME_REGEX" is a regular expression that can be used to parse read names in the BAM file. Read names are parsed to extract three variables: tile/region, x coordinate and y coordinate. These values are used to estimate the rate of optical duplication in order to give a more accurate estimated library size

The ouput file *sample.real.nodup.bam* was finally indexing using samtool (Li et al., 2009) .

***samtools index sample.real.nodup.bam***.

### 3.5.3.9 RECALIBRATION

The recalibration walker is a two step procedure. The first determines the covariates (such as read group, quality score, machine cycle, and nucleotide context) affecting base quality scores in the BAM file and generates recalibration table based on various this covariates. The latter, walking through the BAM file and rewrite the quality scores.

**CountCovariates:**

***java -jar GenomeAnalysisTK.jar***
***-T CountCovariates***
***-l INFO***
***-R REF_GENOME***
***-I sample.real.nodup.bam***
***-cov ReadGroupCovariate***
***-cov QualityScoreCovariate***
***-cov CycleCovariate***
***-cov DinucCovariate***
***-recalFile OUTDIR/sample.real.nodup.cov_data.csv***
***-dP illumina***
***-nt 4***

where:

- “cov ReadGroupCovariate” is the read group this read is a member of.
- “cov QualityScoreCovariate” is the reported base quality score for this base.
- “cov CycleCovariate” is the cycle of for the considered base.
- “cov DinucCovariate” is the combination of the considered base and the previous one.

This step creates a .csv file which is needed to recalibrate reads.

**TableRecalibration:**

***java -jar GenomeAnalysisTK.jar***
***-T TableRecalibration***
***-l INFO***
***-R REF_GENOME***
***-I sample.real.nodup.bam***
***-recalFile OUTDIR/sample.real.nodup.cov_data.csv***
***--out sample.bam***
***-dP illumina***

Finally samtools was used to index the recalibrated BAM file.

***samtools index sample.bam***

#### 3.5.3.10 ANALYZE COVARIATES

***java -Xmx4g -jar AnalyzeCovariates.jar***
***-recalFile OUTDIR/sample.real.nodup.recal.cov_data.csv***
***-outputDir OUTDIR/recalibrated***
***-ignoreQ 5***

This tool produces a PDF files that graphically show the various metrics and characteristics for BAM file given in input to CounCovariates walker. In order to show that any biases in the reported quality scores have been fixed through recalibration was run again on the bam file produced by TableRecalibration walker. In this way is possible to compare the analysis plots generated by pre-recalibration and post-recalibration .csv files.

**CountCovariates:**

***java -jar GenomeAnalysisTK.jar***
***-T CountCovariates***
***-l INFO***
***-R REF_GENOME***
***-I sample.bam***
***-cov ReadGroupCovariate***
***-cov QualityScoreCovariate***
***-cov CycleCovariate***
***-cov DinucCovariate***
***-recalFile OUTDIR/sample.real.nodup.recal.cov_data.csv***
***-dP illumina***
***-nt 4***

**AnalyzeCovariates:**

***java -Xmx4g -jar AnalyzeCovariates.jar***
***-recalFile OUTDIR/sample.real.nodup.recal.cov_data.csv***
***-outputDir OUTDIR/recalibrated***
***-ignoreQ 5***

### 3.5.3.11 UNIFIED GENOTYPER

Finally UnifiedGenotyper walker was ran again on recalibrated bam files to call SNP and indels.

***java -Xmx4g -jar GenomeAnalysisTK.jar***
***-T UnifiedGenotyper***
***-glm BOTH***
***-R REF_GENOME***
***-I sample.bam***
***-o sample.vcf***
***-stand_call_conf 1.0***
***-stand_emit_conf 1.0***
***-A DepthOfCoverage***

***-A AlleleBalance***

***-deletions 1.00***

***-rf BadCigar***

***-nt 4***

### 3.5.3.12 RECALIBRATION

After obtaining BAM and VCF files for all the six samples, raw variants were filtered using RIKURATOR software (**Figure 15**). First the three parameters Call to 25%, Coverage to 8 and Quality to 40 was set, to filter out all variants that are supported by less than 8 reads, with a fraction of alternative call over reference + alternative call less than 25% and with quality less than 40. Subsequently all variants in dbSNP 135 were filtered out leaving only novel variants. Finally common variants between three, four, five or six samples were searched.

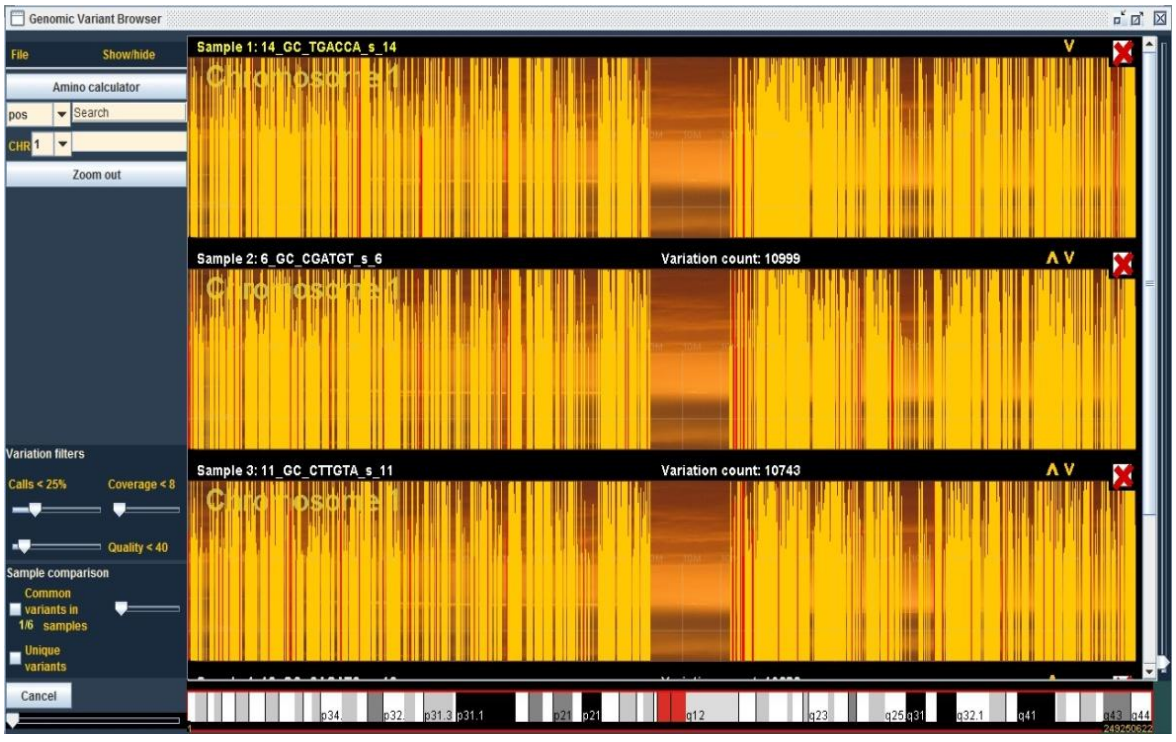



**Figure 15 Screenshot showing the RIKURATOR software**

# 3.6. RESULTS AND DISCUSSION

The results are divided in two sections. The first, presenting results of the pipeline to demonstrate its reliability; the latter, reporting results of the experiment, performed in the Laboratory of Molecular Medicine and Genomics at University of Salerno where I worked, on six patients with clinical features of a rare syndrome called Malignant Migrating Partial Seizures in Infancy (MMPSI).

## 3.6.1 OVERALL STATISTICS OF EXOME SEQUENCING EXPERIMENT

Exome sequencing was performed in six unrelated children showing the clinical features of MMPSI. An average of 8 Gb of sequence was generated per affected individual as paired-end, 72-bp reads. Starting from more than 100 million of total reads, about 80% are uniquely mapped against human genome (**Table 9**).

| | GC_4 | GC_6 | GC_8 | GC_11 | CG_14 | GC_16 |
|---|---|---|---|---|---|---|
| Total reads | 106,007,833 | 125,276,108 | 89,518,042 | 136,240,122 | 122,366,061 | 114,755,522 |
| Uniquely mapped reads (#) | 82,947,391 | 99,368,840 | 63,804,951 | 99,576,132 | 92,136,087 | 86,551,692 |
| Reads in targeted regions (%) | 66.25 | 67.04 | 65.53 | 66.5 | 66.55 | 66.29 |
| Reads in targeted regions +/- 200bp(%) | 84.5 | 84.46 | 84.92 | 84.94 | 84.52 | 84.32 |
| Targeted bases with minimum 10x coverage (%) | 83.49 | 86.04 | 76.42 | 83.01 | 81.51 | 81.33 |
| Targeted bases with minimum 20x coverage (%) | 72.97 | 77.08 | 64.04 | 73.31 | 71.37 | 70.78 |
| Targeted bases with minimum 40x coverage (%) | 57.53 | 62.95 | 46.88 | 59.63 | 57.35 | 56.07 |
| Average coverage (fold) | 71.93 | 87.62 | 52.29 | 83.27 | 76.6 | 71.87 |

**Table 9 Summary statistics for exome sequencing experiment.**

The targeted bases constituted approximately 66% of all bases read. An additional 18% of bases were within 200 base pairs of targeted sequences. About 82% of targeted bases were sufficiently covered (coverage > 8x) to pass the thresholds and used for subsequent analysis.

**Figure 16** shows the comparison of base quality along reads between unmapped and mapped reads. The X-axis is the position in each read, while the Y-axis is the base quality score (Phred quality score). The red and blue lines denote the median and mean of base qualities for each position along with the box plot in yellow. Comparing with unmapped reads, the mapped reads have high base quality, especially for the bases at the end of reads that have very low quality in unmapped reads.

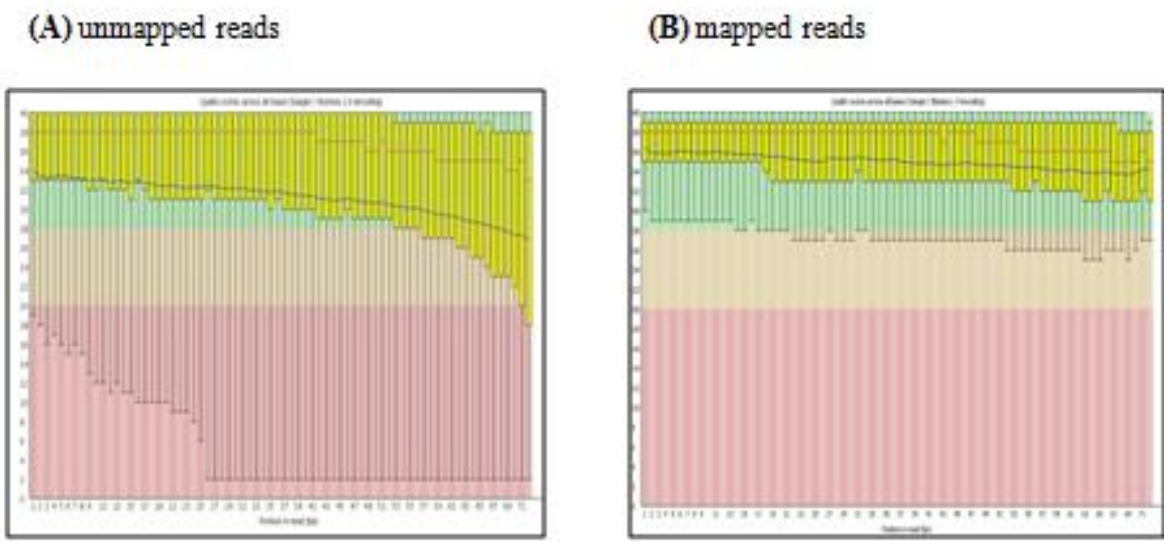



**Figure 16 The box plots of base quality for each position in reads**
The red and blue lines represent the median and mean of Phred score.

## 3.6.2 BASE QUALITY SCORE RECALIBRATION

Base quality score recalibration is performed to avoid the effect of different covariates by adjusting reported base qualities approximate to empirical quality distribution. In the **Figure 17** (**A**) and (**B**), X-axis is the machine cycle (identical to position along reads in the previous figure) whose range is from -72 to 72, -72 to 0 and 0 to 72 denote the positions of two reads in the same read pair, respectively. Y-axis represents the differences between empirical and reported quality score. Before recalibration (**A**) the base quality is overestimated in some positions while underestimated in other ones. After recalibration (**B**), the reported base qualities at all positions are much closer to the empirical qualities, which is reasonable assigned the same weight for SNV calling. The base quality may be underestimated and overestimated respectively on the basis of the result from **Figure 17** (**C**). All the reported base qualities are adjusted approximate to the empirical qualities after recalibration (**Figure**

**17** (**D**)). The base quality may also be affected by the previous base. **Figure 17** (**E**) shows quality distribution of dinucleotide. It is very chaotic and just few of them fit to empirical quality distribution. After base quality recalibration (**F**), the differences between reported and empirical qualities are close to zero. For each covariate, the root mean square error (RMSE) is calculated to evaluate the overall differences between empirical and reported qualities using the formula below:

$$RMSE(\theta_{reported}, \theta_{empirical}) = \sqrt{\frac{\sum_{i=1}^{n} (x_{reported,i} - x_{empirical,i})^2}{n}}$$

n denotes the total number of points in the figure, $x_{reported,i}$ and $x_{empirical,i}$ are the reported and empirical quality for a given point I, respectively. By comparing the value of RMSE before and after base quality score recalibration, is possible to say that the base quality score recalibration can effectively filter out the effect of various covariates and make reported base quality much closer to the empirical base quality which is necessary and crucial for SNV calling.

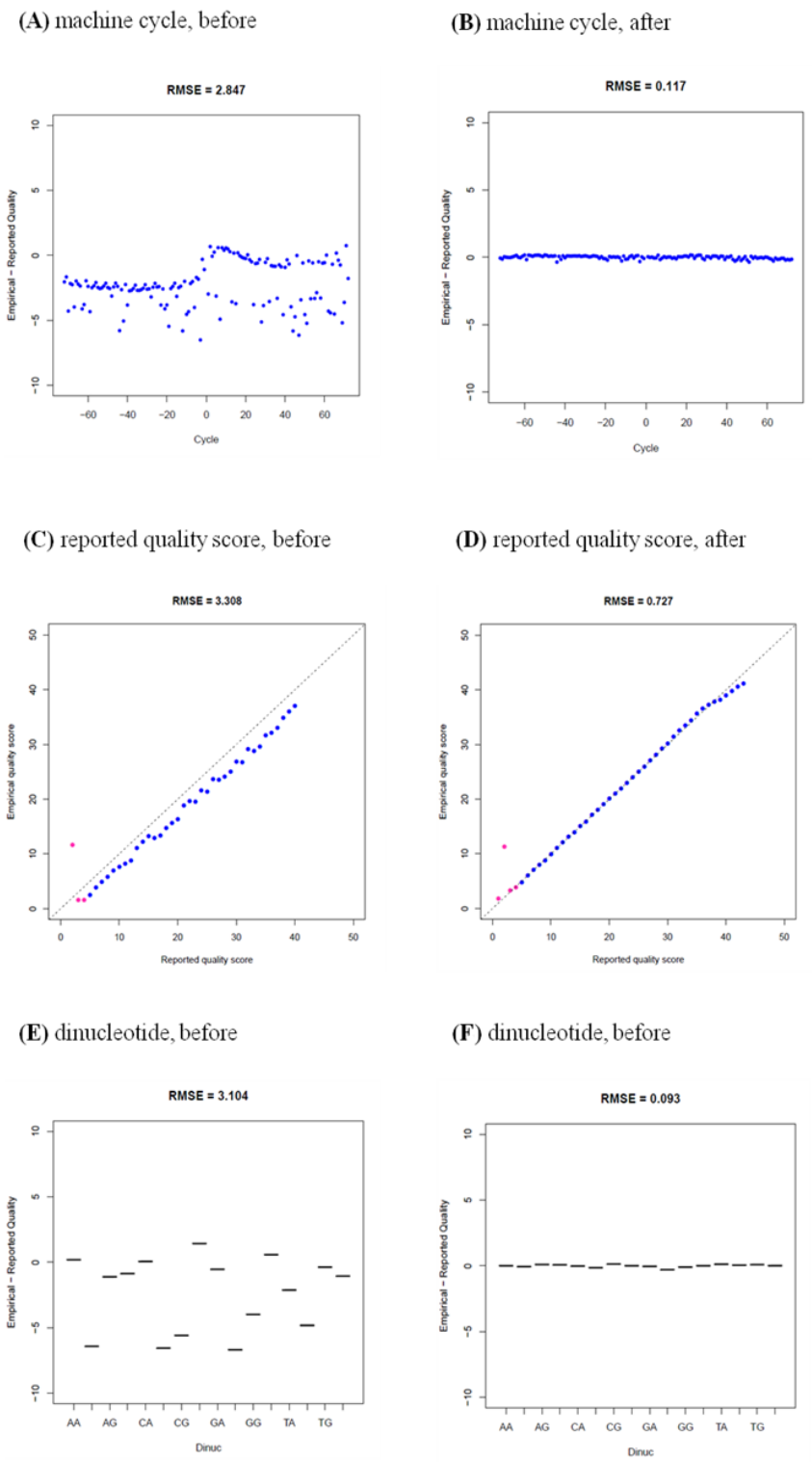



**Figure 17 Quality score recalibration**
The effect of base quality score recalibration on three covariates: machine cycle, reported base quality and dinucleotide for one of the sequenced sample.

### 3.6.3 VARIANTS DETECTION

The last step of the analysis is the identification of variants. Variants are detected based on mapped, local realigned and base quality score recalibrated reads to reduce the number of false positive. SNP and short indel identification were performed with Unified Genotyper tool provided by Genome Analysis Toolkit (see methods 3.4.3.11). Detailed information about each SNP or INDEL call are given in the generated VCF file which include total number of reads covering the site, ratio of reference supporting reads to total number of references (referred to as allele balance), number of identified alleles (referred to as allele count), allele frequencies for each allele, number of reads supporting the site with mapping quality zero, dbSNP id if variant is known, and strand bias.

### 3.6.4 EXPERIMENTAL RESULTS

The developed pipeline was used to analyze data from whole exome sequencing on six patients showing clinical features of MMPSI. Enrichment was performed using the SureSelect Target Enrichment 50Mb protocol followed by 72 base pair massively parallel sequencing using the Genome Analyzer IIx sequencer. The number of raw variants obtained was more than 45,000 in all samples except sample 8 where the number is about 35,000. To avoid false negative results, SNP and INDEL with coverage less than 8, quality less than 40 and call (alternative/reference+alternative) less than 25% were filtered out. Then dbSNP 135 and a European control were used to obtain only the novel variations that were used for the further analysis (**Table 10**).

| | GC_4 | GC_6 | GC_8 | GC_11 | CG_14 | GC_16 |
|---|---|---|---|---|---|---|
| **SNP/short INDEL identified** | 48,443 | 52,957 | 35,184 | 47,795 | 47,386 | 47,887 |
| **After filtering coverage>8, quality>40, call>25%** | 10,592 | 10,999 | 8,769 | 10,743 | 10,671 | 10,683 |
| **After filtering known variations** | 891 | 930 | 761 | 1,067 | 1,387 | 1,422 |
| **After filtering European variations** | 740 | 758 | 625 | 916 | 1,247 | 1,288 |

**Table 10 Summary of variations identified per sample, before and filtering**

In order to identify variants that are clinically significant, common variants between at least 3 samples were studied and different type of variations, non synonymous, framshift coding, stop gained and splice cite, were taken in consideration (**Table 11**).

| | 6/6 | 5/6 | 4/6 | 3/6 |
|---|---|---|---|---|
| **RARE VARIANTS IDENTIFIED** | 125 | 137 | 194 | 357 |
| **NON SYNONYMOUS CODING** | 16 | 19 | 37 | 36 |
| **FRAMESHIFT CODING** | 2 | 2 | 4 | 9 |
| **STOP GAINED** | | 2 | | |
| **SPLICE SITE** | 5 | 7 | 8 | 31 |

**Table 11 Summary of common variants**
Common variants identified in common between 6, 5, 4 and 3 samples associated with variation effect.

### 3.6.4.1 VARIATIONS IN GENES ASSOCIATED WITH MMPSI

After studied novel variations, the attention was focused on mutations, in all the six samples, regarding genes known from literature to be associated with MMPSI (**Table 4**). The first gene evaluated was PLCB1. A homozygous deletion (486 kb) in this gene was discovered by Poduri et al. in one patient with MMPSI. Considering that in this study large structural variation were not investigated, all the other possible variations in this gene were considered before and after filtering for coverage greater than 8. As shown in **Table 12** mutations were found in intronic, splicing site and UTR regions, but no non-synonymous variations were identified.

| | | | COVERAGE ≥8 |
|---|---|---|---|
| | **VARIANT TYPE** | **n°** | **n°** |
| **PLCB1 NOVEL VARIATIONS** | INTRONIC | 159 | 8 |
| | SPLICING SITE | 1 | 1 |
| | UTR | 5 | 2 |
| | TOTAL | 165 | 11 |

**Table 12 PLCB1 mutations**

After that, mutations in SCN1A were considered. Carranza Rojo et al., identified pathogenic genetic abnormalities in 2/15 (13%) patients, including a missense mutation and a deletion that encompasses the entire SCN1A gene as well as other sodium channel subunit genes. Freilich et al., discovered a novel missense mutation in a patient who clinically fulfilled the criteria for malignant migrating partial seizures of infancy. **Table 13** shows 2 novel non synonymous mutations identified in our samples, the first in one patient and the latter in two patients (**Table 14**). Although both variations resulted in a high coverage, the ratio between alternative/alternative+reference was to the limit of call threshold, setted to 25%, in all of them, so the probability to have a false positive result was considered. To validate these variants, Sanger sequencing approach was used (**Figure 18**). As shown in the figure both the variations were not confirmed.

| | | | COVERAGE ≥8 |
|---|---|---|---|
| | **VARIANT TYPE** | **n°** | **n°** |
| **SCN1A NOVEL VARIATIONS** | INTRONIC | 34 | 5 |
| | NON SYNONYMOUS | 2 | 2 |
| | SPLICING SITE | 2 | 2 |
| | TOTAL | 38 | 9 |

**Table 13. SCN1A mutations**

| | | | | ALTERNATIVE/REFERENCE+ALTERNATIVE | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **CHR** | **Position** | **BaseChange** | **AminoChange** | **GC4** | **GC6** | **GC8** | **GC11** | **GC14** | **GC16** |
| **2** | **166,847,950** | **G->T** | **Asn1945Lys** | | | | | **28/99** | |
| **2** | **166,847,956** | **A->T** | **Asn1969Lys** | | | | | **28/97** | **22/105** |

**Table 14 Two novel non-synonymous mutations in SCN1A**
In red is the ratio between the number of read covering the alternative base over the total number of read

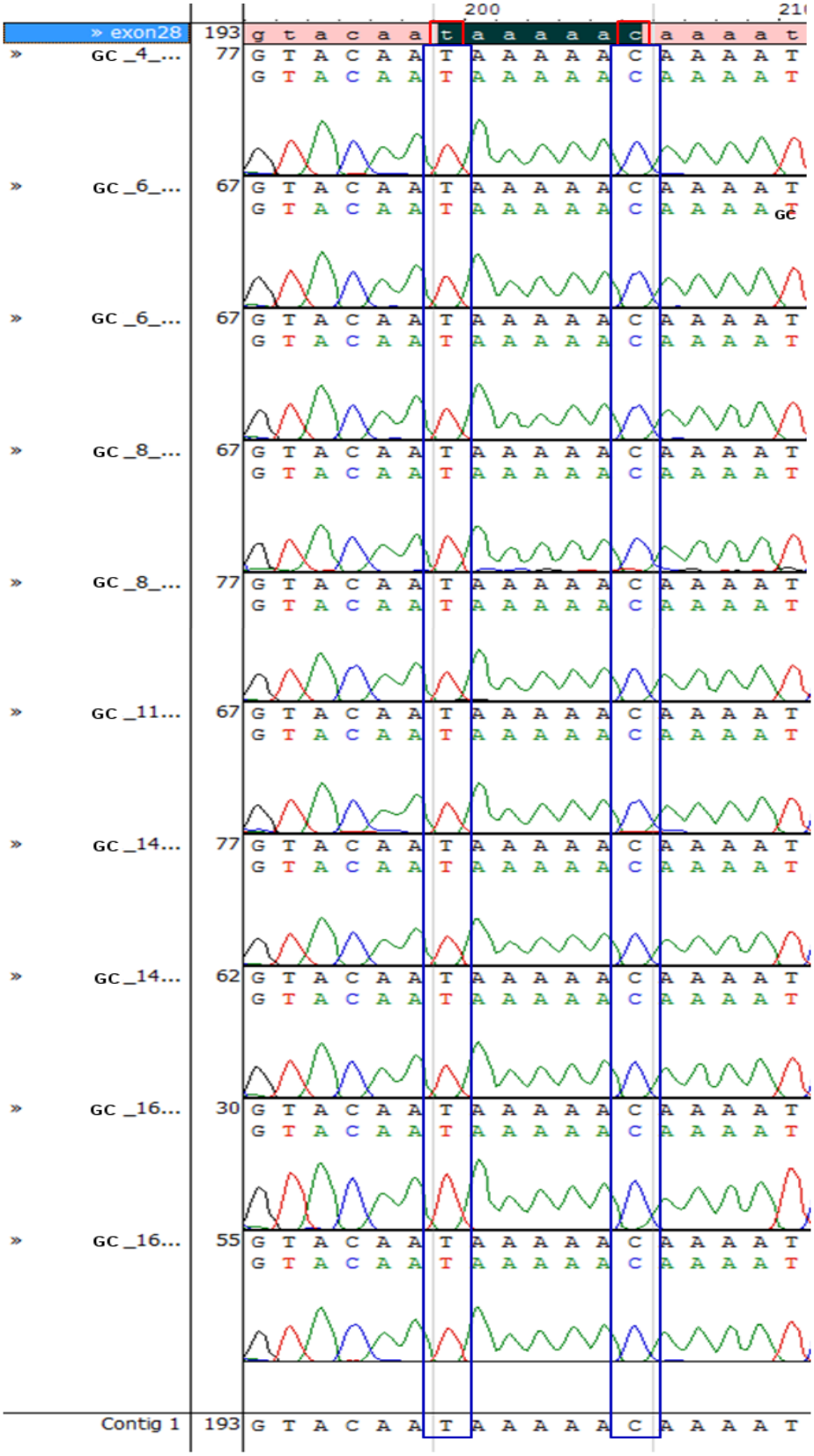



**Figure 18 Sanger Sequencing of SCN1A mutations**
Sanger sequencing validation of the two novel non-synonymous mutations identified in SCN1A (exon 28, chr2:166,847,950 and chr2:166,847,956 )

Finally KCNT1 gene was investigated. While the genes mentioned above had a good coverage, KCNT1 presented several targeted regions with very low coverage (**Figure 19**), leading to needs to validate all the variations found in them.

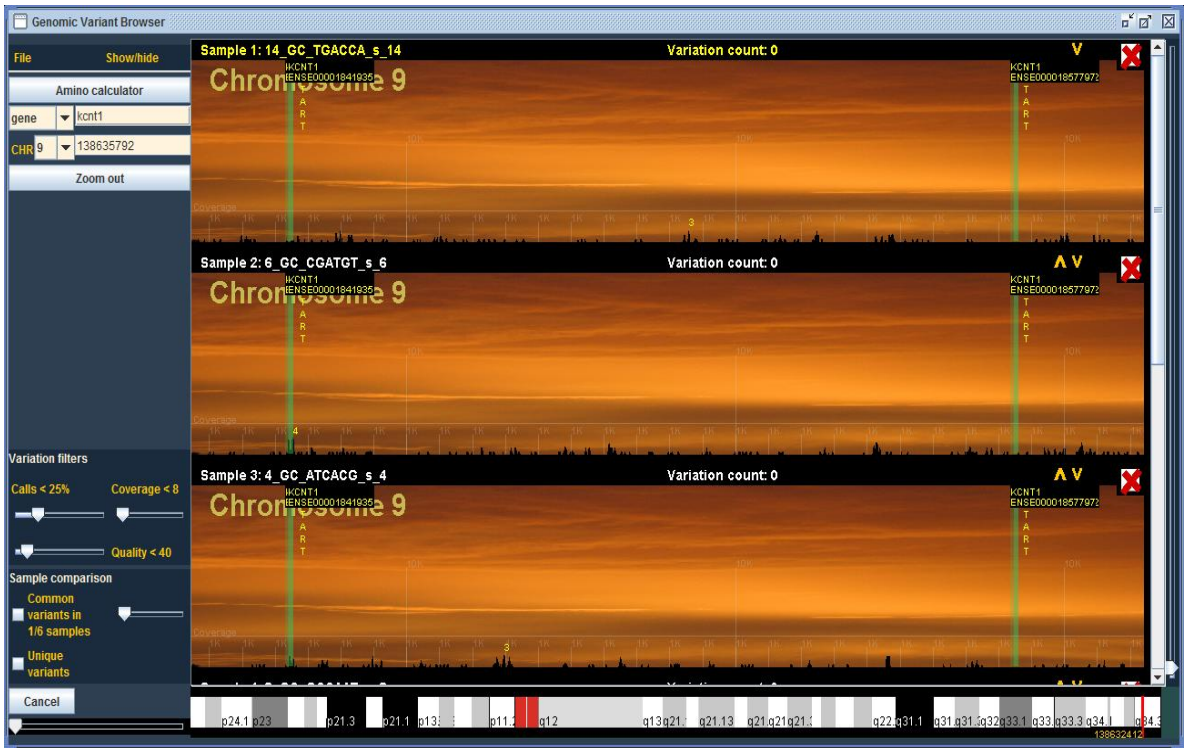



**Figure 19 KCNT1 coverage**
Screenshot of Rikurator in the KCNT1 targeted regions showing the low coverage

Using exome sequencing, Barcia et al. identified de novo KCNT1 gain-of-function mutations in 6 of 12 patients with MMPSI.

| | | | COVERAGE ≥8 |
|---|---|---|---|
| | **VARIANT TYPE** | **n°** | **n°** |
| **KCNT1 NOVEL VARIATIONS** | INTRONIC | 38 | 2 |
| | SYNONYMOUS/UTR | 1 | - |
| | UTR | 1 | - |
| | NON SYNONYMOUS/UTR | 3 | **1** |
| | TOTAL | 43 | - |

**Table 15 KCNT1 mutations**

This gene seems to be a major disease-associated gene for MMPSI; infact, the mutations described by Barcia et al., appear substantially more severe than those reported for other potassium channels.

For this reason all the variations identified were taken in considerations (**Table 15**). Three non-synonimous variants were identified in our sample without filtering for coverage, while after that only one remains in the list of possible candidates (**Table 16**). Considering the fact that one of the limitations of the exome sequencing is the possibility to have low coverage in some targeted regions, all the variants were validate with Sanger sequencing (**Figure 20, 21**).

| | | | | ALTERNATIVE/REFERENCE+ALTERNATIVE | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Chromosome** | **Position** | **BaseChange** | **AminoChange** | **GC4** | **GC6** | **GC8** | **GC11** | **GC14** | **GC16** |
| **9** | **138,660,705** | **G->T** | **Val459Leu/UTR** | | | | | | 1/5 |
| **9** | **138,662,780** | **T->G** | **Leu597Arg/UTR** | | | | 1/1 | | 1/4 |
| **9** | **138,667,199** | **T->C** | **Ser744Pro/UTR** | | | | | | **5/11** |

**Table 16 Three novel non-synonymous mutations in KCNT1**
In red is the ratio between the number of read covering the alternative base over the total number of read.

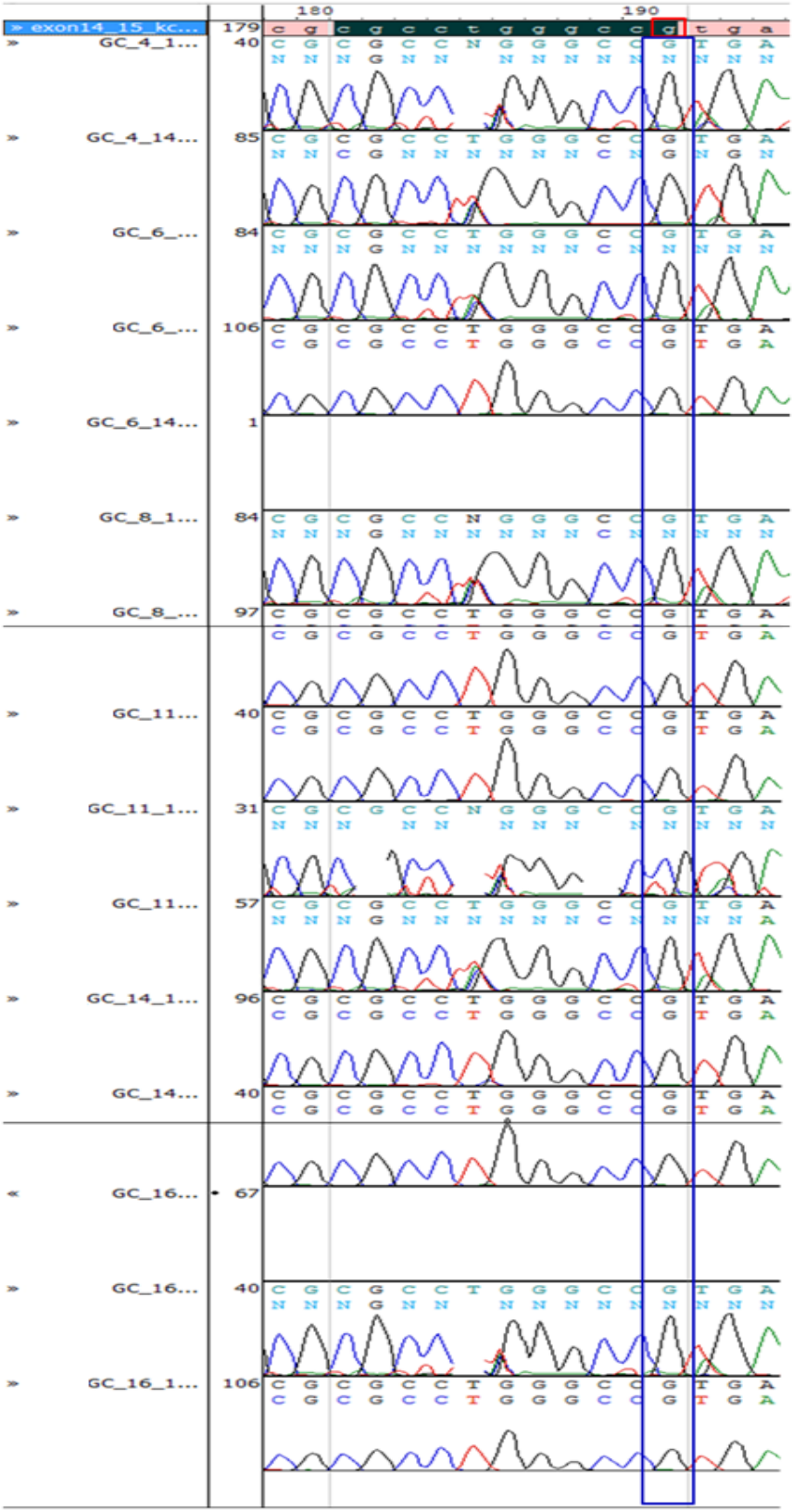



**Figure 20 Sanger sequencing in KCNT1 gene**
Sanger sequencing validation of the first novel non-synonymous mutations identified in KCNT1 (exon 15, chr9: 138,660,705)

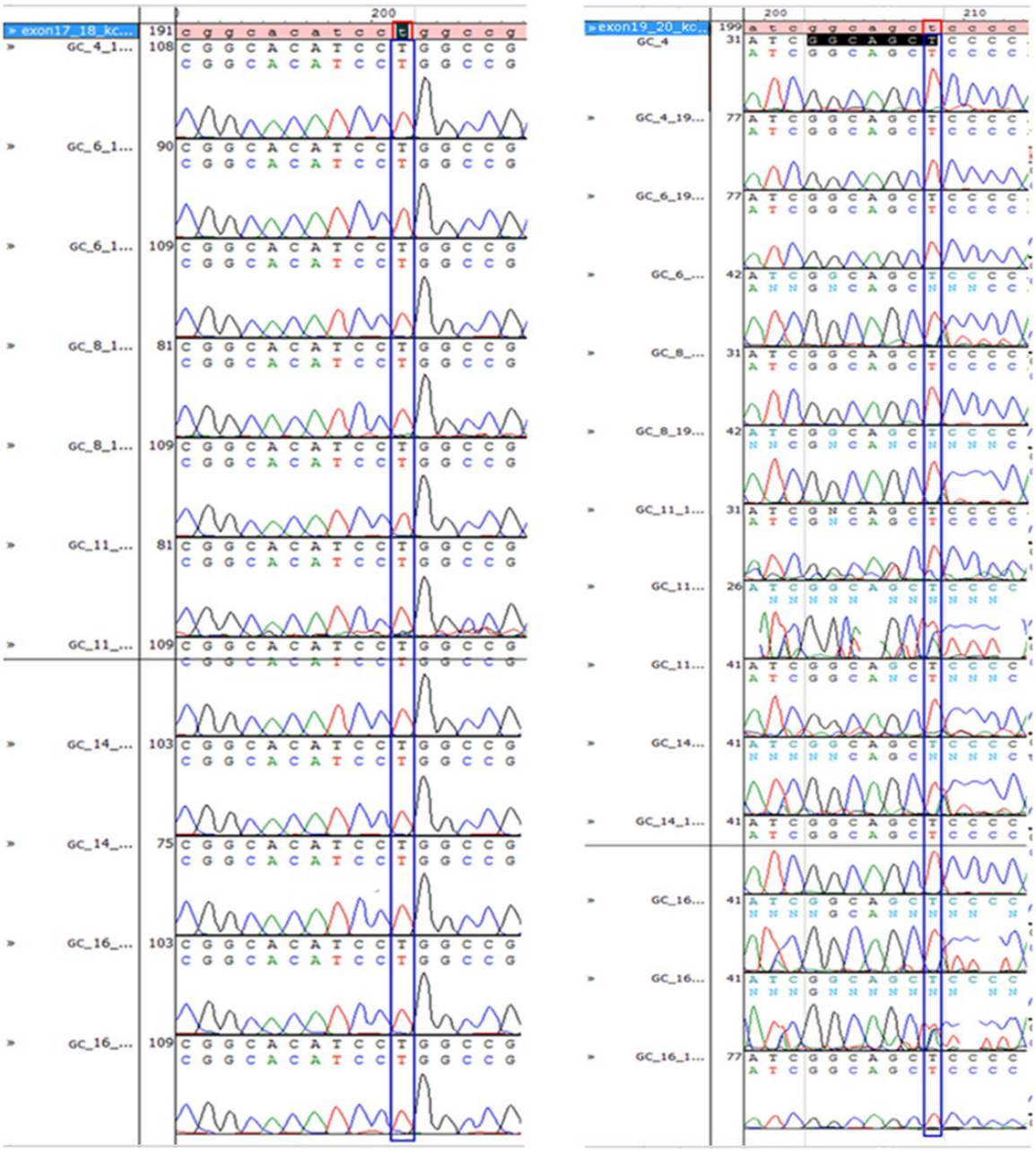


**Figure 21 Sanger sequencing in KCNT1 gene**
Sanger sequencing validation of the other two novel non-synonymous mutations identified in KCNT1 (exon 17, chr9: 138,662,780 and exon 20 chr9: 138,667,199)

## 3.7. CONCLUSIONS

The development of next generation sequencing technologies has provide an unprecedented opportunity to investigate and identify potential disease causing mutations for rare disease. It is a big challenge to accurately identify candidate SNVs and select the subset of functionally important variants from ten thousands of single nucleotide variations. For short reads alignment, sequencing errors are no longer a big challenge for BWT based algorithm. The preprocessing procedure for SNV calling, including PCR removal, indel realignment and base quality score recalibration can efficiently remove read pairs as PCR duplication, adjust false positive SNV calls due to indel misalignment and correct the effects of covariates. Base quality score recalibration can make reported base quality more approximate to the empirical quality by adjusting covariates such as machine cycle, dinucleotide etc. All these procedure can systematically adjust the alignment result from different aspects and make SNV calls more reasonable.

Starting from six unrelated patients, with clinical features of MMPSI, exome sequencing was performed to look for disease causing mutation(s). The pipeline developed was used to analyze data from sequencer. After filtering for known variants, because MMPSI is likely to be genetically heterogeneous and, therefore, not all affected individuals will carry mutations in the same gene, common variants in a subset of patients were investigate. First of all, the attention was focused on novel variants in at least three patients, then, mutations known from literature to be associated with MMSI were taken in considerations. In particular, PLCB1, SCN1A and KCNT1 were considered. No variants were identified in PLCB1, while in SCN1A and KCNT1, 2 and 3 mutations were identified before filtering for coverage respectively. With coverage greater than 8, the mutations in SCN1A were still present in the list, but the percentage of alternative/reference + alternative call was very close to the limit of threshold (25%), leading to the consideration that variants identified can be false positive results. After validation using Sanger sequencing, variations were discarded, confirming our observation.

As with any technology, also exome sequencing has limitations, one of the main is the low coverage in some targeted regions, that can complicate the interpretation of data. This is the case of the last gene considered, KCNT1 that had very low coverage in some targeted regions. Because of its interesting role in MMPSI, all the mutations identified, before and after filetering for coverage, were validated. Also in this case, mutations were discarded.

The last few years have clearly shown how massively parallel sequencing can accelerate the pace of disease gene discovery and revolutionize the study of the genetic bases on several disorders. Despite these numerous successes, there is still much that is unknown about the variants found in the genome. Furthermore technical limitations, such as the low coverage in some targeted regions, lead to false positive results, making difficult to interpret data and need to re-sequence that regions.

# REFERENCES

Ambros V: The functions of animal microRNAs. Nature 2004, 431(7006):350-355.

Anders S, Huber W: Differential expression analysis for sequence count data. Genome Biol 2010, 11(10):R106.

Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 1987 Jun 11;15(11):4593-602

Barcia G, Fleming MR, Deligniere A, Gazula VR, Brown MR, Langouet M, Chen H, Kronengold J, Abhyankar A, Cilio R, Nitschke P, Kaminska A, Boddaert N, Casanova JL, Desguerre I, Munnich A, Dulac O, Kaczmarek LK, Colleaux L, Nabbout R. De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. Nat Genet. 2012 Nov; 44(11):1255-9.

Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 2004, 116(2):281-297.

Bedoyan JK, Kumar RA, Sudi J, Silverstein F, Ackley T, Iyler RK, Christian SL, Martin DM. Duplication 16p11.2 in a child with infantile seizure disorder. Am J Med Genet. 2010 June; A 152A:1567-1574.

Betel D, Wilson M, Gabow A, Marks DS, Sander C: The microRNA.org resource: targets and expression. Nucleic Acids Res 2008, 36(Database issue):D149-153.

Betel D, Koppal A, Agius P, Sander C, Leslie C: Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010, 11(8):R90.

Callinan PA, Feinberg AP. The emerging science of epigenomics. Hum Mol Genet. 2006 Apr 15

Carranza Rojo D, Hamiwka L, McMahon JM, Dibbens LM, Arsov T, Suls A, Stödberg T, Kelley K, Wirrell E, Appleton B, Mackay M, Freeman JL, Yendle SC, Berkovic SF, Bienvenu T, De Jonghe P, Thorburn DR, Mulley JC, Merfford HC, Scheffer IE. De novo SCN1A mutations in migrating partial seizures of infancy. Neurology. 2011 July; 77:380-383.

Cicatiello L, Mutarelli M, Grober OM, Paris O, Ferraro L, Ravo M, Tarallo R, Luo S, Schroth GP et al: Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. Am J Pathol 2010, 176(5):2113-30.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010 Apr;38(6):1767-71.

Coppola G, Plouin P, Chiron C, Robain O, Dulac O. Migrating partial seizures in infancy: a malignant disorder with developmental arrest. Epilepsia. 1995 Oct; 36(10):1017-24.

Coppola G, Veggiotti P, Del Giudice EM, Bellini G, Longaretti F, Taglialatela M, Pascotto A. Mutational scanning of potassium, sodium and chloride ion channels in malignant migrating partial seizures in infancy. Brain Dev. 2006 Mar;28(2):76-9. Epub 2005 Sep 15.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156-8.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8.

Dulac O (2005). Malignant migrating partial seizures in infancy. In Roger J, Bureau M, Dravet Ch, Genton P, Tassinari CA, Wolf P. (Eds) Epileptic syndromes in infancy,

childhood and adolescence, 4th edn. John Libbey Eurotext Ltd, Montrouge, France, pp. 73–76.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006 Dec;38(12):1378-85. Epub 2006 Oct 29.

Esteller M. The necessity of a human epigenome project. Carcinogenesis. 2006 Jun;27(6):1121-5. Epub 2006 May 13.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998 Mar;8(3):186-94.

Ferraro L, Ravo M, Nassa G, Tarallo R, De Filippo MR, Giurato G, Cirillo F, Stellato C, Silvestro S, Cantarella C et al: Effects of estrogen on microRNA expression in hormone-responsive breast cancer cells. Horm Cancer 2011, 2(5):610-621

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM. Ensembl 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D800-6.

Flynt AS, Lai EC: Biological principles of microRNA-mediated regulation: shared themes amid diversity. Nature Reviews Genetics 2008, 9(11):831-842.

Freilich ER, Jones JM, Gaillard WD, Conry JA, Tsuchida TN, Reyes C, Dib-Hajj S, Waxman SG, Meisler MH, Peral PL. Novel SCN1A mutation in a proband with malignant migrating partial seizures of infancy. Arch Neurol. 2011 May; 68:665-671.

Friedman RC, Farh KK, Burge CB, Bartel DP: Most mammalian mRNAs are conserved targets of microRNAs. Genome Res 2009, 19(1):92-105.

Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP: Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol 2011, 18(10):1139-1146.

Giurato G, De Filippo M.R., Rinaldi A., Hashim A., Nassa G., Ravo M., Rizzo F., Tarallo R. and Weisz A. iMir: An Integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by miRNA-Seq. BMC Bioinformatics, submitted.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 2006, 34(Database issue):D140-144.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: miRBase: tools for microRNA genomics. Nucleic Acids Res 2008, 36(Database issue):D154-158.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP: MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell 2007, 27(1):91-105.

Grober OM, Mutarelli M, Giurato G, Ravo M, Cicatiello L, De Filippo MR, Ferraro L, Nassa G, Papa MF, Paris O, Tarallo R, Luo S, Schroth GP, Benes V, Weisz A. Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. BMC Genomics. 2011 Jan 14;12:36.

Gross-Tsur V, Ben-Zeev B, Shalev RS. Malignant migrating partial seizures in infancy. Pediatr Neurol. 2004 Oct;31(4):287-90.

Gupta A, Gartner JJ, Sethupathy P, Hatzigeorgiou AG, Fraser NW: Anti-apoptotic function of a microRNA encoded by the HSV-1 latency-associated transcript. Nature 2006, 442(7098):82-85.

Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM: miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic Acids Res 2011, 39(Web Server issue):W132-138.

He L, Hannon GJ: MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 2004, 5(7):522-531.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 August 15; 25(16): 2078–2079.

Hmaimess G, Kadhim H, Nassogne MC, Bonnier C, van Rijckevorsel K. Levetiracetam in a neonate with malignant migrating partial seizures. Pediatr Neurol. 2006 Jan;34(1):55-9.

http://picard.sourceforge.net

Huang PJ, Liu YC, Lee CC, Lin WC, Gan RR, Lyu PC, Tang P: DSAP: deep-sequencing small RNA analysis pipeline. Nucleic Acids Res 2010, 38(Web Server issue):W385-391.

Huang Q, Gumireddy K, Schrier M, le Sage C, Nagel R, Nair S, Egan DA, Li A, Huang G, Klein-Szanto AJ et al: The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. Nat Cell Biol 2008, 10(2):202-210.

Hutvagner G, McLachlan J, Pasquinelli AE, Balint E, Tuschl T, Zamore PD: A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. Science 2001, 293(5531):834-838.

Isakov O, Ronen R, Kovarsky J, Gabay A, Gan I, Modai S, Shomron N. Novel insight into the non-coding repertoire through deep sequencing analysis. Nucleic Acids Res. 2012 Jun;40(11):e86. doi: 10.1093/nar/gks228. Epub 2012 Mar 9.

Jasny BR, Roberts L. Are We There Yet?, Science. 2003 Oct 24: 587.

Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P: Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science 2005, 309(5740):1577-1581.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009 Sep 1;25(17):2283-5.

Kozomara A, Griffiths-Jones S: miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011, 39(Database issue):D152-157.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: Identification of novel genes coding for small expressed RNAs. Science 2001, 294(5543):853-858.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C,Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C,Stange-Thomann N, Stojanovic N, Subramanian Aet al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860-921. Erratum in: Nature 2001 Aug 2;412(6846):565. Nature 2001 Jun 7;411(6838):720.

Lau NC, Lim LP, Weinstein EG, Bartel DP: An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 2001, 294(5543):858-862.

Ledergerber C, Dessimoz C. Base-calling for next-generation sequencing platforms. Brief Bioinform. 2011 Sep;12(5):489-97.

Lee RC, Ambros V: An extensive class of small RNAs in Caenorhabditis elegans. Science 2001, 294(5543):862-864.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S et al: The nuclear RNase III Drosha initiates microRNA processing. Nature 2003, 425(6956):415-419.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN: MicroRNA genes are transcribed by RNA polymerase II. EMBO J 2004, 23(20):4051-4060.

Lewis BP, Burge CB, Bartel DP: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005, 120(1):15-20.

Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009; 25:1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U: Nuclear export of microRNA precursors. Science 2004, 303(5654):95-98.

Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011 Sep;48(9):580-9.

Marsh E, Melamed SE, Barron T, Clancy RR. Migrating partial seizures in infancy: expanding the phenotype of a rare seizure syndrome. Epilepsia. 2005 Apr;46(4):568-72

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303.

Menges F, Narzisi G, Mishra B. TotalReCaller: improved accuracy and performance via integrated alignment and base-calling. Bioinformatics. 2011 Sep 1;27(17):2330-7.

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M et al: Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 2008, 18(4):610-621.

Nilsen TW: Mechanisms of microRNA-mediated gene regulation in animal cells. Trends Genet 2007, 23(5):243-249.

Paris O, Ferraro L, Grober OM, Ravo M, De Filippo MR, Giurato G, Nassa G,Tarallo R, Cantarella C, Rizzo F, Di Benedetto A, Mottolese M, Benes V, Ambrosino C, Nola E, Weisz A. Direct regulation of microRNA biogenesis and expression by estrogen receptor beta in hormone-responsive breast cancer. Oncogene. 2012 Sep 20;31(38):4196-206.

Poduri A, Chopra SS, Neilan EG, Elhosary PC, Kurian MA, Meyer E, Barry BJ, Khwaja OS, Salih MA, Stödberg T, Scheffer IE, Maher ER, Sahin M, Wu BL, Berry GT, Walsh CA, Picker J, Kothare SV. Homozygous PLCB1 deletion associated with malignant migrating partial seizures in infancy. Epilepsia. 2012 Aug; 53(8):e146-50.

Schwab M, Alitalo K, Klempnauer KH, Varmus HE, Bishop JM, Gilbert F, Brodeur G, Goldstein M, Trent J. Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature. 1983 Sep 15-21;305(5931):245-8.

Silber J, Lim DA, Petritsch C, Persson AI, Maunakea AK, Yu M, Vandenberg SR, Ginzinger DG, James CD, Costello JF et al: miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. BMC Medicine 2008, 6:14.

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. Fluorescence detection in automated DNA sequence analysis. Nature. 1986 Jun 12-18;321(6071):674-9.

Tay YM, Tam WL, Ang YS, Gaughwin PM, Yang H, Wang W, Liu R, George J, Ng HH, Perera RJ et al: MicroRNA-134 modulates the differentiation of mouse embryonic stem cells, where it causes post-transcriptional attenuation of Nanog and LRH1. Stem Cells 2008, 26(1):17-29.

Tili E, Michaille JJ, Cimino A, Costinean S, Dumitru CD, Adair B, Fabbri M, Alder H, Liu CG, Calin GA et al: Modulation of miR-155 and miR-125b levels following lipopolysaccharide/TNF-alpha stimulation and their possible roles in regulating the response to endotoxin shock. J Immunol 2007, 179(8):5082-5089.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K *et al*. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. Erratum in: Science 2001 Jun 5;292(5523):1838.

Wang WC, Lin FM, Chang WC, Lin KY, Huang HD, Lin NS: miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. BMC bioinformatics 2009, 10:328.

Wilmshurst JM, Appleton DB, Grattan-Smith PJ. Migrating partial seizures in infancy: two new cases. J Child Neurol. 2000 Nov;15(11):717-22.

Zamore PD, Haley B: Ribo-gnome: the big world of small RNAs. Science 2005, 309(5740):1519-1524.

Zamponi N, Rychlicki F, Corpaci L, Cesaroni E, Trignani R. Vagus nerve stimulation (VNS) is effective in treating catastrophic 1 epilepsy in very young children. Neurosurg Rev. 2008 Jul;31(3):291-7.

Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, Sun Z, Wu J: mirTools: microRNA profiling and discovery based on high-throughput sequencing. Nucleic Acids Res 2010, 38(Web Server issue):W392-397.

# LIST OF PUBLICATIONS AND ABSTRACTS

- Giurato G.*, **De Filippo M.R.*,** Rinaldi A., Hashim A., Nassa G., Ravo M., Rizzo F., Tarallo R. and Weisz A. iMir: An Integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by miRNA-Seq. BMC Bioinformatics, submitted.

- Cirillo F., Nassa G., Tarallo R., Stellato C., **De Filippo M.R.,** Ambrosino C., Baumann M., Nyman T.A., Weisz A. Molecular Mechanisms of Selective Estrogen Receptor Modulator Activity in Human Breast Cancer Cells: Identification of Novel Nuclear Cofactors of Antiestrogen-ERα Complexes by Interaction Proteomics. J Proteome Res. 2013 Jan 4; 12(1):421-31.

- Ferraro L., Ravo M., Nassa G., Tarallo R., **De Filippo M.R.**, Giurato G., Cirillo F., Stellato C., Silvestro S., Cantarella C., Rizzo F., Cimino D., Friard O., Biglia N., De Bortoli M., Cicatiello L., Nola E., Weisz A. Effects of estrogen on microRNA expression in hormone-responsive breast cancer cells. Horm Cancer. 2012 Jun;3(3):65-78.

- Paris O., Ferraro L., Grober O.M.V., Ravo M., **De Filippo M.R.**, Giurato G., Nassa G., Tarallo R., Cantarella C., Rizzo F., Di Benedetto A., Mottolese M., Benes V., Ambrosino C., Nola E., Weisz A. Direct Regulation of microRNA Biogenesis and Expression by Estrogen Receptor Beta in Hormone-responsive Breast Cancer. Oncogene. 2012 Sep 20;31(38):4196-206.

- Grober O.M.V., Mutarelli M., Giurato G., Ravo M., Cicatiello L., **De Filippo M.R.**, Ferraro L., Nassa G., Papa M.F., Paris O., Tarallo R., Luo S., Schroth G.P., Benes V., Weisz A. Global analysis of ERβ binding to breast cancer cell genome reveals extensive interplay with ERα for target gene regulation. BMC Genomics. 2011 Jan 14;12:36.

- **De Filippo M.R.**, Giurato G., Rizzo F., Greco D., Katainen R., Aaltonen L., Coppola G., Weisz A. Development of an accurate pipeline for exome seqeuncing data analysis. XV Congresso Nazionale SIGU, 21-24 Novembre 2012, Sorrento, Italia.

- **De Filippo M.R.**, Giurato G., Rizzo F., Ravo M., Stellato C., Marchese G., Weisz A., An accurate pipeline for analysis of exome sequencing data, GerIta, "Next generation

sequencing-Application cases and bioinformatics development", 17-19 luglio 2012, Napoli, Italy

- **De Filippo M.R.**, Giurato G., Cantarella C., Rizzo F., Cirillo F., Weisz A., Development of pipeline for Exome Sequencing Data Analysis, BITS 2012, Nine Annual Meeting of Bioinformatics Italian Society,2-4 maggio 2012, Catania, Italy.

- **De Filippo M.R.**, Giurato G., Tarallo R., Ravo M., Rizzo F., Cantarella C., Nassa G., Nola E., Weisz A. Computational approaches for genome-wide mRNA and miRNA expression profiling in human breast cancer cell lines expressiong (ERβ+) or lacking (ERβ-) estrogen receptor beta by microarray hybridization and massively parallel sequencing (miRNA-Seq). BBCC2011, Sixth Annual Meeting Bioinformatica e Biologia Computazionale in Campania, 4 novembre 2011, ISA-CNR, Avellino, Italy. (comunicazione orale).

***** G.G. and MR.DF. equally contributed to this work and should therefore be both considered as first

**Author's covering letter for initial submission**

**Title:** iMir: An Integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by miRNA-Seq

**Authors:**


Giorgio Giurato (ggiurato@unisa.it)
Maria Rosaria De Filippo (mdefilippo@unisa.it)
Antonio Rinaldi (arinaldi@unisa.it)
Adnan Hashim (ahashim@unisa.it)
Giovanni Nassa (gnassa@unisa.it)
Maria Ravo (mravo@unisa.it)
Francesca Rizzo (frizzo@unisa.it)
Roberta Tarallo (rtarallo@unisa.it)
Alessandro Weisz (aweisz@unisa.it)



**Version:** 1 **Date:** 26 March 2013


**Comments:**

Dear Editor,

I submit to your attention the manuscript entitled: "iMIR: AN INTEGRATED PIPELINE FOR HIGH-THROUGHPUT ANALYSIS OF SMALL NON-CODING RNA DATA OBTAINED BY MIRNA-SEQ" by Giorgio Giurato et al. for publication in BMC Bioinformatics.

In this software article we describe an original modular pipeline for comprehensive analysis of small non coding RNA sequencing (miRNA-Seq) data generated by the most widely used next generation sequencing platforms. This tool has proven to be more efficient, time-saving and easy to use and update than currently available methods, in particular when a large number of datasets need to be analyzed for the presence of (known and novel) miRNAs, piRNAs and the other sncRNAs.

For the reasons stated above, we believe that this article might be of interest and useful to a wide audience, such as that represented by the readers of your Journal.

Sincerely Yours,

Alessandro Weisz

# iMir: An Integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by miRNA-Seq

Giorgio Giurato[§1], Maria Rosaria De Filippo[§2], Antonio Rinaldi[1], Adnan Hashim[1], Giovanni Nassa[1], Maria Ravo[1], Francesca Rizzo[1], Roberta Tarallo[1] and Alessandro Weisz[1,3]

[1]Laboratory of Molecular Medicine and Genomics, Department of Medicine and Surgery, University of Salerno, Baronissi, Salerno, Italy.
[2]Fondazione IRCCS SDN, Napoli, Italy.
[3] Division of Molecular Pathology and Medical Genomics, "SS. Giovanni di Dio e Ruggi d'Aragona – Schola Medica Salernitana" University of Salerno Hospital, Salerno, Italy.

[§]G.G. and MR.DF. equally contributed to this work and should therefore be both considered as first author

Correspondence to:
Prof. Alessandro Weisz
Laboratory of Molecular Medicine and Genomics,
University of Salerno,
via Allende, 1
84081 Baronissi (SA), Italy
aweisz@unisa.it
Tel. (39+) 089 965043
Fax (39+) 089 969657

## Abstract

**Background**: Qualitative and quantitative analysis of small non-coding RNAs by next generation sequencing (miRNA-Seq) represents a novel technology increasingly used to investigate with high sensitivity and specificity RNA population comprising microRNAs and other regulatory small transcripts. Analysis of miRNA-Seq data to gather biologically relevant information, i.e. detection and differential expression analysis of known and novel non-coding RNAs, target prediction, etc., requires implementation of multiple statistical and bioinformatics tools from different sources, each focusing on a specific step of the analysis pipeline. As a consequence, the analytical workflow is slowed down by the need for continuous interventions by the operator, a critical factor when large numbers of dataset need to be analyzed at once.

**Results**: We designed a novel modular pipeline (iMir) for comprehensive analysis of miRNA-Seq data, comprising specific tools for adapter trimming, quality filtering, differential expression analysis, biological target prediction and other useful options by integrating multiple open source modules and resources in an automated workflow. As statistics is crucial in deep-sequencing data analysis, we devised and integrated in iMir tools based on different statistical approaches to allow the operator to analyze data rigorously. The pipeline created here proved to be efficient and time-saving than currently available methods and, in addition, flexible enough to allow the user to select the preferred combination of analytical steps. We present here the results obtained by applying this pipeline to analyze simultaneously 6 miRNA-Seq datasets from either exponentially growing or growth-arrested human breast cancer MCF-7 cells, that led to the rapid and accurate identification, quantitation and differential expression analysis of ~450 miRNAs, including several novel miRNAs and isomiRs, as well as identification of the putative mRNA targets of differentially expressed miRNAs. In addition, iMir allowed also the identification of ~70 piRNAs (piwi-interacting RNAs), some of which differentially expressed in proliferating vs growth arrested cells.

**Conclusion**: The integrated data analysis pipeline described here is based on a reliable, flexible and fully automated workflow, useful to rapidly and efficiently analyze high-throughput miRNA-Seq data, such as those produced by the most recent high-performance next generation sequencers.

**Keywords**

Next generation sequencing, miRNA-Seq, data analysis pipeline, breast cancer, small non-coding RNA, microRNA, piwi-interacting RNA.

Journal of proteome research

Article

pubs.acs.org/jpr

# Molecular Mechanisms of Selective Estrogen Receptor Modulator Activity in Human Breast Cancer Cells: Identification of Novel Nuclear Cofactors of Antiestrogen−ERα Complexes by Interaction Proteomics

Francesca Cirillo,[‡,§] Giovanni Nassa,[‡,§] Roberta Tarallo,[‡] Claudia Stellato,[‡] Maria Rosaria De Filippo,[§] Concetta Ambrosino,[∥] Marc Baumann,[⊥] Tuula A. Nyman,[▽] and Alessandro Weisz*[,‡]

[‡]Laboratory of Molecular Medicine and Genomics, Department of Medicine and Surgery, University of Salerno, Baronissi, Salerno, Italy

[§]Fondazione IRCCS SDN, Napoli, Italy

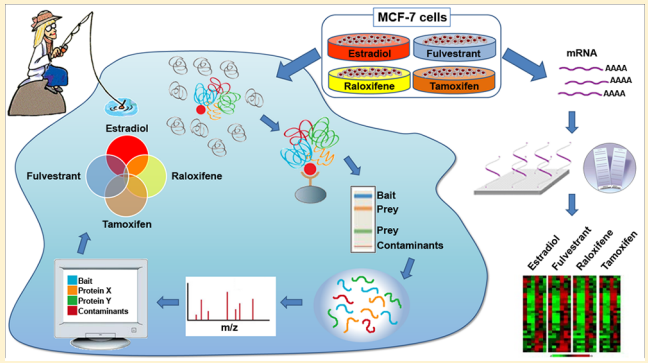
[∥]Department of Biological and Environmental Sciences, University of Sannio, Benevento, Italy

[⊥]Protein Chemistry/Proteomics Unit, Biomedicum, University of Helsinki, Helsinki, Finland

[▽]Protein Chemistry Research Group, Institute of Biotechnology, University of Helsinki, Helsinki, Finland

Supporting Information

**ABSTRACT:** Estrogen receptor alpha (ERα) is a ligand-activated transcription factor that controls key cellular pathways *via* protein−protein interactions involving multiple components of transcriptional coregulator and signal transduction complexes. Natural and synthetic ERα ligands are classified as agonists (17β-estradiol/$E_2$), selective estrogen receptor modulators (SERMs: Tamoxifen/Tam and Raloxifene/Ral), and pure antagonists (ICI 182,780-Fulvestrant/ICI), according to the response they elicit in hormone-responsive cells. Crystallographic analyses reveal ligand-dependent ERα conformations, characterized by specific surface docking sites for functional protein−protein interactions, whose identification is needed to understand antiestrogen effects on estrogen target tissues, in particular breast cancer (BC). Tandem affinity purification (TAP) coupled to mass spectrometry was applied here to map nuclear ERα interactomes dependent upon different classes of ligands in hormone-responsive BC cells. Comparative analyses of agonist ($E_2$)- vs antagonist (Tam, Ral or ICI)-bound ERα interacting proteins reveal significant differences among ER ligands that relate with their biological activity, identifying novel functional partners of antiestrogen−ERα complexes in human BC cell nuclei. In particular, the $E_2$-dependent nuclear ERα interactome is different and more complex than those elicited by Tam, Ral, or ICI, which, in turn, are significantly divergent from each other, a result that provides clues to explain the pharmacological specificities of these compounds.




**KEYWORDS:** *estrogen receptor, antiestrogen, estradiol, tamoxifen, raloxifene, ICI 182,780, breast cancer, histone H3 methyltransferase/DOT1L, Deleted in breast cancer gene 1/KIAA1967, pyruvate dehydrogenase E1 component/PDHA1*

## INTRODUCTION

Estrogens have important physiological effects on the growth, differentiation, and function of hormone dependent tissues, including breast epithelium, uterus, vagina, and ovaries. In addition, they preserve bone mineral density, reducing the risk of osteoporosis; protect the cardiovascular system by reducing cholesterol levels; and modulate cognitive functions and behavior. Estrogens are known to act mainly through a genomic pathway, whereby they influence gene expression and cellular phenotype by diffusing into the cell and binding to estrogen receptors (ERs), which in turn translocate to the nucleus, dimerize, associate with various coregulatory proteins, and subsequently bind to DNA to control transcription of responsive genes.[1−3]

The two known mammalian ER subtypes, ERα[4] and ERβ,[5] are characterized by distinct, but often overlapping, tissue distribution. They share sequence homology within the DNA binding domain and hormone recognition region, but they have different transcriptional activation properties, suggesting that

**Received:** August 8, 2012
**Published:** November 21, 2012

each of them interacts with a unique set of nuclear factors and plays different roles in the regulation of gene expression.[6−8] In the normal mammary gland, proliferating cells express low levels of ERα, which is downregulated, through an ubiquitin proteasome pathway, in the presence of its cognate ligand 17β-estradiol ($E_2$). In contrast, ERα is readily detected in the majority (70%) of human mammary tumors. This feature plays a crucial role in the progression of the human BC, because it facilitates epithelial mammary cells to switch from a hormone-dependent to a hormone-independent status. Excessive/deregulated expression of ERα is thus a risk factor for BC development, and it is associated with tumor responsiveness to hormonal treatment.[9]

The fact that estrogens promote tumorigenesis and cancer progression led to the development of endocrine therapies for BC treatment. Synthetic compounds that either act as ERα antagonists (antiestrogens) or block the function of aromatases (enzymes that catalyze locally androgen conversion to estrogen) have been developed. Antiestrogens are designed to antagonize hormone-mediated proliferation and ERα target gene expression in mammary tumor cells by competitively inhibiting $E_2$ binding to the receptor. Such synthetic compounds are important and widely used therapeutic agents.[10] Among antiestrogens, it is possible to distinguish two major classes of drugs, depending upon their functional effects. The "selective estrogen receptor modulators" (SERMs) are able to act both as receptor agonists and antagonists (partial antagonists), depending on cellular and gene promoter context as well as on the ER isoform involved. The "selective estrogen receptor downregulators/disruptors" (SERDs), instead, are pure antiestrogens (full antagonists), capable of completely blocking ERα activity, increasing receptor turnover, and/or disrupting its nuclear localization, with a concomitant reduction in ER expression in the cell both *in vitro* and *in vivo*. In general, this class of compounds is used in second line therapy against advanced BC in patients who develop resistance to SERM treatment.[11]

All these molecules are able to bind the ligand-binding domain (LBD) of ERs, whose transcriptional activity is mediated by two separate activation functions (AFs): AF-1, regulated by growth factors acting through the MAP kinase signaling pathway,[12] and AF-2, responsive to agonist ligands.[13] Binding of agonists mainly triggers AF-2 activity, whereas binding of antagonists does not.[14] This difference relies on specific structural features of ER bound to the different compounds, as repositioning of helix 12 of $E_2$-liganded ERα creates a coactivator binding site that is, instead, absent in ER−antiestrogen complexes.[15,16] Mutational and structural analyses revealed that the recognition surface created in the presence of agonist ligands is necessary in order to allow coactivators to bridge ER to the RNA pol II transcriptional machinery.[17−19]

The first antiestrogen introduced in clinical practice was Tamoxifen (also referred to in the literature as Nolvadex), the SERM prototype,[20,21] a nonsteroidal antiestrogen that antagonizes the action of estrogen and is used for treatment of all stages of ER-positive BC,[22−25] that however can adapt to chronic exposure to this drug developing resistance.[26,27] It is noteworthy that Tamoxifen acts as an estrogenic compound on bone, blood lipids, and the endometrium,[28] increasing the risk of endometrial cancer and thrombotic events. Raloxifene (also reported as LY 156,758, keoxifene, LY 139, 481-HCL, or Evista) is a second-generation SERM, designed with the aim to develop a new hormone replacement therapy to prevent osteoporosis, while decreasing endometrial and breast cancer risk as a beneficial side effect. Raloxifene and Tamoxifen exhibit similar binding affinity for the ER, but the former shows a higher estrogenic activity on bone cells.[29] SERM-bound ERα is able to dimerize and translocate to the nucleus, acquiring a specific conformation that allows coactivator recruitment to the receptor, with helix 12 directly affecting the structure and a partial reorganization of AF-2. This, which implies an incomplete recruitment of the AF-2 coactivators, is likely to be the structural basis of the specific biological activities of the SERM-activated receptor.

Fulvestrant (also known as ICI 182,780 and Faslodex), the SERD prototype, is a steroidal molecule devoid of estrogen-like activity that was designed to treat patients with ERα-positive breast tumors that developed SERM resistance.[30] In these cases, the absence of agonist activity in SERDs is thought to overcome the drug resistance. Indeed, proliferation of tamoxifen-resistant BC cell lines is strongly inhibited by fulvestrant.[31,32] Correct alignment of ERα helix 12 is prevented by SERDs, which therefore affects AF-2 function. Furthermore, this conformation results in loss of receptor dimerization, accelerated ER ubiquitination, and shuttling to the proteasome for degradation.[31,33]

Clinical evidence shows that more than 30% receptor-positive mammary tumors are unexpectedly nonresponsive to endocrine treatments. The reasons for such failure have been suggested to depend on the functions of ER and/or the intracellular signaling pathway controlled by estrogens. In this regard, dissection of the ER signaling networks in hormone-responsive BC cells, a useful approach to identify the molecular mechanisms of cell responsiveness to estrogen, may provide new insights on resistance of breast tumors to endocrine therapies.

Interaction proteomics led so far to the identification of a large number of $E_2$−ERα interactors in BC cell nuclei, including transcriptional coregulators and components of the nuclear actin pathway.[34−37] The main purpose of this study was to apply this technology to map the nuclear interactomes of ERα bound to the antiestrogenic compounds commonly used for BC treatment (i.e., ICI, Ral, and Tam), aiming at providing new mechanistic information to help explain the pharmacological activities of these drugs in BC cells *in vitro* and *in vivo*.

## EXPERIMENTAL PROCEDURES

### Cell Cultures

The human hormone-responsive mammary carcinoma cell line MCF-7 (Clontech-Takara) was cultured in Dulbecco's modified Eagle's medium containing 1 mg/mL D-glucose (Sigma-Aldrich) and supplemented with 2 mM L-glutamine, 10% FBS (HyClone), 25 units/mL penicillin, 25 units/mL streptomycin, 250 ng/mL amphotericin B, and 100 μg/mL G418 (standard growth conditions).

To study protein complexes assembly upon ligand treatments, cells were estrogen deprived (starved) by exchanging the medium to Dulbecco's modified Eagle's medium without phenol red (Sigma-Aldrich) supplemented with 2 mM L-glutamine and 5% stripped serum (dextran-coated charcoal-treated FBS) 5 days prior to performing the ligand treatments and to harvesting the cells, as described.[38]

MCF-7 cells were used to generate stable clones expressing TAP (control cells) or C-TAP-ERα (TAP-ERα expressing cells) as described earlier.[36]


dx.doi.org/10.1021/pr300753u | J. Proteome Res. 2013, 12, 421−431

### Preparation of Nuclear Extracts

The cells were harvested by scraping, washed twice in cold 1× PBS, collected by centrifugation at 1000*g*, and resuspended in 3 volumes with respect to the cell pellet of hypotonic buffer (20 mM HEPES pH 7.4, 5 mM NaF, 10 mM sodium molybdate, 0.1 mM EDTA, 1 mM PMSF, and 1× protease inhibitor mixture (Sigma-Aldrich)). Upon incubation on ice for 15 min, 0.5% Triton X-100 was added, and a cytosolic fraction was discarded after centrifugation of the samples at 15,000*g* for 30 s at 4 °C. The nuclear pellet was first washed twice in hypotonic buffer to remove any residual cytosolic contaminations and then was resuspended in 1 volume of nuclear lysis buffer (20 mM HEPES pH 7.4, 25% (v/v) glycerol, 420 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, 1× protease inhibitor mixture (Sigma-Aldrich), and 1 mM PMSF), incubated for 30 min at 4 °C on a rotating platform. The nuclear extract was clarified by centrifugation at 15,000g, for 30 min at 4 °C and then was diluted by adding 2 volumes of nuclear lysis buffer w/o NaCl. The nuclear extracts were assayed, and nonsignificant cross-contamination between the two cellular compartments could be detected by Western Blotting using an anti-$\alpha$ tubulin antibody.[36]

### Western Blotting

Western blot analyses were performed using standard protocols as previously described.[39] In detail, protein samples were denatured, separated on a 7 or 10% polyacrylamide and 0.1% SDS (SDS-PAGE), and electrotransferred onto a nitrocellulose membrane (Whatman GmbH-Schleicher & Schuell). The membrane was blocked using 5% (w/v) fat-free milk powder in 1× TBS supplemented with 0.1% (v/v) Tween20 (TBS-T). The used primary antibodies were as follows: rabbit antihuman ER$\alpha$ (sc-543, HC-20, Santa Cruz Biotechnology), rabbit anti-TAP (CAB1001, Thermo Scientific-Pierce), rabbit anti-$\alpha$-tubulin (T6199, Sigma Aldrich), mouse anti-$\beta$-actin (A1978, Sigma Aldrich), rabbit anti-$\alpha$-tubulin (T6074, Sigma Aldrich), mouse anti-DBC1/3G4 (#5857, Cell Signaling), rabbit anti-Nucleophosmin (ab52644, Abcam), rabbit anti-DOT1L/KMT4 (ab72454, Abcam), mouse anti-DBC1/3G4 (#5857, Cell Signaling), and mouse anti-Pyruvate Dehydrogenase E1-alpha subunit (ab110334, Abcam).

All antibodies were first tested to evaluate specificity and sensitivity. After extensively washing with TBS-T, the immunoblotted proteins were incubated with the appropriate horseradish peroxidase-conjugated secondary antibodies (GE Healthcare) and were detected by enhanced chemiluminescence (ECL Kit, GE Healthcare) and exposure to a medical X-ray film (FujiFilm).

### Isolation of ER$\alpha$ Nuclear Partners by Tandem Affinity Purification

Control and TAP-ER$\alpha$ expressing cells (approximately $6 \times 10^8$ cells in 500 $cm^2$ plates) were used for each tandem affinity purification (TAP) procedure. The cells were starved and stimulated with $1 \times 10^{-8}$ M ligand (17$\beta$-estradiol/$E_2$, 4-hydroxytamoxifen/Tam, Raloxifene/Ral, or Fulvestrant/ICI; all from Sigma-Aldrich) for 1 h. Cells were harvested, extensively washed with ice-cold 1× PBS, and lysed as described above. Nuclear extracts were incubated with 6 $\mu$L/mg protein IgG-Sepharose beads (IgG-Sepharose 6 Fast Flow, GE Helthcare) at 4 °C for 4 h on a rotating platform. Before incubation, the beads were equilibrated in 10 volumes of TEV buffer (50 mM Tris-HCl, pH 8.0, 0.5 mM EDTA, 1 mM DTT, 0.1% Triton X-100, and 150 mM NaCl), and washed four times with 20 volumes of IPP150 buffer (20 mM HEPES, pH 7.5, 8% glycerol, 150 mM NaCl, 0.5 mM $MgCl_2$, 0.1 mM EDTA, and 0.1% Triton X-100) at 4 °C for 15 min. At the end of the incubation, the unbound proteins were collected by centrifugation and the beads were washed with 100 volumes of IPP150 and 30 volumes of TEV buffer in a Poly-Prep Chromatography column (0.8 cm × 4 cm, Bio-Rad) at 4 °C. Thereafter, 4 bead volumes of TEV buffer containing 1 unit of TEV protease/$\mu$L of beads (Invitrogen) were added and, following incubation for 2 h at 16 °C on a shaking platform (Thermomixer, Eppendorf), the eluted proteins were collected by sedimentation.

### Nano LC-MS/MS Analysis of TEV Eluates

The partially purified protein samples from the different experimental points were concentrated by precipitation with acetone/TCA, dried, sonicated, and resuspended in Laemmli buffer followed by SDS-PAGE and visualization with Silver Staining, as previously described.[39]

All lanes on the gels were excised and were sliced into six pieces, and the proteins were in-gel digested with trypsin solution (Sequencing grade Modified Trypsin, Promega) and incubated at 37 °C overnight as described earlier.[40,41]

The resulting peptides were acidified and dissolved by addition of 0.1% TFA (Sigma-Aldrich) and analyzed by LC-MS/MS using an Ultimate 3.000 nano-LC (Dionex, Sunnyvalle, CA, USA) and a QSTAR Elite hybrid quadrupole TOF-MS (Applied Biosystems/MDS Sciex, CA, USA) with nano-ESI ionization. The LC-MS/MS samples were first loaded on a ProteCol C18 trap column (10 mm × 150 $\mu$m, 3 $\mu$m, 120 Å) (SGE Incorporated, Austin, Texas, USA), followed by peptide separation on a PepMap100 C18 analytical column (15 cm × 75 $\mu$m, 5 $\mu$m, 100 Å) (LC Packings/Dionex) at 200 nL/min. The separation gradient consisted of 0−50% B in 50 min, 50% B for 3 min, 50−100% B in 2 min, and 100% B for 3 min (buffer A: 0.1% formic acid; buffer B: 0.08% formic acid in 80% acetonitrile). MS data were acquired using Analyst QS 2.0 software. The information-dependent acquisition method consisted of a 0.5 s TOF MS survey scan of $m/z$ 400−1400. From every survey scan, the two most abundant ions with charge states +2 to +4 were selected for product ion scans. Once an ion was selected for MS/MS fragmentation, it was put on an exclusion list for 60 s.

The LC-MS/MS data were searched against SwissProt release 22062011 (529056 sequences; 187423367 residues; Taxonomy *Homo sapiens* (human): 20236 sequences) for all samples using the in-house Mascot (version 2.2, Matrix Science) through the ProteinPilot 2.0.1 interface. The criteria for Mascot searches were the following: human-specific taxonomy, trypsin digestion with one missed cleavage allowed, and oxidation of methionine as a variable modification and carbamidomethylation as a fixed modification. For the LC-MS/MS spectra the maximum precursor ion mass tolerance was 50 ppm and the MS/MS fragment ion mass tolerance was 0.2 Da, and peptide charge states of +1, +2 or +3 were used. All reported protein identifications were statistically significant because, instead of a Standard Scoring, a MudPIT scoring was used which automatically filters low scoring peptide masses. Mascot search results are listed in the Supporting Information (Tables S5−S8 for $E_2$, ICI, Ral, and Tam, respectively). Moreover, raw MS/MS fragmentation data for single peptide-based protein identification are included in Supporting Information Tables S9−S12 for $E_2$, ICI, Ral, and Tam, respectively. To eliminate the redundancy of proteins that


dx.doi.org/10.1021/pr300753u | *J. Proteome Res.* 2013, 12, 421−431

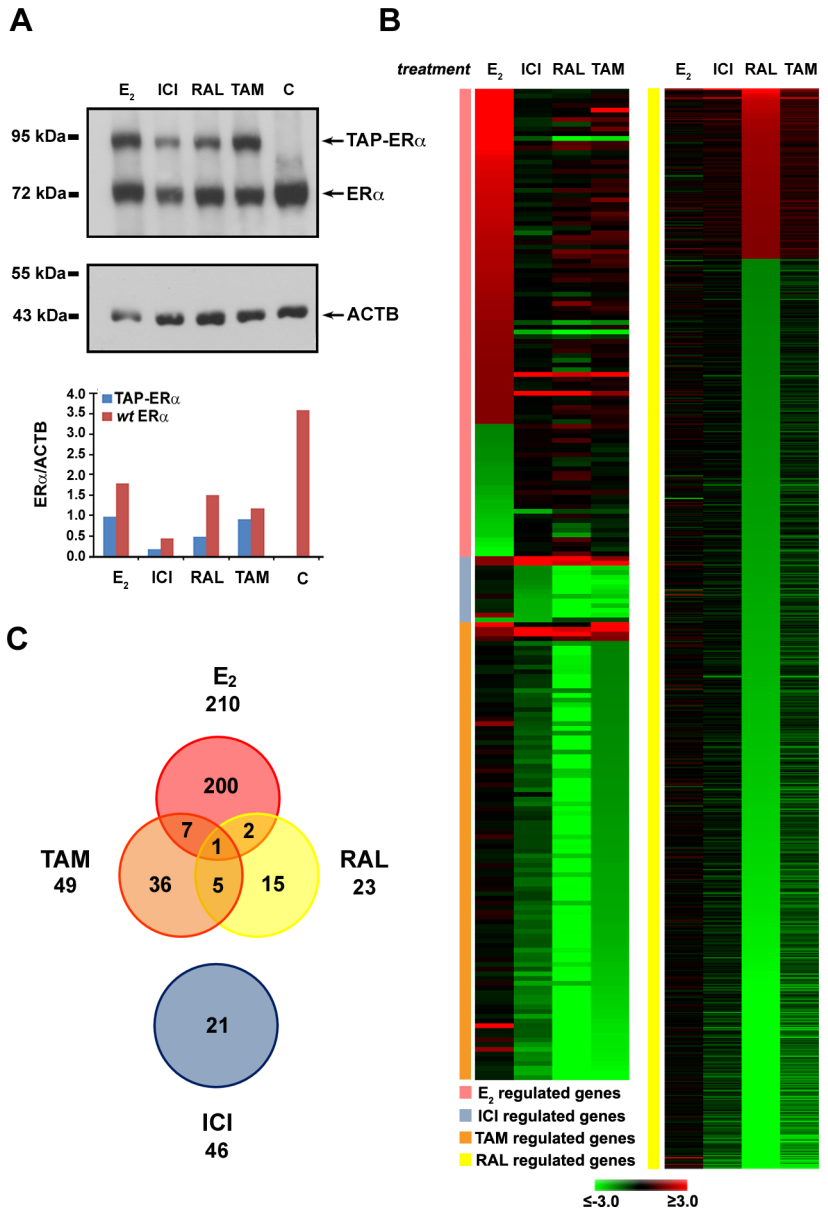



**Figure 1.** Interaction proteomics and genomics of agonist- and antagonist-bound ER$\alpha$ in MCF-7 cells. (A) Western blot analysis of ER$\alpha$ and TAP-ER$\alpha$ in MCF-7 nuclear extracts, normalized to nuclear $\beta$-actin (ACTB) concentration in the same samples, from cells treated with $E_2$, ICI, Ral, and Tam ($10^{-8}$ M; 1 h). (C) Wild-type MCF-7 cells, not expressing TAP-ER$\alpha$, stimulated with $E_2$, used as negative control. Relative quantitation of TAP-ER$\alpha$ and ER$\alpha$ compared to ACTB is shown in the histogram. (B) Heatmaps summarizing the results of transcriptome analyses of TAP-ER$\alpha$-expressing MCF-7 cells treated ($10^{-8}$ M; 12 h) with $E_2$ (top-left), ICI (midleft), Tam (bottom-left), or Ral (right). For each ligand, gene expression fold-changes were calculated with respect to the untreated control. (C) Venn diagram showing overlaps of interactomes identified following cell treatment with $E_2$, Tam, or Ral ($10^{-8}$ M; 1 h). Numbers reported below each symbol indicate the total number of specific interactors identified in purified samples by MS. The number within the isolated circle at the bottom of the panel indicates interactors specific for ICI treated cells.

appear in the database under different names and accession numbers, the single protein member with the highest protein score (top rank) was selected from multiprotein families for the identification results.

## Gene Ontology Analyses

Statistically over-represented biological processes were identified among the sets of proteins identified by MS analyses in each of the four experimental conditions by means of GOFFA,[42] a bioinformatics tool for the functional analysis of genomic and proteomic data, developed for ArrayTrack, that starting from a list of genes or proteins identifies Gene Ontology (GO) terms associated with each of them. GOFFA determines the statistical significance of a GO term using Fisher's Exact Test. For this study, the list of genes expressed in MCF-7 cells and identified by microarray-mediated gene expression profiling (see below), was used as a reference and for each dataset the GO terms over-represented respect to the reference with a $p$-value $\leq 0.05$ were selected. In addition, GO analysis was performed also using as a reference a list of proteins identified experimentally in MCF-7 cells (listed in Supporting Information Table S14-A), by means of the Gene Ontology Enrichment Analysis and Visualization Tool (GORILLA; http://cbl-gorilla.cs.technion.ac.il).


dx.doi.org/10.1021/pr300753u | J. Proteome Res. 2013, 12, 421−431

### Protein Complexes Immunoprecipitation

For immunoprecipitation of endogenous ERα or DOT1L, to nuclear extracts from MCF-7 cells (800−2000 μg proteins) was added 2.0−2.5 μg/mg protein specific Abs (rabbit antihuman ERα: sc-543, HC-20, Santa Cruz Biotechnology and rabbit anti-DOT1L/KMT4: A300-954A, Bethyl), and the mixture was incubated for 1−3 h at 4 °C with stirring via rotation; then Protein A/G Plus-Agarose was added for 1 h. Immunoprecipitated proteins were collected by centrifugation, and after extensive washing, the beads were resuspended in Laemmli buffer and subject to SDS-PAGE and Western blotting as described previously.[36]

### RNA Purification

Total RNA was extracted from TAP-ERα expressing cells, using the standard RNA extraction with TRI Reagent (Sigma-Aldrich) method, as described.[43] Cells were starved and total RNA was extracted after stimulation with $1 \times 10^{-8}$ M ligand ($E_2$, Tam, Ral, or ICI) or ethanol vehicle for 12 h. In each case RNA extracted from two independent biological replicates was used. Before use, the RNA concentration of each sample was assayed with a ND-1000 spectrophotometer (NanoDrop) and its quality assessed with the Agilent 2100 Bioanalyzer with an Agilent RNA 6000 Nano kit (Agilent Technologies).

### RNA Expression Profiling

For mRNA expression profiling, 500 ng of total RNA was reverse transcribed, as described previously[44] and used for synthesis of cDNA and biotinylated cRNA according to the Illumina TotalPrep RNA Amplification Kit (Ambion, Cat. no. IL1791) protocol. For each sample, 750 ng of cRNA was hybridized for 18 h at 58 °C on Illumina Human HT-12v4 BeadChips (Illumina Inc.), as described earlier,[43] and subsequently scanned with the Illumina iScan. Data analyses were performed with GenomeStudio software v2011.1 (Illumina Inc.), by comparing all values obtained at each time point against the 0 h values. Data were normalized with the quantile normalization algorithm, and genes were considered as detected if the detection $p$-value was lower than 0.01. Statistical significance was calculated with the Illumina DiffScore, a proprietary algorithm that uses the bead standard deviation to build an error model. Only genes with a DiffScore $\leq -30$ and $\geq 30$, corresponding to a $p$-value of 0.001, were considered as statistical significant.[43,44] Raw microarray data have been deposited, in a format complying with the Minimum Information About a Microarray Gene Experiment (MIAME) guidelines of the Microarray Gene Expression Data Society (MGED), in the EBI ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) with Accession Number E-TABM-1196.

## RESULTS AND DISCUSSION

### Evaluation of Ligand Effects on the Intracellular Localization of Wild-Type and TAP-Tagged ERα

The cellular model used in this study was derived from hormone-responsive human BC MCF-7 cells, naturally expressing ERα and widely used to investigate signal transductions by ERs in BC and to test the pharmacological effects of ER ligands. MCF-7 cells were stably transfected with an expression vector encoding ERα fused at the C-terminus with a TAP tag[45,46] that can act as a "bait" for isolation of native ER-containing multiprotein complexes by tandem affinity purification (TAP). These cells have been used successfully for mapping and functional characterization of $E_2$-induced ERα nuclear interactome by TAP.[36,37,47] As antiestrogens have been described to influence in different ways the cellular level and/or localization of ERα, the behavior of the exogenous fusion protein with respect to the endogenous receptor was assessed upon cell treatment with either $E_2$, ICI, Ral, or Tam. In all cases, the nuclear levels of TAP-ERα were assessed by Western blotting and compared to those relative to endogenous ERα in the same samples or, for $E_2$, in *wt* MCF-7 cells (Figure 1A). Results, summarized in the histogram, show that the exogenous receptor behaves similarly to the endogenous one in all cases, as demonstrated previously by Ambrosino et al.[36] Interestingly, comparing the results obtained for edogenous receptor in $E_2$-treated *wt* and TAP-ERα cells, the former exhibit higher ERα levels, possibly due to inhibition of endogenous receptor expression by the exogenous protein.[36] The pure antiestrogen ICI disrupts nuclear-cytoplasmic shuttling of both ERα and TAP-ERα, possibly by inducing proteasome-dependent ER degradation,[48,49] as 1 h treatment of the cells with this compound causes a modest increase of receptor concentration in the nuclear extracts, when compared to that elicited by $E_2$ (Figure 1A). On the other hand, the two SERMs (Ral and Tam) induce substantial nuclear accumulation of both receptor forms, with Tam being less effective than $E_2$, but more than Ral, as described earlier for endogenous ERα in this cell type.[50] Kinetic evaluation of nuclear translocation of ER was performed after 1, 6, and 12 h of treatment with each of the compounds studied by WB analysis of ERα and TAP-ERα both in the nucleus and in the cytoplasm, with β-actin and α-tubulin, respectively, used as controls. Results, reported in Supporting Information Figure S1, show that while $E_2$ and SERMs induce a substantial accumulation of both ERs in the nuclear compartment for up to 12 h of treatment, ICI effects are not only less pronounced but also dynamic, since nuclear ER concentration decreases between 1 and 6 h and then rises again at 12 h.

### Effects of Estrogen and Antiestrogens on the Transcriptome of TAP-ERα Expressing Cells

ERα exerts its biological effects through several mechanisms, that all converge on regulation of target gene expression. As the different ligands affect recruitment of coregulators on ERα, their influence on the ER dependent transcriptome in MCF-7 cells was investigated by gene expression profiling with oligonucleotide microarrays. To this end, TAP-ERα expressing MCF-7 cells were stimulated for 12 h with either $E_2$, ICI, Ral, or Tam, and total RNA was extracted, fluorescently labeled, and analyzed on high-density oligonucleotide microarrays. This time of stimulation was chosen in order to better evaluate early, primary responses to the ligands, with respect to late, secondary events in TAP-ERα cells, as shown earlier for $E_2$ in *wt* MCF-7[3] and another hormone-responsive BC cell line.[51] Untreated TAP-ERα cells were used as a control.

The results obtained are summarized in the heatmaps reported in Figure 1B, where data are reported relative to mRNAs that showed a ≥2 fold-change, with respect to the control, in response to stimulation with each of the four compounds tested. In each case, changes in expression of the same mRNA under all four conditions are also reported, side-by-side, to highlight similarities and differences in gene response to different ER ligands. Results show that gene activation clearly prevails over inhibition in response to $E_2$ stimulation, as $^2/_3$ of the transcripts show significantly higher levels in treated vs untreated cells, while, on the contrary, gene down-regulation events appear predominant following anti-


dx.doi.org/10.1021/pr300753u | J. Proteome Res. 2013, 12, 421−431

**Table 1. Proteins Identified Specifically in Nuclear Extracts from SERM-Treated Cells**

| SwissProt ID | protein name | gene name | peptides matched | | sequence coverage (%) | | MOWSE score | |
|---|---|---|---|---|---|---|---|---|
| | | | Ral | Tam | Ral | Tam | Ral | Tam |
| B2RTY4 | Myosin-IXa | MYO9A | 6 | 6 | 2 | 2 | 30 | 27 |
| O95140 | Mitofusin-2 | MFN2 | 2 | 3 | 2 | 3 | 28 | 30 |
| P08559 | pyruvate dehydrogenase E1 component subunit $\alpha$, somatic form | PDHA1 | 5 | 1 | 11 | 3 | 38 | 48 |
| Q5VUG0 | Scm-like with four MBT domains protein 2 | SFMBT2 | 2 | 1 | 1 | 1 | 30 | 30 |
| Q8TEK3 | histone-lysine *N*-methyltransferase, H3 lysine-79 specific | DOT1L | 5 | 2 | 3 | 0 | 28 | 27 |

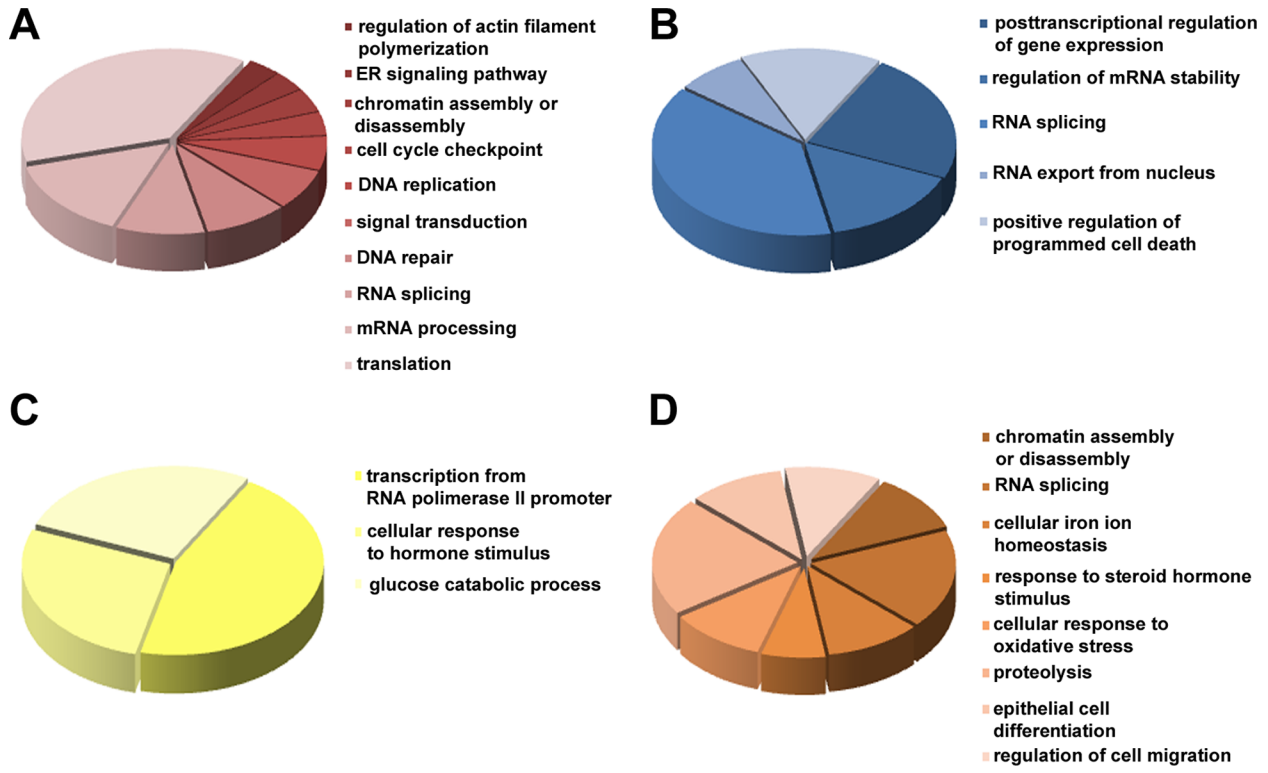



**Figure 2.** Gene ontology analysis of the biological processes involving the proteins interacting with $E_2$- (A), ICI- (B), Ral- (C), or Tam-bound (D) ER$\alpha$. For each treatment a pie chart highlights the most significant cellular processes involving the proteins recruited to the receptor by each of the ligands studied.

estrogen treatment, independently from the nature of the drug used. As expected, under these conditions most of the genes regulated by estrogen do not respond similarly to antiestrogens. When comparing the overall responses of the MCF-7 cell transcriptome to the three antiestrogens tested, the effects appear strikingly different. First of all, the total number of genes responding to ICI is very low, when compared to the responses elicited by two SERMs, in agreement with the low nuclear ER concentration in the presence of this ligand and the known effects of ICI in hormone-responsive BC cells,[52] with ICI-regulated genes generally responding similarly also to Tam and Ral. Furthermore, Tam-responsive genes are more numerous than ICI-regulated ones and show similar regulation by Ral, but not by $E_2$. This is in accordance with the observation that these SERMs promote ER$\alpha$ translocation to the nucleus and induce very similar conformational changes of the receptor, that are different from those promoted by binding of the cognate hormone. The effects of Ral on the cell transcriptome, however, are much more evident and, in most cases, unique to this drug.

Taken together, the results shown in Figure 1 reproduce correctly the known biological effects of these drugs in hormone-responsive BC cells, that are here confirmed to be strictly related to the chemical structure of each compound and its ER binding properties, to indicate that the experimental model described here is suitable to identify functional protein partners of antiestrogen-bound ER$\alpha$ by interaction proteomics.

## Identification of Proteins Recruited by ER$\alpha$ in the Nucleus of BC Cells in Response to $E_2$, ICI, Ral, or Tam

In order to identify ER$\alpha$ partner proteins specifically recruited by the receptor upon binding of an agonist ($E_2$) or of different antagonist (ICI, Ral, Tam) ligands, partially purified ER$\alpha$ protein complexes isolated from native MCF-7 cell nuclear extracts were subjected to MS analysis (nanoLC-MS/MS). As control, *wt* cells, lacking the TAP-tagged receptor, were subject to the same purification–identification protocol and all proteins identified in these samples were considered not specific, and, for this reason, when present, they were discarded from the lists of specific ER$\alpha$ interactors and not considered further, as described previously.[37] In addition, when focusing on SERMs-specific ER$\alpha$ complexes, the proteins identified in ICI-treated samples were subtracted from the number of the Ral and Tam interactors and listed as a separate set. Two biological replicates were performed, and when the resulting MS data were analyzed


dx.doi.org/10.1021/pr300753u | *J. Proteome Res.* 2013, 12, 421–431

separately, a very good reproducibility was observed (>60% identified proteins in common between the replicas), suggesting reliability of the purification procedure. To identify ligand-specific ER-associated nuclear proteins, the MS results from the two biological replicates were combined and analyzed together, to obtain more robust data sets. Results of this analysis, detailed in the Experimental Procedures section, are listed in Supporting Information Tables S1−S4 for $E_2$, ICI, Ral, Tam, respectively, and are summarized in Figure 1C.

Receptor activation by $E_2$ resulted in interaction with a set of nuclear proteins (210) significantly larger than those observed with ICI (46), Ral (23), or Tam (49). This result, that confirms our previous observations,[36,37] is likely to be due to the optimal receptor conformation promoted by the agonist, as well as to formation of stable ER homodimers, that provide an efficient docking site for interacting proteins. Furthermore, this could be explained, at least in part, by the relatively higher concentration of ERα in these samples, which may have facilitated isolation and/or MS identification of interacting proteins. It is worth mentioning that the number of molecular partners of ERα in the sample treated with $E_2$ for 1 h is comparable to what was previously reported for a 2 h stimulation with $E_2$ of the same cells, but the two sets share only about 50% of proteins,[37] in agreement with the highly dynamic ER−protein interactions occurring on BC cell chromatin during the earlier phases of hormonal stimulation.[53]

The vast majority of the interactors identified are ligand specific and, as mentioned above, their number in SERM-treated cells is significantly lower than that in $E_2$-stimulated cells, with the compositions of the Ral and Tam interactomes being rather different from each other and clearly distinct from those of ICI samples. Indeed, comparative analysis of the lists of ERα interactors identified with the four compounds tested shows that the majority of them are ligand-specific: 200 for $E_2$, 15 for Ral, 36 for Tam, and 21 for ICI (Figure 1C). Interestingly, five proteins not present in the $E_2$ treated samples are specific to both SERMs (Table 1), suggesting that they might represent specific SERM effectors (see also below). It is worth mentioning that while the patterns of interactors detected with the two SERMs and $E_2$ are very different from each other, the SERD appears to promote recruitment of a relatively larger number of proteins in common with $E_2$ (19/46), a result that could relate to the fact that ICI, unlike SERMs, has a steroid structure like that of $E_2$ and could therefore induce a conformational change on the receptor that, to some extent, is structurally closer to that elicited by binding of the natural hormone.

The number of interactors identified does not seem to correlate only with concentration of ERα in nuclear extracts or purified samples, since, for example, the amount of receptor in Tam samples is only slightly lower than that in $E_2$ samples but much higher than that in Ral or ICI samples, while the difference in number of binding proteins identified in $E_2$ vs Tam samples is significant and that in Tam vs Ral or ICI samples is very small (compare results in Figure 1C with those in Figures 1A and 2A). On the basis of this observation, the known differences in biological activity in the BC cells of the compounds tested, and the results reported in Figure 1B, we suggest that the lists of ERα interactors identified in this study, the majority of which were not shown before to be partners of the antiestrogen-bound receptor, represent a snapshot of the early and specific functional complexes formed by this protein in BC cell nuclei upon binding to antiestrogens, exploitable now to identify the molecular mechanisms that determine the variegated pharmacological effects of these drugs in BC cells.

Functional analysis of the biological processes that involve the ERα interacting proteins identified in this study by gene ontology highlights significant differences between agonist ($E_2$)-, SERD-, and SERM-specific interactomes (Figure 2 and Supporting Information Table S13), that reflect also known effects of these ligands in BC cells. In particular, estrogen promotes recruitment by ER of proteins involved in DNA replication and cell cycle progression, chromatin remodeling, gene transcription, and RNA splicing and actin cytoskeleton organization, while components of the ICI-dependent interactome participate in the control of mRNA stability and translation and in regulation of apoptosis, all processes associated with the cytostatic actions of this drug. On the other hand, the proteins specifically bound to SERM-ER are specifically involved not only in regulation of gene expression and signal transduction but also in proteolysis, epithelial cell differentiation, cell migration, and response to oxidative stress. While most of these functions remain to be elucidated in the context of hormone-responsive BC cells, this result confirms the existence of common pathways controlled by estrogen and SERMs in this cell type, clearly distinct from those specifically affected by ICI. Similar differences between the four lists of interactors were observed also when GO term enrichment analysis was performed using as background (reference) a list of MCF-7 proteins detected experimentally, obtained by combining published results (see Supporting Information Tables S14-A to E for background list, $E_2$, ICI, Ral, Tam, respectively).

To validate the results obtained by mass spectrometry, WB analysis was carried out using a selection of antibodies directed against some of the most interesting proteins exhibiting ligand-specific association with ERα. Among these, we selected the ICI-ERα specific interactor KIAA1967 protein, also known as Deleted in breast cancer gene 1, whose expression in MCF-7 cell nuclei was unaffected by ligand treatment (Figure 3A). In agreement with the MS results, KIAA1967 was prevalently detected by WB in purified TAP−ERα complexes from SERD-treated cells (Figure 3B), despite the low concentration of ERα in these samples. A slight amount of this protein in $E_2$ and Tam samples was close to that detected in the control sample, despite the very high concentration of receptor under these conditions, confirming a preferential interaction of KIAA1967 with SERD-bound ERα. Recruitment of KIAA1967 by ICI-liganded ERα may have important antitumor effects, as this protein has been shown to be able to interact directly with SIRT1 and to inhibit the activity of this enzyme both *in vitro* and *in vivo*.[54] SIRT1 is involved in cancer cell growth and survival, due also to its antiapoptotic activity[55,56] and ability to silence tumor suppressor genes.[57] Interestingly, KIAA1967 has been found overexpressed in cancer cells,[58−61] suggesting that recruitment of KIAA1967 by ICI-liganded ERα may target key cancer genes, resulting in their silencing by SIRT1.

Pyruvate dehydrogenase E1 component subunit α (PDHA1) appears, instead, to be a preferred interactor of ERα−Ral and, to a lesser extent, −Tam complexes, as confirmed by WB (Figure 3). Surprisingly, a significant increase of PDHA1 can be observed in the crude nuclear extracts upon treatment of the cells with Ral and, to a lesser extent, with Tam, but not with $E_2$ ICI (Figure 3A). This appears to be the result of drug-induced nuclear translocation of the protein, as the total cellular concentration of PDHA1 did not change significantly following treatment (Supporting Information Figure S2) and WB analysis

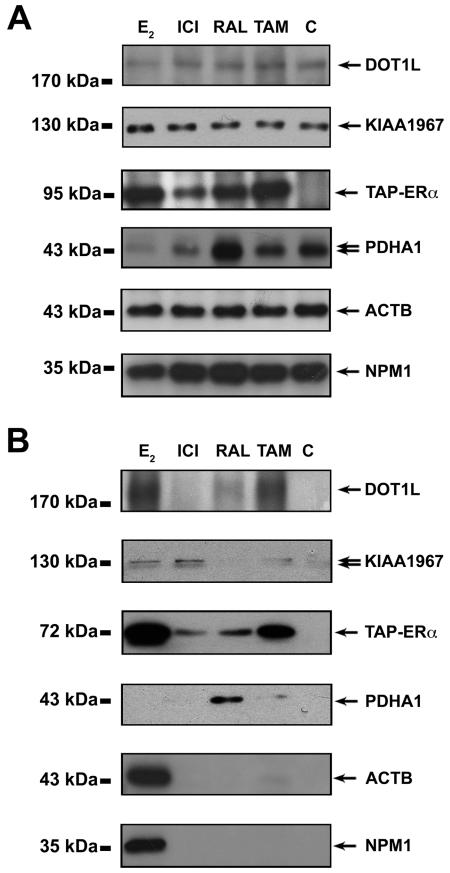



**Figure 3.** Western blot analysis of selected ER nuclear interactors identified by TAP. (A) Whole nuclear extracts from wild-type MCF-7 (C) or TAP-ER$\alpha$ cells stimulated with $E_2$, ICI, Ral, or Tam ($10^{-8}$ M; 1 h). (B) Validation of MS data. Confirmation of TAP-ER$\alpha$ interaction with DOT1L (DOT1-like, histone H3 methyltransferase), KIAA1967 (Deleted in breast cancer gene 1/KIAA1967), PDHA1 (pyruvate dehydrogenase (lipoamide) alpha 1), ACTB (nuclear $\beta$-actin), or NPM1 (Nucleophosmin/Nucleolar phosphoprotein B23/Numatrin), measured in partially purified samples (bound or eluate TAP fractions). Double arrows indicate the presence of two bands detected by the antibodies against human PDHA1 and KIAA1967 proteins.

of the cytoplasm fractions showed reduction of the cytoplasmic concentration of PDHA1 in correspondence with its increase in the nuclear compartment (data not shown). Pyruvate dehydrogenases, that exert a pivotal role in cellular metabolism, were recently assigned an additional function in the nucleus as coactivators in STAT5-dependent gene transcription in response to interleukin IL-3. This was reported specifically for the pyruvate dehydrogenase E2 component (PDHE2), which interacts both in the nucleus and in the cytoplasm with the E1 component (PDHE1),[62] suggesting that PDHA1 may contribute to the function of PDHE1 as coregulator of SERM–ER$\alpha$ complexes for target gene regulation.

Nuclear levels of $\beta$-actin (ACTB) and Nucleophosmin (Nucleolar phosphoprotein B23/Numatrin NPM1), two well characterized functional partners of estrogen-activated ER$\alpha$ in BC cell nuclei,[36] were, instead, not affected by treatments (Figure 3A), but these proteins were clearly detected in $E_2$ stimulated samples (Figure 3B), in agreement with the MS results, and to a much lower extent, in Tam-treated samples, in agreement with previously published results[36,37,63] and with the MS output data, that in Tam samples detected peptides from this protein but assigned a low MOWSE score. In view of the role that $\beta$-actin plays in regulation of gene expression, by recruitment of chromatin remodeling complexes and a positive effect on RNA polymerase II activity, and the known role of Nucleophosmin in ribosome biogenesis, a reduced recruitment of these proteins to antiestrogen-bound ER$\alpha$ will result in reduction of receptor effects on the above-mentioned processes, which might explain the differences in gene regulation shown in Figure 1B and the antiestrogenic effects of SERMs and SERDs in BC cells both *in vitro* and *in vivo*. Interestingly, an additional SERM-spcecific interactor, Myosin-IXa, is itself a component of the actin-based motors involved in intracellular movements and, in particular, in collective epithelial cell migration that facilitate formation and maintenance of continuous cell layers. In MCF-7 cells, estrogens promote acquisition of mesenchymal-like features associated with metastasis development and stimulate movement of a subset of estrogen-treated cells as cell clusters (collective motility). Antiestrogens, such as Tam, prevent both phenomena.[64] Myosin-IXa has been suggested to locally regulate Rho proteins and assembly of thin actin bundles associated with nascent cell–cell adhesion, which is required to sustain the collective migration of epithelial cells. Recruitment of this protein by Ral- and Tam-bound ER$\alpha$, identified here in the nucleus, could also occur in the extranuclear compartment, where it may result in reduction of the collective cell migration. Alternatively, binding of this protein to SERM-ER could result in accumulation of this protein in the nucleus, diverting it from its activities outside this compartment.

Another interesting SERM-specific ER$\alpha$ interactor discovered here is the DOT1-like, histone H3 methyltransferase DOT1L protein, a histone code "writer" lacking the SET domain. DOT1L, that is responsible for regulating gene expression through histone-methylation (H3K79),[65] can bind to several MLL-fusion partners found in acute leukemia and, through this binding, is thought to promote oncogenesis.[66,67]

In order to further investigate DOT1L-ER$\alpha$ interaction by an independent experimental approach, coimmunoprecipitations were performed. *wt* MCF-7 cells were stimulated with $E_2$, ICI, Ral, or Tam ($10^{-8}$ M, 1 h) or the vehicle alone (V). Whole nuclear protein extracts were immunoprecipated with specific antibodies against either DOT1L or ER$\alpha$, and the immunoprecipitates were analyzed by Western Blotting with both Abs. The results shown in Figure 4 confirm preferential DOT1L interaction not only with SERM- but also with ICI-bound ER$\alpha$.

Interestingly, inhibition of DOT1L has been shown to be a valid therapeutic strategy in tumor treatment.[68] Recruitment of DOT1L by antiestrogen–ER$\alpha$ complexes (Figures 3B and 4 and Table 1) could thus play a role in controlling the enzymatic activity of DOT1L and, therefore, modulate its downstream targets.

## ■ CONCLUSIONS

This study provides for the first time a comparative analysis of the effects of antiestrogens on the nuclear ER$\alpha$ interactome of hormone-responsive human BC cells.

The results clearly show that the protein complexes recruited by ER$\alpha$ upon estrogen ($E_2$) and antiestrogen (ICI, Ral, Tam) stimulation share few components, as the majority of the receptor partners identified appear to be ligand-specific. This evidence points to the possibility, suggested by a number of indirect observations, that estrogenic and antiestrogenic compounds may induce different biological effects in BC cells


dx.doi.org/10.1021/pr300753u | *J. Proteome Res.* 2013, 12, 421–431

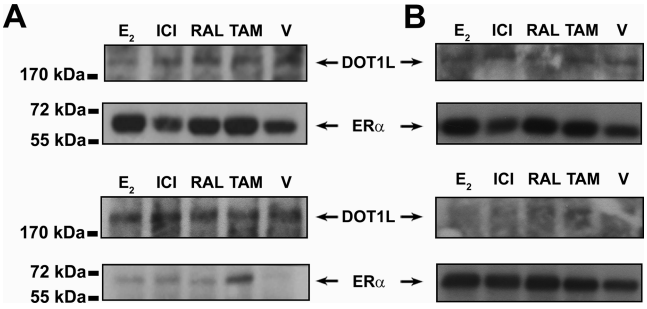



**Figure 4.** DOT1L-ER$\alpha$ coimmunoprecipitation. (A) Immunoprecipitation with anti-DOT1L Abs. Upper panel: whole nuclear extracts from wild-type MCF-7 cells stimulated with $E_2$, ICI, Ral, Tam ($10^{-8}$ M; 1 h), or vehicle alone (V); lower panel: immunoprecipitates from the same samples. (B) Immunoprecipitation with anti- ER$\alpha$ Abs. Upper panel: whole nuclear extracts from wild-type MCF-7 cells stimulated with $E_2$, ICI, Ral, Tam ($10^{-8}$ M; 1 h), or vehicle alone (V); lower panel: immunoprecipitates from the same samples.

*via* ER$\alpha$ by promoting recruitment to the receptor of specific molecular partners. Comparison of the number of interactors shared between two receptor complexes to the total number of interactors identified suggests that ER$\alpha$ complexes recruited upon SERM stimulation share a relatively higher number of common interactors. This result is in agreement with the possibility of a direct relationship between the structure of the compound, the molecular composition of the interactome, and the biological effects elicited by the receptor. The known functions of several proteins identified here open new venues to investigate the molecular mechanisms underlying SERM inhibition of BC cells proliferation and promotion of cell death and to understand the events that lead to loss of breast tumor sensitivity to antiestrogen-based therapies.

## ■ ASSOCIATED CONTENT

### Supporting Information

List of proteins identified experimentally in MCF-7 cells; Mascot search results from MS data of the two biological replicates, combined and analyzed together; raw MS/MS fragmentation data for single peptide-based protein identification; results of kinetic evaluation of nuclear translocation of ER$\alpha$; GO analysis results of the biological processes involving the ER$\alpha$ ligand-stimulated interactors. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*Address: Laboratory of Molecular Medicine and Genomics, University of Salerno, via S. Allende, 184081 Baronissi, Salerno, Italy. E-mail: aweisz@unisa.it. Tel. (39+) 089 965043. Fax (39+) 089 969657.

### Author Contributions

[§]F.C. and G.N. equally contributed to this work and should, therefore, both be considered as first author.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors wish to thank Dr.s Rosario Casale and Maria Francesca Papa for technical assistance. Research supported by: Italian Association for Cancer Research (Grant IG-8586), Italian Ministry for Education, University and Research (Grant PRIN 2008CJ4SYW_004 and 2010LC747T_002), University of Salerno (Fondi FARB 2011-2012), Fondazione con il Sud (Grant 2009-PdP-22), Union for International Cancer Control (ICRETT fellowship ICR/11/025/2011 to FC), Federation of European Biochemical Societies (FEBS Short-term Fellowship to GN) and Fondazione IRCCS SDN. FC is a PhD student in "Molecular Pathology and Physiopathology" University of Napoli 'Federico II', CS is a PhD student in "Molecular Oncology, Experimental Immunology and Innovative Therapy Development" University of Catanzaro 'Magna Graecia' and MRDF is a PhD student in "Computational Biology and Bioinformatics" University of Napoli 'Federico II'.

## ■ REFERENCES

(1) McKenna, N. J.; O'Malley, B. W. Nuclear receptors, coregulators, ligands, and selective receptor modulators: making sense of the patchwork quilt. *Ann. N. Y. Acad. Sci.* **2001**, *949*, 3−5.

(2) Cicatiello, L.; Addeo, R.; Sasso, A.; Altucci, L.; Petrizzi, V. B.; Borgo, R.; Cancemi, M.; Caporali, S.; Caristi, S.; Scafoglio, C.; Teti, D.; Bresciani, F.; Perillo, B.; Weisz, A. Estrogens and progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of cyclin D1 to its own gene promoter. *Mol. Cell. Biol.* **2004**, *24*, 7260−7274.

(3) Cicatiello, L.; Mutarelli, M.; Grober, O. M.; Paris, O.; Ferraro, L.; Ravo, M.; Tarallo, R.; Luo, S.; Schroth, G. P.; Seifert, M.; Zinser, C.; Chiusano, M. L.; Traini, A.; De Bortoli, M.; Weisz, A. Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. *Am. J. Pathol.* **2010**, *176*, 2113−2130.

(4) Greene, G. L.; Gilna, P.; Waterfield, M.; Baker, A.; Hort, Y.; Shine, J. Sequence and expression of human estrogen receptor complementary DNA. *Science* **1986**, *231*, 1150−1154.

(5) Mosselman, S.; Polman, J.; Dijkema, R. ER beta: identification and characterization of a novel human estrogen receptor. *FEBS Lett.* **1996**, *392*, 49−53.

(6) Speirs, V.; Adams, I. P.; Walton, D. S.; Atkin, S. L. Identification of wild-type and exon 5 deletion variants of estrogen receptor beta in normal human mammary gland. *J. Clin. Endocrinol. Metab.* **2000**, *85*, 1601−1605.

(7) Barkhem, T.; Carlsson, B.; Nilsson, Y.; Enmark, E.; Gustafsson, J.; Nilsson, S. Differential response of estrogen receptor alpha and estrogen receptor beta to partial estrogen agonists/antagonists. *Mol. Pharmacol.* **1998**, *54*, 105−112.

(8) McInerney, E. M.; Weis, K. E.; Sun, J.; Mosselman, S.; Katzenellenbogen, B. S. Transcription activation by the human estrogen receptor subtype beta (ER beta) studied with ER beta and ER alpha receptor chimeras. *Endocrinology* **1998**, *139*, 4513−4522.

(9) McCarty, K. S., Jr.; Kinsel, L. B.; Georgiade, G.; Leight, G.; McCarty, K. S., Sr. Long-term prognostic implications of sex-steroid receptors in human cancer. *Prog. Clin. Biol. Res.* **1990**, *322*, 279−293.

(10) Bentrem, D.; Dardes, R.; Liu, H.; MacGregor-Schafer, J.; Zapf, J.; Jordan, V. Molecular mechanism of action at estrogen receptor alpha of a new clinically relevant antiestrogen (GW7604) related to tamoxifen. *Endocrinology* **2001**, *142*, 838−846.

(11) Waters, E. A.; McNeel, T. S.; Stevens, W. M.; Freedman, A. N. Use of tamoxifen and raloxifene for breast cancer chemoprevention in 2010. *Breast Cancer Res. Treat.* **2012**, *134* (2), 875−80.

(12) Kato, S.; Endoh, H.; Masuhiro, Y.; Kitamoto, T.; Uchiyama, S.; Sasaki, H.; Masushige, S.; Gotoh, Y.; Nishida, E.; Kawashima, H.; Metzger, D.; Chambon, P. Activation of the estrogen receptor through phosphorylation by mitogen-activated protein kinase. *Science* **1995**, *270*, 1491−1494.

(13) Kumar, V.; Green, S.; Stack, G.; Berry, M.; Jin, J. R.; Chambon, P. Functional domains of the human estrogen receptor. *Cell* **1987**, *51*, 941−951.
(14) Berry, M.; Metzger, D.; Chambon, P. Role of the two activating domains of the oestrogen receptor in the cell-type and promoter-context dependent agonistic activity of the anti-oestrogen 4-hydroxytamoxifen. *EMBO J.* **1990**, *9*, 2811−2818.
(15) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927−937.
(16) Greschik, H.; Flaig, R.; Renaud, J. P.; Moras, D. Structural basis for the deactivation of the estrogen-related receptor gamma by diethylstilbestrol or 4-hydroxytamoxifen and determinants of selectivity. *J. Biol. Chem.* **2004**, *279*, 33639−33646.
(17) Kumar, R.; Johnson, B. H.; Thompson, E. B. Overview of the structural basis for transcription regulation by nuclear hormone receptors. *Essays Biochem.* **2004**, *40*, 27−39.
(18) Wrenn, C. K.; Katzenellenbogen, B. S. Structure-function analysis of the hormone binding domain of the human estrogen receptor by region-specific mutagenesis and phenotypic screening in yeast. *J. Biol. Chem.* **1993**, *268*, 24089−24098.
(19) Feng, W.; Ribeiro, R. C.; Wagner, R. L.; Nguyen, H.; Apriletti, J. W.; Fletterick, R. J.; Baxter, J. D.; Kushner, P. J.; West, B. L. Hormone-dependent coactivator binding to a hydrophobic cleft on nuclear receptors. *Science* **1998**, *280*, 1747−1749.
(20) Clemons, M.; Danson, S.; Howell, A. Tamoxifen ("Nolvadex"): a review. *Cancer Treat Rev.* **2002**, *28*, 165−180.
(21) Ward, H. W. Anti-oestrogen therapy for breast cancer: a trial of tamoxifen at two dose levels. *Br. Med. J.* **1973**, *1*, 13−14.
(22) Strasser-Weippl, K.; Goss, P. E. Advances in adjuvant hormonal therapy for postmenopausal women. *J. Clin. Oncol.* **2005**, *23*, 1751−1759.
(23) Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer. *Cochrane Database Syst. Rev.*. 2008; CD000486.
(24) Fisher, B.; Dignam, J.; Bryant, J.; Wolmark, N. Five versus more than five years of tamoxifen for lymph node-negative breast cancer: updated findings from the National Surgical Adjuvant Breast and Bowel Project B-14 randomized trial. *J. Natl. Cancer Inst.* **2001**, *93*, 684−690.
(25) Smith, M. R. Selective estrogen receptor modulators to prevent treatment-related osteoporosis. *Rev. Urol.* **2005**, *7* (Suppl 3), S30−5.
(26) Sengupta, S.; Jordan, V. C. Selective estrogen modulators as an anticancer tool: mechanisms of efficiency and resistance. *Adv. Exp. Med. Biol.* **2008**, *630*, 206−219.
(27) Johnston, S. R. Acquired tamoxifen resistance in human breast cancer–potential mechanisms and clinical implications. *Anticancer Drugs* **1997**, *8*, 911−930.
(28) Moen, M. D.; Keating, G. M. Raloxifene: a review of its use in the prevention of invasive breast cancer. *Drugs* **2008**, *68*, 2059−2083.
(29) Johnston, S. R. Endocrine manipulation in advanced breast cancer: recent advances with SERM therapies. *Clin. Cancer Res.* **2001**, *7* (12 Suppl), 4376s−4387s discussion 4411s-4412s..
(30) Howell, A. Pure oestrogen antagonists for the treatment of advanced breast cancer. *Endocr.-Relat. Cancer* **2006**, *13*, 689−706.
(31) Carlson, R. W. The history and mechanism of action of fulvestrant. *Clin. Breast Cancer* **2005**, *6* (Suppl1), S5−8.
(32) Gradishar, W. J.; Sahmoudv, T. Current and future perspectives on fulvestrant. *Clin. Breast Cancer* **2005**, *6* (Suppl 1), S23−9.
(33) Pike, A. C.; Brzozowski, A. M.; Walton, J.; Hubbard, R. E.; Thorsell, A. G.; Li, Y. L.; Gustafsson, J. A.; Carlquist, M. Structural insights into the mode of action of a pure antiestrogen. *Structure* **2001**, *9*, 145−153.
(34) Schultz-Norton, J. R.; Ziegler, Y. S.; Likhite, V. S.; Yates, J. R.; Nardulli, A. M. Isolation of novel coregulatory protein networks associated with DNA-bound estrogen receptor alpha. *BMC Mol. Biol.* **2008**, *9*, 97.
(35) Cheng, P. C.; Chang, H. K.; Chen, S. H. Quantitative nanoproteomics for protein complexes (QNanoPX) related to estrogen transcriptional action. *Mol. Cell. Proteomics.* **2010**, *9*, 209−224.
(36) Ambrosino, C.; Tarallo, R.; Bamundo, A.; Cuomo, D.; Franci, G.; Nassa, G.; Paris, O.; Ravo, M.; Giovane, A.; Zambrano, N.; Lepikhova, T.; Jänne, O. A.; Baumann, M.; Nyman, T. A.; Cicatiello, L.; Weisz, A. Identification of a hormone-regulated dynamic nuclear actin network associated with estrogen receptor alpha in human breast cancer cell nuclei. *Mol. Cell. Proteomics* **2010**, *9*, 1352−1367.
(37) Tarallo, R.; Bamundo, A.; Nassa, G.; Nola, E.; Paris, O.; Ambrosino, C.; Facchiano, A.; Baumann, M.; Nyman, T. A.; Weisz, A. Identification of proteins associated with ligand-activated estrogen receptor $\alpha$ in human breast cancer cell nuclei by tandem affinity purification and nano LC-MS/MS. *Proteomics* **2011**, *11*, 172−179.
(38) Addeo, R.; Altucci, L.; Battista, T.; Bonapace, I. M.; Cancemi, M.; Cicatiello, L.; Germano, D.; Pacilio, C.; Salzano, S.; Bresciani, F.; Weisz, A. Stimulation of human breast cancer MCF-7 cells with estrogen prevents cell cycle arrest by HMG-CoA reductase inhibitors. *Biochem. Biophys. Res. Commun.* **1996**, *220*, 864−870.
(39) Nassa, G.; Tarallo, R.; Ambrosino, C.; Bamundo, A.; Ferraro, L.; Paris, O.; Ravo, M.; Guzzi, P. H.; Cannataro, M.; Baumann, M.; Nyman, T. A.; Nola, E.; Weisz, A. A large set of estrogen receptor $\beta$-interacting proteins identified by tandem affinity purification in hormone-responsive human breast cancer cell nuclei. *Proteomics* **2011**, *11*, 159−165.
(40) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68*, 850−858.
(41) Nyman, T. A.; Rosengren, A.; Syyrakki, S.; Pellinen, T. P.; Rautajoki, K.; Lahesmaa, R. A proteome database of human primary T helper cells. *Electrophoresis* **2001**, *22*, 4375−4382.
(42) Sun, H.; Fang, H.; Chen, T.; Perkings, R.; Tong, W. GOFFA: Gene Ontology For Functional Analysis - Software for gene ontology-based functional analysis of genomic and proteomic data. *BMC Bioinformatics* **2006**, *7*, S23.
(43) Grober, O. M.; Mutarelli, M.; Giurato, G.; Ravo, M.; Cicatiello, L.; De Filippo, M. R.; Ferraro, L.; Nassa, G.; Papa, M. F.; Paris, O.; Tarallo, R.; Luo, S.; Schroth, G. P.; Benes, V.; Weisz, A. Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. *BMC Genomics* **2011**, *12*, 36.
(44) Paris, O.; Ferraro, L.; Grober, O. M.; Ravo, M.; De Filippo, M. R.; Giurato, G.; Nassa, G.; Tarallo, R.; Cantarella, C.; Rizzo, F.; Di Benedetto, A.; Mottolese, M.; Benes, V.; Ambrosino, C.; Nola, E.; Weisz, A. Direct regulation of microRNA biogenesis and expression by estrogen receptor beta in hormone-responsive breast cancer. *Oncogene.* **2012**, [Epub ahead of print].
(45) Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **1999**, *17*, 1030−1032.
(46) Gavin, A. C.; Bösche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J. M.; Michon, A. M.; Cruciat, C. M.; Remor, M.; Höfert, C.; Schelder, M.; Brajenovic, M.; Ruffner, H.; Merino, A.; Klein, K.; Hudak, M.; Dickson, D.; Rudi, T.; Gnau, V.; Bauch, A.; Bastuck, S.; Huhse, B.; Leutwein, C.; Heurtier, M. A.; Copley, R. R.; Edelmann, A.; Querfurth, E.; Rybin, V.; Drewes, G.; Raida, M.; Bouwmeester, T.; Bork, P.; Seraphin, B.; Kuster, B.; Neubauer, G.; Superti-Furga, G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*, 141−147.
(47) Nassa, G.; Tarallo, R.; Guzzi, P. H.; Ferraro, L.; Cirillo, F.; Ravo, M.; Nola, E.; Baumann, M.; Nyman, T. A.; Cannataro, M.; Ambrosino, C.; Weisz, A. Comparative analysis of nuclear estrogen receptor alpha and beta interactomes in breast cancer cells. *Mol. Biosyst.* **2011**, *7*, 667−676.
(48) Dauvois, S.; Danielian, P. S.; White, R.; Parker, M. G. Antiestrogen ICI 164,384 reduces cellular estrogen receptor content by increasing its turnover. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 4037−4041.

(49) Dauvois, S.; White, R.; Parker, M. G. The antiestrogen ICI 182780 disrupts estrogen receptor nucleocytoplasmic shuttling. *J. Cell Sci.* **1993**, *106*, 1377−1388.
(50) Kocanova, S.; Mazaheri, M.; Caze-Subra, S.; Bystricky, K. Ligands specify estrogen receptor alpha nuclear localization and degradation. *BMC Cell Biol.* **2010**, *11*, 98.
(51) Cicatiello, L.; Scafoglio, C.; Altucci, L.; Cancemi, M.; Natoli, G.; Facchiano, A.; Iazzetti, G.; Calogero, R.; Biglia, N.; De Bortoli, M.; Sfiligoi, C.; Sismondi, P.; Bresciani, F.; Weisz, A. A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. *J. Mol. Endocrinol.* **2004**, *32*, 719−775.
(52) Scafoglio, C.; Ambrosino, C.; Cicatiello, L.; Altucci, L.; Ardovino, M.; Bontempo, P.; Medici, N.; Molinari, A. M.; Nebbioso, A.; Facchiano, A.; Calogero, R. A.; Elkon, R.; Menini, N.; Ponzone, R.; Biglia, N.; Sismondi, P.; De Bortoli, M.; Weisz, A. Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cells. *J. Cell Biochem.* **2006**, *98*, 1163−1184.
(53) Métivier, R.; Penot, G.; Hübner, M. R.; Reid, G.; Brand, H.; Kos, M.; Gannon, F. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell* **2003**, *115*, 751−763.
(54) Kim, J. E.; Chen, J.; Lou, Z. DBC1 is a negative regulator of SIRT1. *Nature* **2008**, *451*, 583−586.
(55) Ford, J.; Jiang, M.; Milner, J. Cancer-specific functions of SIRT1 enable human epithelial cancer cell growth and survival. *Cancer Res.* **2005**, *65*, 10457−10463.
(56) Ota, H.; Tokunaga, E.; Chang, K.; Hikasa, M.; Iijima, K.; Eto, M.; Kozaki, K.; Akishita, M.; Ouchi, Y.; Kaneki, M. Sirt1 inhibitor, Sirtinol, induces senescence-like growth arrest with attenuated Ras-MAPK signaling in human cancer cells. *Oncogene* **2006**, *25*, 176−185.
(57) Pruitt, K.; Zinn, R. L.; Ohm, J. E.; McGarvey, K. M.; Kang, S. H.; Watkins, D. N.; Herman, J. G.; Baylin, S. B. Inhibition of SIRT1 reactivates silenced cancer genes without loss of promoter DNA hypermethylation. *PLoS Genet.* **2006**, *2*, e40.
(58) Lim, C. S. SIRT1: tumor promoter or tumor suppressor? *Med. Hypotheses* **2006**, *67*, 341−344.
(59) Yeung, F.; Hoberg, J. E.; Ramsey, C. S.; Keller, M. D.; Jones, D. R.; Frye, R. A.; Mayo, M. W. Modulation of NF-kappaB-dependent transcription and cell survival by the SIRT1 deacetylase. *EMBO J.* **2004**, *23*, 2369−2380.
(60) Kuzmichev, A.; Margueron, R.; Vaquero, A.; Preissner, T. S.; Scher, M.; Kirmizis, A.; Ouyang, X.; Brockdorff, N.; Abate-Shen, C.; Farnham, P.; Reinberg, D. Composition and histone substrates of polycomb repressive group complexes change during cellular differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 1859−1864.
(61) Chen, W. Y.; Wang, D. H.; Yen, R. C.; Luo, J.; Gu, W.; Baylin, S. B. Tumor suppressor HIC1 directly regulates SIRT1 to modulate p53-dependent DNA-damage responses. *Cell* **2005**, *123*, 437−448.
(62) Chueh, F. Y.; Leong, K. F.; Cronk, R. J.; Venkitachalam, S.; Pabich, S.; Yu, C. L. Nuclear localization of pyruvate dehydrogenase complex-E2 (PDC-E2), a mitochondrial enzyme, and its role in signal transducer and activator of transcription 5 (STAT5)-dependent gene transcription. *Cell. Signalling* **2011**, *23*, 1170−1178.
(63) Hu, Q.; Kwon, Y. S.; Nunez, E.; Cardamone, M. D.; Hutt, K. R.; Ohgi, K. A.; Garcia-Bassets, I.; Rose, D. W.; Glass, C. K.; Rosenfeld, M. G.; Fu, X. D. Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 19199−19204.
(64) Planas-Silva, M. D.; Waltz, P. K. Estrogen promotes reversible epithelial-to-mesenchymal-like transition and collective motility in MCF-7 breast cancer cells. *J. Steroid Biochem. Mol. Biol.* **2007**, *104*, 11−21.
(65) Steger, D. J.; Lefterova, M. I.; Ying, L.; Stonestrom, A. J.; Schupp, M.; Zhuo, D.; Vakoc, A. L.; Kim, J. E.; Chen, J.; Lazar, M. A.; Blobel, G. A.; Vakoc, C. R. DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells. *Mol. Cell. Biol.* **2008**, *28*, 2825−2839.
(66) Bernt, K. M.; Armstrong, S. A. A role for DOT1L in MLL-rearranged leukemias. *Epigenomics* **2011**, *3*, 667−670.
(67) Chang, M. J.; Wu, H.; Achille, N. J.; Reisenauer, M. R.; Chou, C. W.; Zeleznik-Le, N. J.; Hemenway, C. S.; Zhang, W. Histone H3 lysine 79 methyltransferase Dot1 is required for immortalization by MLL oncogenes. *Cancer Res.* **2010**, *70*, 10234−10242.
(68) Daigle, S. R.; Olhava, E. J.; Therkelsen, C. A.; Majer, C. R.; Sneeringer, C. J.; Song, J.; Johnston, L. D.; Scott, M. P.; Smith, J. J.; Xiao, Y.; Jin, L.; Kuntz, K. W.; Chesworth, R.; Moyer, M. P.; Bernt, K. M.; Tseng, J. C.; Kung, A. L.; Armstrong, S. A.; Copeland, R. A.; Richon, V. M.; Pollock, R. M. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell.* **2011**, *20*, 53−56.

HORM CANC
DOI 10.1007/s12672-012-0102-1

# Effects of Oestrogen on MicroRNA Expression in Hormone-Responsive Breast Cancer Cells

**Lorenzo Ferraro · Maria Ravo · Giovanni Nassa · Roberta Tarallo · Maria Rosaria De Filippo · Giorgio Giurato · Francesca Cirillo · Claudia Stellato · Silvana Silvestro · Concita Cantarella · Francesca Rizzo · Daniela Cimino · Olivier Friard · Nicoletta Biglia · Michele De Bortoli · Luigi Cicatiello · Ernesto Nola · Alessandro Weisz**



**Abstract** Oestrogen receptor alpha (ERα) is a ligand-dependent transcription factor that mediates oestrogen effects in hormone-responsive cells. Following oestrogenic activation, ERα directly regulates the transcription of target genes via DNA binding. MicroRNAs (miRNAs) represent a class of small noncoding RNAs that function as negative regulators of protein-coding gene expression. They are found aberrantly expressed or mutated in cancer, suggesting their crucial role as either oncogenes or tumour suppressor genes. Here, we analysed changes in miRNA expression in response to oestrogen in hormone-responsive breast cancer MCF-7 and ZR-75.1 cells by microarray-mediated expression profiling. This led to the identification of 172 miRNAs up- or down-regulated by ERα in response to 17β-oestradiol, of which 52 are similarly regulated by the hormone in the two cell models investigated. To identify mechanisms by which ERα exerts its effects on oestrogen-responsive miRNA genes, the oestrogen-dependent miRNA

**Electronic supplementary material** The online version of this article (doi:10.1007/s12672-012-0102-1) contains supplementary material, which is available to authorized users.

L. Ferraro · M. Ravo · G. Nassa · R. Tarallo · G. Giurato · C. Stellato · S. Silvestro · C. Cantarella · F. Rizzo · A. Weisz (✉)
Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno,
via S. Allende 1,
84081 Baronissi, Salerno, Italy
e-mail: aweisz@unisa.it

L. Ferraro · F. Cirillo · L. Cicatiello · E. Nola · A. Weisz
Department of General Pathology, Second University of Naples,
vico L. De Crecchio 7,
80138 Napoli, Italy

M. R. De Filippo
Fondazione IRCCS SDN,
via Gianturco 113,
80143 Naples, Italy

D. Cimino
Molecular Biotechnology Center and Department of Oncological Sciences, University of Turin,
via Nizza 52,
10126 Turin, Italy

O. Friard · M. De Bortoli
Center for Molecular Systems Biology, University of Turin,
via Accademia Albertina 13,
10123 Turin, Italy

N. Biglia
Department of Obstetrics and Gynecology, Mauriziano 'Umberto I' Hospital, University of Turin,
Largo Turati 62,
10128 Turin, Italy

A. Weisz
Division of Molecular Pathology and Medical Genomics, 'SS. Giovanni di Dio e Ruggi d'Aragona' Hospital,
University of Salerno,
via San Leonardo,
84131 Salerno, Italy

Published online: 25 January 2012

expression profiles were integrated with global in vivo ERα binding site mapping in the genome by ChIP-Seq. In addition, data from miRNA and messenger RNA (mRNA) expression profiles obtained under identical experimental conditions were compared to identify relevant miRNA target transcripts. Results show that miRNAs modulated by ERα represent a novel genomic pathway to impact oestrogen-dependent processes that affect hormone-responsive breast cancer cell behaviour. MiRNome analysis in tumour tissues from breast cancer patients confirmed a strong association between expression of these small RNAs and clinical outcome of the disease, although this appears to involve only marginally the oestrogen-regulated miRNAs identified in this study.

**Keywords** Oestrogen receptor · Breast cancer · MicroRNA · Cell cycle · Gene expression

## Introduction

The steroid hormone 17β-oestradiol (E2) is a key regulator of growth and differentiation in the mammary gland [1, 2] where it is involved in the pathogenesis and clinical outcome of breast cancer (BC) [3]. In normal and transformed mammary epithelial cells, the biological effects of E2 are mediated primarily by oestrogen receptor alpha (ERα), a ligand-inducible transcription factor of the nuclear receptor gene superfamily. Following oestrogenic activation, ERα mediates transcription by interacting directly with specific oestrogen response elements (EREs) located in the promoter/enhancer region of its target genes; it can also interact with other transcription factor complexes like Fos/Jun [4] or SP-1 [5] influencing transcription of genes whose promoters do not harbour ERE (tethering). This leads to transcriptional activation or repression of target genes involved in important cellular function such as cell cycle control, differentiation and apoptosis [6–8]. Alternatively, oestrogens are able to trigger rapid and transient cellular responses via ERα crosstalk with different signal transduction pathways in the cytoplasm [9, 10]. The cellular response to oestrogens involves multiple biological events, including transcription, RNA stability and post-translational modifications [11]. MicroRNAs (miRNAs) are a class of small RNAs of 23 nucleotides (nt) in length, which coordinate a broad range of gene expression programs mainly through modulation of gene regulation [12]. There are over 1,700 identified miRNAs in the human genome that are, themselves, subject to regulation at both transcriptional and post-transcriptional level. MiRNAs are encoded in several regions of the genome, both in protein coding and non-coding transcription units. Approximately 50% of miRNAs are derived from non-coding RNA transcripts and have their own promoters, while an additional 40% are located within the introns of protein coding genes and share the same transcriptional control of the host genes [13, 14]. Moreover, many miRNAs are encoded in the genome as clusters that can range from 2 to 19 miRNA hairpins, encoded in tandem and in close proximity to each other [15]. After being transcribed, miRNA carrier transcripts (pri-miRNA [16]) undergo a step-wise processing: The long miRNA transcript is cleaved into the nucleus by Drosha into pre-miRNA [17], exported into the cytoplasm and there cleaved by Dicer into a miRNA–miRNA* duplex [18]; mature miRNAs are loaded into microRNA-induced silencing complex (miRISC), which interferes with the transfer of transcriptome information into proteome output via RNA-induced gene silencing [19–21]. Regardless of the mechanisms, each miRNA can potentially regulate gene expression of hundreds of genes, and on the other hand, a single transcript can be targeted by multiple miRNAs [22, 23]. In fact, almost one third of the protein-coding genes are under the regulation of miRNAs, and, as a consequence, many miRNAs seem to play a crucial role in different biological processes such as differentiation, proliferation and cell death in a context-dependent way [24]. Not surprisingly, aberrant miRNA expression is a hallmark of several diseases, including cancer [25]. Several studies have established a role of miRNAs in the pathogenesis of BC, showing a link between E2/ERs and microRNAs expression either in BC cell lines or in cancerous breast tissues [26–30]. These studies indicate that miRNAs can act as either oncogenes or oncosuppressors [31]. In addition, identification of differential expression profiles of miRNAs between normal and neoplastic breast tissue or among human BC subgroups confirms the hypothesis of a possible involvement of miRNAs in tumour development and progression [32–35]. The expression of miRNAs has been examined in different BC cell lines and biopsies. Among these, the most consistently deregulated miRNAs following E2 treatment were miR-206, miR-125a/b, miR-17-5p, miR-34a and some member of let-7 family that may act as tumour suppressor genes [36]. Some recent reports, in fact, have shown that E2 treatment leads to alteration of miR-206 whose expression levels decrease in ERα-positive human BC tissues. This miRNA is able to suppress ESR1 expression and to inhibit growth of MCF-7 BC cells [37]. Moreover, miR-17-5p represses the translation of AIB1 mRNA, thereby inhibiting the function of E2F1 and ERα. Down-regulation of AIB1 by miR-17-5p results in the suppression of oestrogen-stimulated proliferation and oestrogen/ER-independent BC cell proliferation [38, 39]. MiRNA let-7, one of the first discovered members of let-7 family, is poorly expressed or deleted in many human cancers as well as miR-34a, which has been shown to be transcriptionally regulated also by p53 [40]. On the other hand, miR-21, miR-155 and miR-10b may act as oncogenes, being consistently found

over-expressed in cancer. Consistent with these findings, miR-21 was also shown to be highly up-regulated in breast tumours compared to the normal breast tissues, suggesting its oncogenic role [41]. Moreover, miR-21 together with let-7 and miR-98 are involved in a negative-regulatory loop that controls c-Myc, E2F1 and E2F2 protein levels [28].

A number of genes involved in BC progression have been identified by in silico analyses and then experimentally proven to be targets of miRNAs that are deregulated in breast tumours [42]. Furthermore, it has been shown that miRNA deregulation in BC can occur not only at the transcriptional level but also at the processing level. In fact, it was recently reported that E2 is able to up-regulate Dicer1 gene expression in ERα-positive BC cells [28]. Furthermore, some miRNAs, including miR-221/222 and miR-29a (highly expressed in ERα negative BC cells), directly repress ERα and Dicer1 expression; in contrast, miR-200c (highly expressed in ERα-positive BC cells) increases Dicer1 levels [43].

In this study, we performed a time-course analysis of oestrogen-regulated miRNAs in MCF-7 and ZR-75.1 cell lines, with the aim to identify all miRNAs showing identical kinetics and type of response to the hormone in this BC cell model. In order to identify relevant miRNA target transcripts, we then performed a functional analysis of the E2-regulated miRNAs by integrating data from both miRNA and mRNA expression profiles obtained under identical experimental conditions. To determine the mechanisms by which ERα exerts its effects on target miRNAs, data derived from global analyses of ERα in vivo binding sites to the genome upon E2 stimulation in MCF-7 cells were integrated with hormone-responsive miRNome data. Finally, to investigate the role of miRNA expression in primary breast tumours, we analysed in silico miRNomes from tumour specimens from patients with divergent clinical outcomes following surgical treatment to evaluate existing correlations between the expression patterns of oestrogen-responsive miRNAs identified in cellular models of the disease and clinical–pathological parameters of BC.

## Materials and Methods

### Cell Culture and Imunoblotting

Human hormone-responsive BC cells MCF-7 Tet-Off (Clontech-Takara, Saint-Germain-en-Laye, France) and ZR-75.1 (ATCC CRL-1500) were grown in Dulbecco's modified Eagle's medium (Sigma-Aldrich, Milan, Italy) supplemented with 10% foetal bovine serum (HyClone, Cramlington, UK) and antibiotics: 100 U/ml penicillin, 100 mg/ml streptomycin, 250 ng/ml Amfotericin-B and 50 μg/ml G418. Cells were routinely tested for mycoplasma infection using MycoAlert mycoplasma detection kit (Cambrex BioScience, Rockland, ME, USA). For G0–G1 synchronisation, cells were plated at 20–40% confluence in steroid-free medium (phenol red-free Dulbecco's modified Eagle's medium with 5% foetal bovine serum, pre-treated with dextran-coated charcoal and antibiotics) and maintained for 4 days with replacement of the same fresh medium before stimulation with $10^{-8}$ M 17β-estradiol (+E2) or EtOH (vehicle) as negative control. ER expression in cell lines was assayed by sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis and immunoblotting of total protein extracts, using rabbit polyclonal anti-ERα (sc-543, Santa Cruz Biotechnology, Heidelberg, Germany) as previously described [44].

### Patients

Thirty-two frozen tumour specimens were selected from a former cohort [45]. They were obtained from patients who underwent primary surgical treatment between 1988 and 2001 at a median age of 57 (range, 32–79). Twenty-two cases were oestrogen receptor (ER) positive and treated in the adjuvant setting with 20 mg tamoxifen daily for 5 years alone or in combination, while 10 cases were ER negative. The average follow-up was of 89 months. ER and PgR status was determined by immuno-histochemical stainings; samples were defined positives when tumours contained more than 10% positive cells. Details are provided as Supplementary Table S3.

### RNA Purification

Total RNA was extracted from control (+EtOH, −E2) and from hormone-stimulated (+E2) cell cultures with Trizol (Invitrogen, Carlsbad, CA, USA), as described previously [46]. In each case, cells were collected from multiple parallel cultures and pooled before RNA extraction as described before [47]. For tumour samples, after surgical removal, total RNA was isolated with Concert Cytoplasmic RNA Reagent (Invitrogen Life Technologies, Carlsbad, CA, USA) from 20 to 50 mg tumour tissues, according to the manufacturer's guidelines. Frozen tumours were placed in this reagent and homogenised using a ball mill (MM200, Retsch, Düsseldorf, Germany). The suspension was centrifuged at 14,000×*g* for 5 min at 4°C, then lysed with 0.1 ml of 10% SDS followed by 0.3 ml of 5M sodium chloride and 0.2 ml of chloroform for milliliters of reagent. The lysate was centrifuged at 14,000×*g* for 15 min at 4°C, and the upper aqueous phase was removed and combined with 0.8 volume of isopropyl alcohol for 10 min at room temperature. The RNA was recovered by centrifugation, washed with 75% ethanol and finally dissolved in RNase-free water. Before use, RNA concentration in each sample was

determined with a ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA) and quality assessed with Agilent 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA, USA).

### Microarray Analyses

For miRNA expression profiling, technical replicates were produced. For MCF-7 cells, 800 ng of total RNA were fluorescently labelled and amplified in triplicate to be then pooled for the hybridisation; for ZR-75.1 cells and breast tumour samples, the same concentration of RNA were fluorescently labelled, amplified and hybridised at least in duplicate. Hybridisation reactions were performed with Illumina v2 MicroRNA Expression BeadChips, according to the protocols provided by the array manufacturer (Illumina Inc., San Diego, CA, USA). The human microRNA panel used comprises 1,145 probes designed on miRNA sequence present in miRBase database (Release 12.0) and on additional novel content derived using Illumina sequencing technology. For mRNA expression profiling, 500 ng total RNA from MCF-7 and ZR-75.1 were reverse transcribed, as described previously [48, 49], used for synthesis of complementary DNA and biotinylated complementary RNA (cRNA), according to the Illumina TotalPrep RNA amplification kit (Ambion, Austin, TX, USA; category number IL1791) protocol. For each sample, 700 ng of cRNA were hybridised for 18 h at 55°C on Illumina HumanWG-6 version 2.0 BeadChips containing 48701 probes (Illumina Inc., San Diego, CA, USA), according to the manufacturer's protocol. The BeadChips were scanned using Illumina BeadArray Reader 500 according to the manufacturer's standard methods.

### MicroRNA Expression Profiling Data Analysis

For data analysis, the fluorescence intensity files were loaded into the Illumina GenomeStudio v2009.1 software for quality control and expression analysis. First, the quantile normalisation algorithm was applied on the raw datasets to correct systematic errors. This normalisation equalises distribution, median and mean of probe intensities among all samples, as the normalised distribution is chosen by averaging each quantile across samples. For differential expression analysis, technical replicates were grouped together, and miRNAs with a detection $p<0.01$, corresponding to a false positive rate of 1%, were considered as expressed. Statistical significance was calculated with the Illumina DiffScore, a proprietary algorithm that uses the bead standard deviation to build an error model. Only miRNAs with a DiffScore $\leq -20$ and $\geq 20$, corresponding to a $p=0.01$, were considered as statistical significant. The microarray data were deposited in the Array Express repository database with the following accession numbers: E-TABM-1194.

### Breast Tumour Data Analysis

Raw data were normalised as described above using GenomeStudio v. 2009.1, and only probes with $p<0.05$ were selected for further analysis. A total of 1,021 probes had at least one valid call in a tumour sample. Two tumour samples had <10 valid data and were not included. Furthermore, to avoid unbalancing, only probes showing valid calls in at least 20 tumour samples were used in clustering and differential expression analysis (739 probes). Data were normalised to the median value in all tumour samples ($N=30$) and converted to log (log ratio). Unsupervised hierarchical clustering was performed using TMev package (MultiExperiment Viewer, at www.t4m.org/mev) using the HCL with Pearson's correlation with complete or average linkage, depending on sample/gene number. Differential expression analysis was carried out using the SAM routine in the same package, using 5,000 permutations and Tusher's S0 [50]. Clusters defined by the HLC were then compared in terms of overall survival (OS) or disease-free survival (DFS) using Kaplan–Meier log rank correlation. Survival curves were generated with PASW 18.0 statistical software.

### Quantitative Real-time RT-PCR

Total RNA was extracted from MCF-7 cells before and after stimulation for 72 h with $10^{-8}$ M E2 as described above. Mature miRNAs were reverse transcribed using a miRNA-specific stem loops and reverse transcriptase before real-time PCR performed using Taqman microRNA assays (assay ID: 2439, 2445, 2441, 2126, 2174, 2440, 2333; Applied Biosystems), as described by Grober et al. [51]. RNU49, unaffected by hormone treatment (data not shown), was used as an internal control to normalise all data using the Taqman RNU49 assay (Applied Biosystems). All real-time PCR runs were performed on a MJ Research PTC-200 Opticon Instrument.

### ChIP-Seq Data Analysis

For determining the ERα genome-wide binding sites, we have re-analysed the ChIP-Seq data (accession number E-MTAB-131) previously published by Cicatiello et al. [48]. Enriched regions from MCF-7 cells stimulated with 17β-estradiol for 45min (+E2) were compared with the same from MCF-7 cultured maintained in steroid-free medium (−E2), with FindPeaks [52], setting the value of subpeaks parameter to 0.5. For selecting only the most relevant sites, we have applied the first quartile as statistical cut-off. For evaluating the nearest ERα binding sites to the miRNA genes within 50 kb, we have used windowBed of the suite programs BEDTools [53].

### MicroRNA Target Prediction and Functional Analysis of Their Predicted MRNA Targets

For comprehensive prediction of miRNA target genes we used TargetScan, release 5.1 (www.targetscan.org). To identify statistically over-represented 'biological process' Gene Ontology terms among sets of selected mRNA targets, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov) functional annotation tool [54, 55]. To this aim we used as background data coming from gene expression profiling experiments previously performed [48] on the same cell lines and under the same experimental conditions investigated in this study.

### MiRNA Localisation in Host Genes

We annotated the genomic position and context of microRNAs in human genome. This analysis was performed with RegionMiner [56] application of the Genomatix software suite, which generates statistic annotation and data for chromosomal regions.

## Results

### Characterisation of ERα Positive Breast Cancer Cell Lines

The two ERα-positive breast carcinoma cell lines MCF-7 and ZR-75.1 were used as in vitro model of hormone-responsive BC; ERα expression was monitored by Western blot analysis, using as control extracts from ERα negative SKBR3 cells (Fig. 1a). In these cell lines, E2 deprivation induces G1 arrest [57], a quiescence status readily overrun by administration of physiological concentrations of E2. This determines resumption of cell cycle progression, mediated by regulation of cell cycle control pathways [58]. Indeed, analysis of hormone-deprived MCF-7 and ZR-75.1 cell cultures before and after oestrogen (E2 $10^{-8}$ M) stimulation show the timed accumulation of cyclins that characterise cell cycle progression (Fig. 1b). In MCF-7 cells (upper panel), oestrogen-induced accumulation of the G1 cyclin D1 is detectable already after 4 h of stimulation, remaining high throughout the pre-replicative phase, and to a lower extent also during the S phase. Cyclin E2 concentration is also affected by hormone stimulation, showing a 1.5- to 2-fold induction after 10 h and lasting for up to 24 h. The cellular levels of the S-G2 phase cyclins A2 and B1 progressively accumulate in a latter time. The results were also confirmed in ZR-75.1 cells (lower panel of Fig. 1b), which show a good degree of oestrogen dependence and a similar pattern of cyclin gene activation occurring, however, faster than in MCF-7 cells. These molecular responses to E2, occurring in both cell lines, confirm the direct stimulatory action of oestrogen under the conditions used for this study.

### Identification of Oestrogen-Regulated MiRNAs

In order to study the effects of the oestrogen on miRNAs expression, we performed a time-course analysis following oestrogen in both BC cell lines selected for this study. To this aim, total RNA was extracted from MCF-7 and ZR-75.1 cells before and after different time points of stimulation

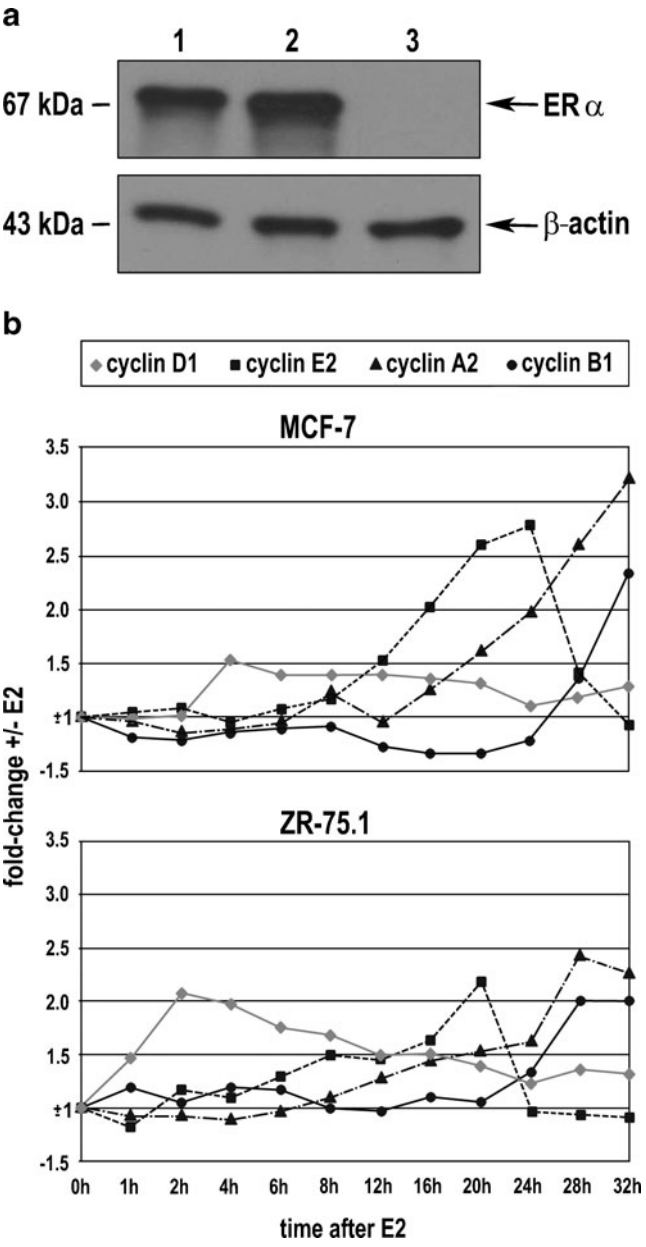



**Fig. 1** Characterization of ERα-positive breast cancer cell lines. **a** Expression level of ERα by Western blot analysis of protein extracts from MCF-7 (*lane 1*), ZR-75.1 (*lane 2*) and SKBR3 (*lane 3*) cells. **b** Monitoring of cyclin gene expression in G1-synchronised MCF-7 and ZR-75.1 cell lines following stimulation with a mitogenic dose of E2

with a mitogenic dose of E2 (6, 12, 24 and 72 h), and a global analysis of miRNA expression profiles was carried out by microarray hybridisation as described in 'Materials and Methods', using a platform (Illumina MicroRNA Expression Beadchip) detecting most known and characterised miRNAs. Results indicate that, in agreement with what previously reported in this cell type [28, 30, 43, 48, 59, 60], this hormone can indeed affect significantly the intracellular concentration of specific miRNAs. Among all known miRNAs, we identified 230 showing significant changes in expression in response to E2 ($p \leq 0.01$), including 110 that changed in both cell lines (52 responding similarly and 58 showing opposite changes), 51 restricted to MCF-7 cells and 69 restricted to ZR-75.1 cells. The 52 miRNAs showing concordant regulation in the two cell lines, 25 down- and 27 up-regulated (Fig. 2 central panel), were considered good candidates to investigate miRNA involvement in oestrogen signalling and for this reason selected for further analysis. In Fig. 3 are displayed in graphic format the actual values of the fluctuations detected for 14 miRNAs showing representative expression profiles. Reliability of the microarray platform used for this study was throughly controlled in our laboratory in a previous study [30]; however, a test was performed by real-time RT-PCR on seven miRNAs in MCF-7 cells 72 h after hormone (Supplementary Fig. S1). The results confirm a good correlation between Q-PCR and microarray measurements.

## Putative Targets of E2-Regulated miRNAs and Functional Analysis of their Predicted Targets

In order to evaluate the functional roles of the 52 miRNAs responsive to E2 treatment in both MCF-7 and ZR-75.1 cells, we performed an in silico functional analysis exploiting previously described mRNA expression profiling data obtained in the same cell lines under comparable experimental conditions [48]. To this aim, we first searched for mRNAs putative targets of these oestrogen-regulated miRNAs, results show that ~30% of all expressed mRNAs indeed represent potential targets of these small RNAs [48], and the same is true for 33% mRNAs found regulated by oestrogen in the same study. In order to identify biological processes likely to be influenced by ERα via miRNAs in our cell lines, a Gene Ontology analysis was performed by DAVID tool, using as background the list of expressed mRNAs identified previously [48]. The results reported in Supplementary Fig. S2 show how several cellular processes were found statistically enriched by ERα-responsive miRNAs, including those known to be affected by ERα, such as response to hormonal stimuli, regulation of transcription and cell proliferation, and other that represent key cellular processes in tumour cells, such as cell migration, adhesion and differentiation. Furthermore, starting from the assumption that miRNA up-regulation might result in down-regulation of its mRNA targets, and vice versa, we investigated the existence of dynamic inverse relationships between miRNA and mRNA levels in oestrogen-stimulated cells. To evaluate this possibility, we first searched for the presence of perfect or imperfect matches between the seed sequence of each regulated miRNA and the untranslated region (UTR) of their putative mRNA targets. Interestingly, we could find only perfect complementarity between these two sequences, a finding that strongly supports the possibility of the existence of a pathway controlling mature mRNA half-life in hormone-responsive BC cells, whereby E2-activated ERα exerts a post-transcriptional control on its target gene activity via specific miRNAs. Indeed, when considering all mRNAs targeted by a single miRNA, we observed a clear inverse relationship between changes in miRNA concentration and that of the corresponding target mRNAs in about 50% of the cases 125 mRNAs of the 252 identified putative targets (listed in Supplementary Table S1).

## E2-Regulated miRNAs Associated with ERα-Binding Sites or Located in the Intragenic Region of Oestrogen-Responsive Genes

Current understanding of microRNA biogenesis indicates that expression of these small RNAs can be modulated either during their transcription or through the multiple steps leading to their maturation. In order to identify the transcriptional mechanisms by which activated ERα exerts its effects on oestrogen-responsive miRNAs, we integrated two global genomic analyses of ERα in vivo binding sites [48, 61] with the miRNome expression profile obtained in this study. In this way, we could observe that some E2-responsive miRNAs indeed display ERα binding sites within 10 kb of the transcription unit (Table 1). In particular, six of these miRNA genes are associated with one or more ERα binding sites, suggesting a mechanism for direct regulation of RNA biogenesis exerted also in this case by chromatin-bound ER. However, recent findings suggest that the receptor is able to perform its action even when its binding site is not in close proximity of the regulated genes. In order to evaluate this possibility, we expanded our search to 50 kb around miRNA genes, leading to the identification of 18 more miRNA genes linked to one or more ERα-binding sites (Supplementary Table S2). As approximately 50% of miRNAs are transcribed from introns of protein-coding genes, while the others are intergenic [62], we searched for E-regulated protein-coding genes harbouring miRNA genes and aligned the results with those relative to E2-regulated miRNAs. The results are reported in Table 2 to show that six E2-regulated miRNAs are located within oestrogen-dependent genes. Interestingly, in most cases, the miRNA follows the same regulation trend of the harbouring gene.

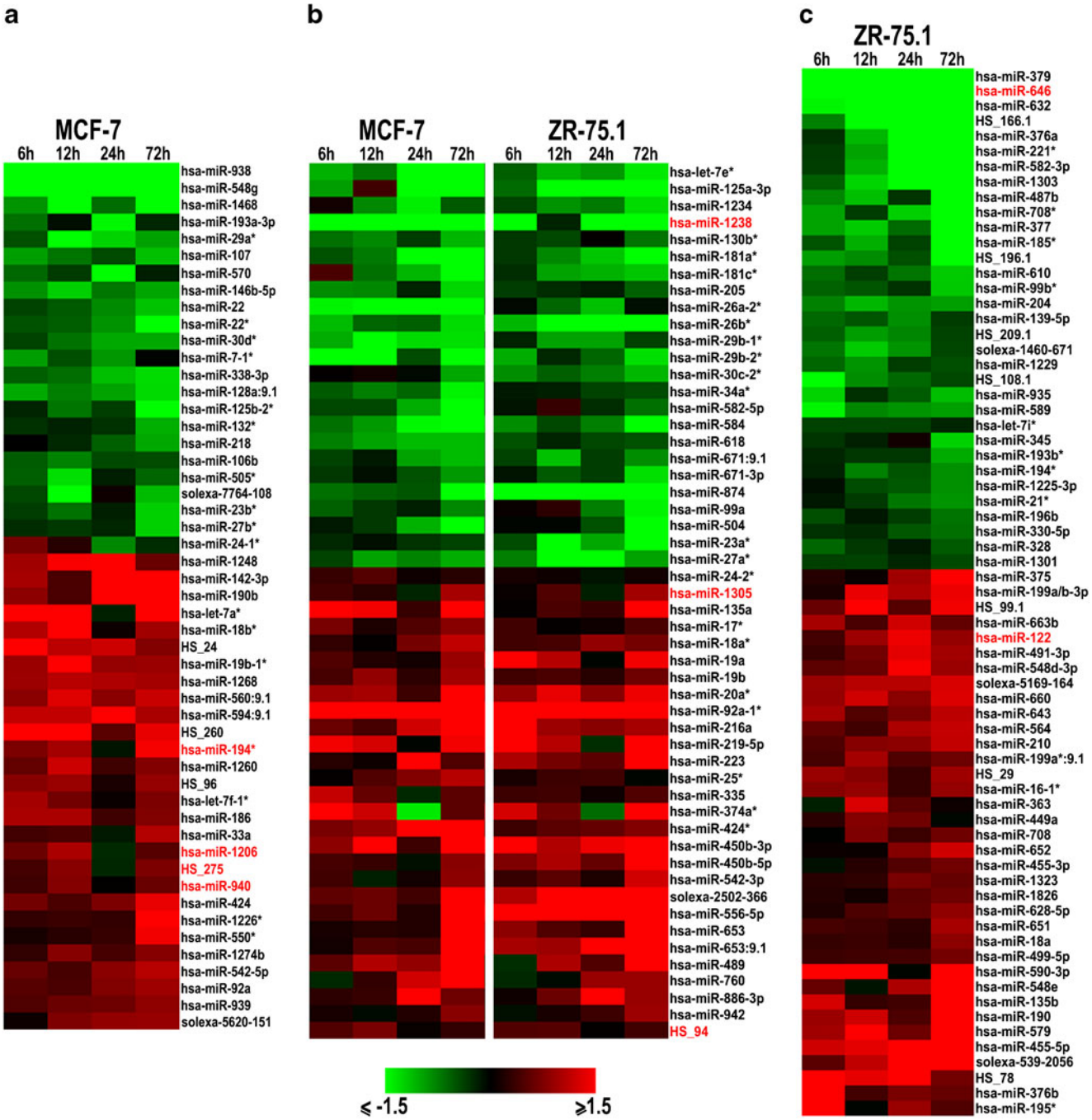



**Fig. 2** Effects of ERα on miRNome of human breast cancer cell lines. Time-course analysis of miRNA expression profiles in MCF-7 and ZR-75.1 cell lines after cells exposure to E2 for the indicated times. Oestrogen-regulated miRNAs are grouped as follows: regulated in MCF-7 cells only (**a**), in both cell lines (**b**) or in ZR-75.1 cells only (**c**). Data displayed represent the ratio between the fluorescence intensity values of each miRNA at the indicated time after exposure to $10^{-8}$ M E2 vs the corresponding 0 h time point. MicroRNAs marked in *red* represent those regulated in BC cell lines in vitro and those displaying differential expression between primary breast tumour subgroups (see Fig. 4)

## Regulation of miRNA Expression in Breast Tumour Samples

We next evaluated expression of the hormone-regulated miRNAs identified here in primary BC tissue samples. For this, we performed miRNA expression profiling in breast tumour samples as described in 'Materials and Methods'. All miRNA probes that showed valid calls in at least 20 tumour samples were used in the statistical analysis for correlation with clinic–pathological parameters. Of these 739 probes, 161 were regulated by oestrogen in at least one cell line, whereas 49 were

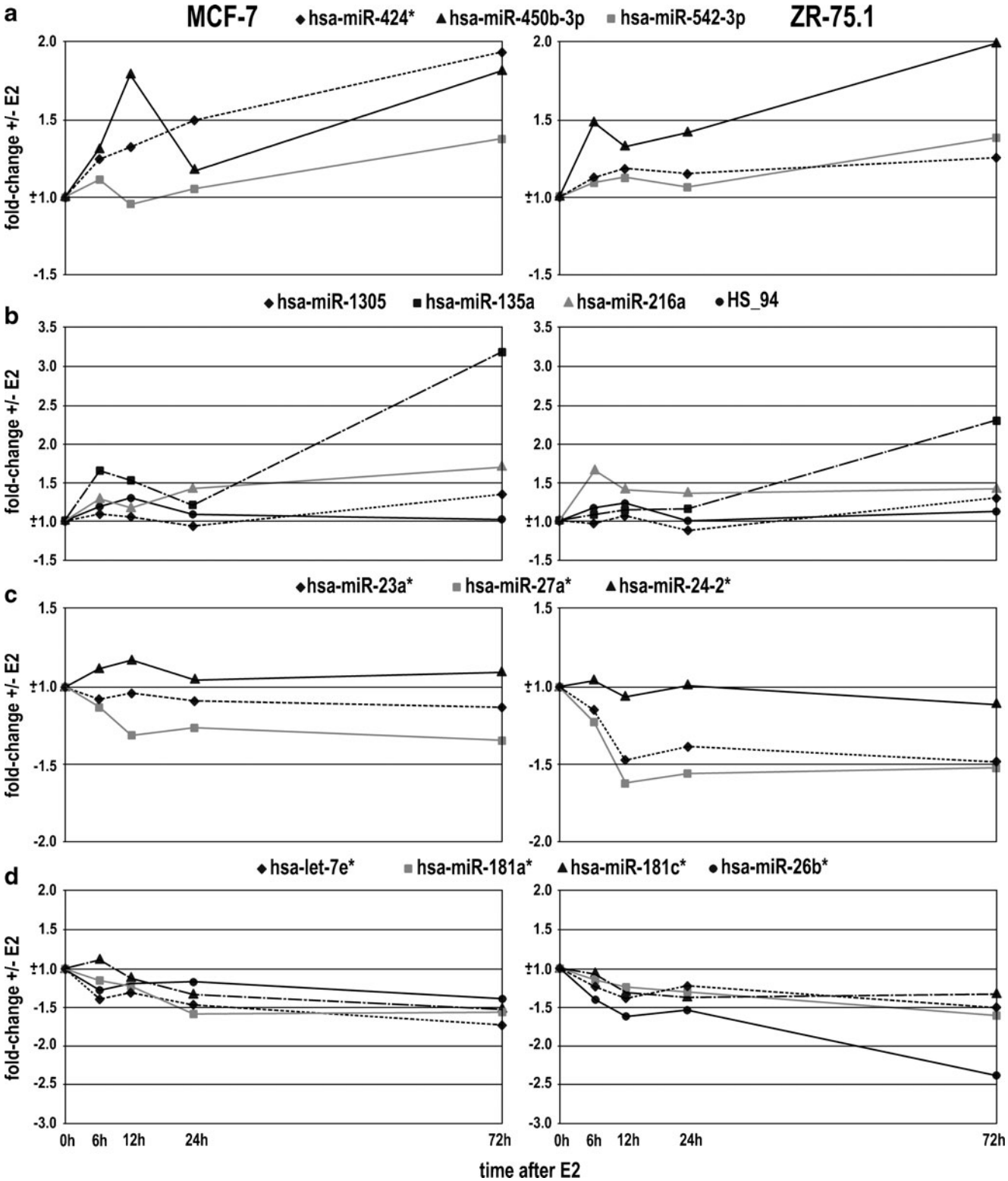



**Fig. 3** Graphic representation of timed fold-change variations of selected E2-regulated miRNAs in MCF-7 and ZR-75.1 cell lines. MicroRNAs shown are members of clusters (**a**, **c**) or are encoded as independent transcript (**b**, **d**)

regulated in both cell lines examined. First, we considered whether the expression profile of in vitro regulated miRNAs (161 probes) could discriminate between ER-positive and ER-negative breast tumours. However, both *t* test and SAM analysis failed to evidence any correlation. On the contrary, unsupervised hierarchical

**Table 1** ERα-binding sites within 10 kb from loci encoding E2-regulated miRNAs

| miRNA gene | Closest ERα binding site[a,b] | Closest ERα binding site[a,b] | Closest ERα binding site[a,c] | Closest ERα binding site[a,c] |
|---|---|---|---|---|
| hsa-miR-135a-2 | – | 4773 (chr12:96486593-7225) | −7265 (chr12:96474451-56) | 8546 (chr12:96490366-0460) |
| hsa-miR-181c | – | – | −3445 (chr19:13842950-3068) | 5623 (chr19:13852245-2362) |
| hsa-miR-23a | −6791 (chr19:13815263-6060) | 191 (chr19:13807531-8210) | −7212 (chr19:13815685-99) | 786 (chr19:13807588-7661) |
| hsa-miR-27a | −6933 (chr19:13815263-6060) | 44 (chr19:13807531-8210) | −7354 (chr19:13815685-99) | 639 (chr19:13807588-7661) |
| hsa-miR-24-2 | 0 (overlapping) (chr19: 13807531-8210) | 0 (overlapping) (chr19: 13807531-8210) | −7649 (chr19:13815685-99) | 486 (chr19:13807588-7661) |
| hsa-miR-26b | – | 2814 (chr2:218978503-9023) | – | – |

[a] Distance in bps

[b] ChIP-Seq data from Cicatiello et al. [48]

[c] ChIP-Seq data from Fullwood et al. [61]

clustering of these probes gave two well-separated branches (Fig. 4a) that correlated with disease-free survival (disease-free or relapsing; DFS) when all the patients were considered ($P<0.05$ in the case of DFS; Fig. 4b), while marginal significance was observed for overall survival (surviving and not surviving; OS). Limiting the analysis to the group of patients receiving Tamoxifen treatment did not increase separation of survival curves, whereas DFS retained marginal significance, this was lost for OS (Fig. 4c). A second kind of analysis was performed to see whether in vitro regulated probes were present among those differentially expressed in defined groups of samples. Again, SAM analysis failed to evidence differentially expressed probes in ER-positive versus ER-negative samples. On the contrary, groups defined by the DFS status and by the OS status were well differentiated. For DFS, we chose 47 probes differentially expressed (median false significant genes=7.84): of these, five were regulated in vitro in at least one cell line (hsa-miR-122, -194*, -1238, -1305 and HS_94). For OS, 63 probes were selected (median FSG06.3): Of these, nine were regulated in at least one cell line (hsa-miR-122, -194*, -646, -940, 1206, 1238, -1305, HS_94 and HS_275). MiRNAs belonging to these groups are shown in Supplementary Table S4.

## Discussion

MicroRNAs represent a class of small non-coding RNAs that control gene expression by targeting mRNAs and thereby triggering either translation repression or RNA degradation. Among human diseases, it has been shown that miRNAs are aberrantly expressed or mutated in cancer, suggesting that they may play a crucial role as a class of oncogenes or tumour suppressor genes. Multiple lines of evidence show the involvement of specific miRNAs in the pathogenesis of BC, where they may represent an alternative molecular mechanism that could impact the onset, development and progression of this hormone-responsive disease. In this study, we investigated the role of oestrogen and its nuclear receptor ERα in modulating miRNA expression in human BC cells. Once chosen robust and validated cellular models for our study, we performed a time-course profiling analysis to identify miRNAs whose levels are affected by E2 in both MCF-7 and ZR-75.1 cells. This led to the identification, among the 1,145 probes on the miRNA microarray used, of 172 E2-regulated miRNAs (15% of total). Of these small RNAs, 52 resulted commonly regulated in both cell lines, 51 were regulated only in MCF-7 and 69 regulated only in ZR-75.1 cells. These three sets, characterised by a defined kinetic of response to E2, cluster in two concordant groups of significantly down- or up-regulated miRNAs.

**Table 2** E2-regulated miRNAs in MCF-7 and ZR-75.1 cell lines located in the intragenic region of hormone-regulated genes

| miRNA gene | E2 effect on miRNA | Host gene[a] | E2 effect on host gene |
|---|---|---|---|
| hsa-miR-25 (chr7:99.529.199-202) | Up | MCM7 (chr7:99.528.340-99.537.363) | Up |
| hsa-miR-26a (chr12:56.504.649-742) | Down | CTDSP2 (chr12:56.499.977-56.527.014) | Down |
| hsa-miR-424 (chrX:133.508.310-407) | Up | MGC16121 (chrX:133.505.073-133.508.326) | Up |
| hsa-miR-618 (chr12:79.853.646-743) | Down | LIN7A (chr12:79.715.302-79.855.825) | Down |
| hsa-miR-760 (chr1:94.084.976-5055) | Up | BCAR3 (chr1:937.999.37-940.852.94) | Down |
| hsa-miR-942 (chr1:117.438.788-873) | Up | TTF2 (chr1:117.404.472-117.447.014) | Up |

[a] Oestrogen-dependent gene regulation data from Cicatiello et al. [48]

Interestingly, our microarray data reveal that the miRNA 'star' strand, until recently considered the carrier strand devoid of biological significance but now know to be fully functioning and independently controlled by the Ago 2, is most often regulated respect to the corresponding 'non-star' counterpart, which often do not show significant changes when compared to the control (0h-EtOH). The abundance of star sequences in our datasets could be explained by RISC incorporation of star arms, due to the thermodynamic stability of the miRNA–miRNA* duplex. These alternate mature forms share similar evolutionary and structural signatures, and show similar relationships with target 3' UTRs. The results obtained here are also in agreement with the involvement of oestrogen and its receptors in miRNA maturation kinetics, as recently demonstrated for ERβ in BC cells by Paris et al. [30]. As these alternate miRNA species target different transcripts respect to non-star strands, they increase the number of targets for each miRNA gene and may constitute also a powerful evolutionary mechanism for the emergence of new miRNAs.

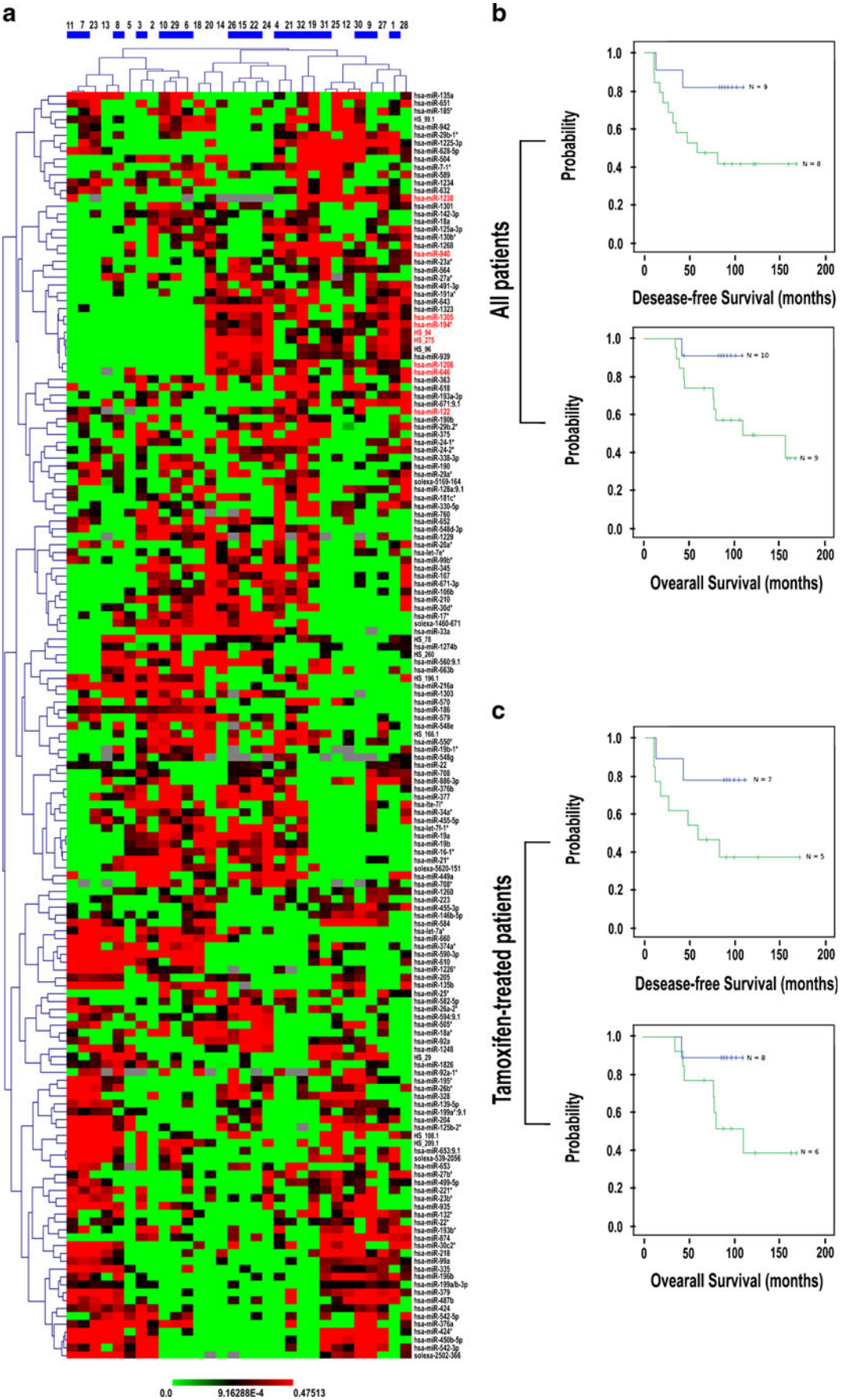
Considering the biological significance of the results obtained, we focused our attention on E2-regulated miRNAs in common between the two BC cell lines analysed as they represent a good starting point to investigate the role of miRNAs in modulating the final responses to oestrogen of hormone-responsive genes in BC cells and to understand in depth the spectrum of molecular mechanisms dependent upon oestrogen in BC. Among the many observations made, we noted that, together with miRNAs transcribed as independent transcripts, several miRNA gene clusters show the same changes in response to E2 in both cell lines, suggesting a global effect of ligand-activated receptor on the entire gene locus. Moreover, we observed that most regulated miRNAs exhibit a significant variation in expression 24–72 h after E2 stimulation. This may be due to the fact that increase/decrease of pri- and pre-miR expression level occurs early after E2 stimulation, while mature miRNA expression level start to vary later, as shown by Castellano et al. [58], who focused their attention on the cluster 17–92 and demonstrated that if the pri-miR levels are quickly increased during the first 3 h of oestrogen stimulation, while the mature forms result mainly increased only after 24–72 h of oestrogen treatment.

As partial confirmation of the results described here, we observed that a number of miRNAs found oestrogen-responsive in our expression profiling assays were identified also in other studies performed in BC cells. E2, for example, was found to increase expression of some members of the miR-17-92 cluster, as well as of miR-424*, miR-450b-3p/5p and miR-542-3p and of miR-25* [28, 59]. Conversely, it was shown to decrease expression of miR-34a*, let-7e*, miR-125a-3p, miR-181a*, miR-181c*, miR-26b*, miR-618 and star members of the miR-23a/27a/24-2 cluster [28, 60].

**Fig. 4** Analysis of miRNA expression in 30 breast cancer samples. **a** ▶ Hierarchical clustering of 161 miRNAs oestrogen-regulated in vitro in at least one cell line. The blue bars denote ER+ samples. *Numbers at the top* indicate individual tumour number (see Supplemental Table S3 for details). The two main branches were used in survival analysis: **b** Kaplan–Meier survival analysis and log rank test concerning all 30 samples ($p<0.05$ for DFS, top; $p<0.08$ for OS, *bottom*) or **c** limited to patients receiving adjuvant Tamoxifen ($p<0.09$ for DFS, top; NS for OS, *bottom*). 'N' on the curves denotes the censored events in each group. The miRNAs marked in *red* are in common between the miRNAs classifying breast tumours and those E2-regulated in BC cell lines

Since miRNAs exert their actions on target mRNAs at the post-transcriptional level, a variation of their cellular levels upon E2 stimulation could have important functional roles. As a consequence, identification of miRNA targets is crucial to understand the functional significance of oestrogen-mediated miRNA expression changes in BC cells. For this reason, we performed an in silico target analysis on E2-regulated miRNAs and found how almost 33% of E2-responsive mRNAs represent putative targets of regulated miRNAs. To identify inverse relationships between changes in oestrogen-responsive miRNA levels and those of the corresponding target mRNAs, we took in account all mRNAs targeted by a single miRNA. Results reveal 125 mRNAs that show and correlation respect to the matched miRNA (Supplementary Table S1), with a significant prevalence ($p<0.003$) for down-regulated mRNAs (14%, ranging from 6% to 23% for each of the E2-responsive gene clusters described in these same cell lines [48]), respect to up-regulated ones (8%, ranging between 3% and 13% in the activated gene clusters). These results indicate a deep impact of E2/ERα on BC cell transcriptome regulation via E2-responsive miRNAs. Considering the biological significance of this finding, we searched biological functions reflecting the activity of mRNAs that are significantly over-represented among all predicted target transcripts by Gene Ontology analysis. The results reported in Supplementary Fig. S2 show that several biological processes are indeed controlled *via* the miRNA-dependent pathway described here. Indeed, the presence of target genes involved in the cell proliferation, gene transcription, signal transduction and apoptosis indicates that ERα affects in this way a number of key cellular processes through this post-transcriptional regulatory mechanism where ERs appear to influence the activity of target genes through miRNA-mediated post-transcriptional regulation of the activity of gene networks.

Concerning the mechanism for miRNA regulation by the oestrogen receptor, starting from expression profiling and ERα ChIP-Seq and ChIA-Pet data, we propose at least two different mechanisms for hormone-mediated regulation of miRNA gene activity. One is represented by ERα binding in proximity of miRNA-encoding genes that suggests a direct


a
11 7 23 13 8 5 3 2 10 29 6 18 20 14 26 15 22 24 4 21 32 19 31 25 12 30 9 27 1 28
0.0
9.16288E-4
0.47513
b
All patients
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
N = 9
N = 8
Desease-free Survival (months)
Probability
N = 10
N = 9
Overall Survival (months)
c
Tamoxifen-treated patients
Probability
N = 7
N = 5
Desease-free Survival (months)
Probability
N = 8
N = 6
Overall Survival (months)

involvement of the receptor in transcriptional regulation of pri-miR synthesis by transcription. As shown in Tables 1 and S2, six E2-regulated miRNAs have an ERα binding site within 10 kb from the miRNA gene at a major distance (50 kb), but still compatible with the mechanism cited above. The presence of binding sites for the receptor upstream and/or downstream of E2-downregulated miR-27a, one of the components of the miR-23a-27a/24-2 cluster, is particularly interesting as targets of this miRNA identified based upon changes in protein levels after treatment with a 27a antagomiR include the transcriptional co-factor ZBTB10/RINZF. ZBTB10 is a repressor of the SP1 transcription factor, which is found overexpressed in a variety of cancers and is thought to play a role in the G0–G1 to S phase progression in BC cells [63]. Similarly, direct upregulation of miR-135a of is interest, considering that this miRNA has been found to be oncogenic, able to regulate APC and Jak2 gene transcripts, which both encode proteins involved in cell survival and proliferation, angiogenesis and immune evasion [64, 65]. The second E2-mediated miRNA regulation mechanism suggested by the results reported here is represented by the involvement of E2-inducibile expression of mRNA-encoding genes that harbour microRNA genes in their intronic regions. In order to investigate this mechanism, we controlled first which E2-regulated miRNAs were located in host mRNA genes and, subsequently, which of these genes were E2-regulated. In this way, we identified 6 E2-regulated miRNAs that are encoded by a host gene that, in all but one case, is itself regulated by the hormone with the same kinetics (Table 2). These considerations do not exclude, of course, other potential mechanisms by which ligand-activated ERα might affect miRNA expression levels. For example, several studies have demonstrated the presence of a crosstalk of the miRNA maturation pathways with intracellular signaling molecules as p53 [66], Smad proteins [67] and ER itself. Recent findings, in fact, suggest a role for E2/ERα action in the biogenesis of miRNAs, where ERα is able to down-regulate miRNA expression blocking Drosha-mediated processing of a subset of miRNAs by binding to Drosha in a p68/p72-dependent manner and inducing the dissociation of the microprocessor complex from pri-miRNA [68].

Analysis of miRNA expression in clinical samples confirmed the general finding that these noncoding RNAs show significant associations with clinical outcome more than protein coding genes [69, 70]. In fact, significant survival curve separations were observed in terms of both disease-free and overall survival, despite the fact that the cohort of patients examined here is small ($N$=30). Notably, the group of miRNAs regulated by oestrogen in vitro was found comparably effective in discriminating survival, although the small number of samples does not allow to evidence specific effects in the group of ER-positive, Tamoxifen-treated samples. In contrast, we could not find any association of regulated miRNAs with ER status. One likely explanation for this is that miRNA expression levels in tumours respond to many different stimuli aside from oestrogen receptor-mediated regulation. The same finding was in part true for protein coding RNAs [71]. For the reason stated above, we are not surprised by the fact that there is no direct correlation between hormone-regulated miRNAs and the miRNA profile that correlates with clinical follow-up. Indeed, the number of RNAs found is not significant, as any other set of corresponding size will contain similar numbers of responsive probes.

Overall, our data indicate that miRNA expression play a key role in oestrogen-dependent functions in BC and possibly other cell types, suggesting that miRNA modulation by ERα represents a novel genetic pathway controlled by these steroid hormones that could impact oestrogen-dependent breast tumour biology and, thereby, influence the clinical and pharmacological profile of the disease.

**Acknowledgements** The authors thank Rosario Casale and Maria Francesca Papa for technical assistance and Claudio Scafoglio for critically reading the revised manuscript. Work supported by: European Union (CRESCENDO I.P., contract number LSHM-CT2005-018652), Italian Association for Cancer Research (grant IG-8586), Ministry for Education, University and Research (grants PRIN 2008CJ4SYW_004) and University of Salerno (Fondi FARB 2011). CC, FR, GN and RT are fellows of Fondazione con il Sud, MR is supported by a 'Vladimir Ashkenazy' fellowship of Italian Association for Cancer Research, MRDF is PhD student of the Research Doctorate 'Computational Biology and Bioinformatics' of the University of Napoli 'Federico II' supported by Fondazione IRCCS SDN, FC is PhD student of the Research Doctorate 'Molecular Pathology and Physiopathology' of the University of Napoli 'Federico II', GG and SS are PhD students of the Research Doctorate 'Experimental Physiopathology and Neurosciences' of the Second University of Napoli and CS is PhD student of the Research Doctorate 'Molecular Oncology, Experimental Immunology and Innovative Therapy Development' of the 'Magna Graecia' University of Catanzaro.

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

1. Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Hartman J, Tujague M et al (2007) Estrogen receptors: how do they signal and what are their targets. Physiol Rev 87:905–931
2. Hall JM, Couse JF, Korach KS (2001) The multifaceted mechanisms of estradiol and estrogen receptor signaling. J Biol Chem 276:36869–36872
3. Bai Z, Gust R (2009) Breast cancer, estrogen receptor and ligands. Arch Pharm (Weinheim) 342:133–149
4. Kushner PJ, Agard DA, Greene GL, Scanlan TS, Shiau AK, Uht RM, Webb P (2000) Estrogen receptor pathways to AP-1. J Steroid Biochem Mol Biol 74:311–317
5. Saville B, Wormke M, Wang F, Nguyen T, Enmark E, Kuiper G, Gustafsson JA et al (2000) Ligand-, cell-, estrogen receptor

subtype (alpha/beta)-dependent activation at GC-rich (Sp1) promoter elements. J Biol Chem 275:5379–5387
6. McDonnell DP, Norris JD (2002) Connections and regulation of the human estrogen receptor. Science 296:1642–1644
7. Klinge CM, Jernigan SC, Mattingly KA, Risinger KE, Zhang J (2004) Estrogen response element-dependent regulation of transcriptional activation of estrogen receptors alpha and beta by coactivators and corepressors. J Mol Endocrinol 33:387–410
8. Chen GG, Zeng Q, Tse GM (2008) Estrogen and its receptors in cancer. Med Res Rev 28:954–974
9. Manavathi B, Kumar R (2006) Steering estrogen signals from the plasma membrane to the nucleus: two sides of the coin. J Cell Physiol 207:594–604
10. Safe S, Kim KJ (2008) Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways. J Mol Endocrinol 41:263–275
11. Gronemeyer H, Gustafsson JA, Laudet V (2004) Principles for modulation of the nuclear receptor superfamily. Nat Rev Drug Discov 3:950–964
12. Kim VN (2005) MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol 6:376–385
13. Xiong H, Qian J, He T, Li F (2009) Indipendent transcription of miR-281 in the intron of ODA in *Drosophila melanogaster*. Biochem Biophys Res Commun 378:883–889
14. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14:1902–1910
15. Saini HK, Griffiths-Jones S, Enright AJ (2007) Genomic analysis of human microRNA transcripts. Proc Natl Acad Sci USA 104:17719–17724
16. Lee Y, Kim M, Han JJ, Yeom KH, Lee S, Baek SH, Kim VN (2004) MicroRNA genes are transcribed by RNA polymerase II. EMBO J 23:4051–4060
17. Denli AM, Tops BB, Plasterk RH, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. Nature 432:231–235
18. Lee Y, Jeon K, Lee JT, Kim S, Kim VN (2002) microRNA maturation: stepwise processing and subcellular localization. EMBO J 21:4663–4670
19. Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R (2005) Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell 123:631–640
20. Engels BM, Hutvagner G (2006) Principles and effects of microRNA-mediated post-transcriptional gene regulation. Oncogene 25:6163–6169
21. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5:522–531
22. Linsen S, Tops B, Cuppen E (2008) miRNAs small changes, widespread effects. Cell Res 18:1157–1159
23. Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120:15–20
24. Zhao Y, Srivastava D (2007) A developmental view of microRNA function. Trends Biochem Sci 32:189–197
25. Ventura A, Jacks T (2009) MicroRNAs and cancer: short RNAs go a long way. Cell 136:586–591
26. Tessel MA, Krett NL, Rosen ST (2010) Steroid receptor and microRNA regulation in cancer. Curr Opin Oncol 22:592–597
27. Klinge CM (2009) Estrogen regulation of miRNA expression. Curr Genomics 10:169–183
28. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, Brown M et al (2009) Estradiol-regulated microRNAs control estradiol response in breast cancer cells. Nucleic Acids Res 37:4850–4861
29. Yang Z, Wang L (2011) Regulation of microRNA expression and function by nuclear receptor signaling. Cell Biosci 1:31
30. Paris O, Ferraro L, Grober OMV, Ravo M, De Filippo MR, Giurato G, Nassa G, Tarallo et al. (2012) Direct regulation of microRNA biogenesis and expression by estrogen receptor beta in hormone-responsive breast cancer. Oncogene. doi:10.1038/onc.2011.583
31. Esquela-Kerscher A, Slack FJ (2006) Oncomirs—microRNAs with a role in cancer. Nat Rev Cancer 6:259–269
32. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E et al (2005) MicroRNA gene expression deregulation in human breast cancer. Cancer Res 65:7065–7070
33. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A et al (2005) MicroRNA expression profiles classify human cancers. Nature 435:834–838
34. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ, Barbosa-Morais NL et al (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. Genome Biol 8:R214
35. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR et al (2007) Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. Cancer Res 67:2456–2468
36. O'Day E, Lal A (2010) MicroRNAs and their target gene networks in breast cancer. Breast Cancer Res 12:201
37. Adams BD, Furneaux H, White BA (2007) The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor alpha (ERalpha) and represses ERalpha messenger RNA and protein expression in breast cancer cell lines. Mol Endocrinol 21:1132–1147
38. Hossain A, Kuo MT, Saunders GF (2006) Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. Mol Cell Biol 26:8191–8201
39. Yu Z, Wang C, Wang M, Li Z, Casimiro MC, Liu M, Wu K et al (2008) A cyclin D1/microRNA 17/20 regulatory feedback loop in control of breast cancer cell proliferation. J Cell Biol 182:509–517
40. Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C, Ambros VR et al (2007) Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. Cancer Res 67:2456–2468
41. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R et al (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci USA 103:2257–2261
42. Gusev Y, Schmittgen TD, Lerner M, Postier R, Brackett D (2007) Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. BMC Bioinforma 8(Suppl 7):S16
43. Cochrane DR, Cittelly DM, Howe EN, Spoelstra NS, McKinsey EL, LaPara K, Elias A et al (2010) MicroRNAs link estrogen receptor alpha status and dicer levels in breast cancer. Horm Cancer 1:306–319
44. Ambrosino C, Tarallo R, Bamundo A, Cuomo D, Franci G, Nassa G, Paris O et al (2010) Identification of a hormone-regulated dynamic nuclear actin network associated with estrogen receptor alpha in human breast cancer cell nuclei. Mol Cell Proteomics 9:1352–1367
45. Cimino D, Fuso L, Sfiligoi C, Biglia N, Ponzone R, Maggiorotto F, Russo G et al (2008) Identification of new genes associated with breast cancer progression by gene expression analysis of predefined set of neoplastic tissues. Int J Cancer 123:1327–1338
46. Scafoglio C, Ambrosino C, Cicatiello L, Altucci L, Ardovino M, Bontempo P, Medici N et al (2006) Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cells. J Cell Biochem 98:1163–1184
47. Cicatiello L, Scafoglio C, Altucci L, Cancemi M, Natoli G, Facchiano A, Iazzetti G et al (2004) A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. J Mol Endocrinol 32:719–775

48. Cicatiello L, Mutarelli M, Grober OMV, Paris O, Ferraro L, Ravo M, Tarallo R et al (2010) A gene network controlled by estrogen receptor α in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. Am J Pathol 176:2113–2130
49. Ravo M, Mutarelli M, Ferraro L, Grober OMV, Paris O, Tarallo R, Vigilante A et al (2008) Quantitative expression profilino of highly degraded RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays. Lab Inv 88:430–440
50. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci 98:5116–5121
51. Grober OMV, Mutarelli M, Giurato G, Ravo M, Cicatiello L, De Filippo MR, Ferraro L et al (2011) Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. BMC Genomics 12:36
52. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics 24:1729–1730
53. Qiunian AR, Hall IN (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842
54. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3
55. da Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc 4:44–57
56. Weigelt K, Moehle C, Stempfl T, Weber B, Langmann T (2008) An integrated workflow for analysis of ChIP-chip data. Biotechniques 45:131–132
57. Cicatiello L, Scafoglio C, Altucci L, Cancemi M, Natoli G, Facchiano A, Iazzetti G et al (2004) A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. J Mol Endocrinol 32:719–775
58. Cicatiello L, Addeo R, Altucci L, Belsito Petrizzi V, Boccia V, Cancemi M, Germano D et al (2000) The antiestrogen ICI 182,780 inhibits proliferation of human breast cancer cells by interfering with multiple, sequential estrogen-regulated processes required for cell cycle completion. Mol Cell Endocrinol 165:199–209
59. Castellano L, Giamas G, Jacob J, Coombes RC, Lucchesi W, Thiruchelvam P, Barton G et al (2009) The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response. Proc Natl Acad Sci U S A 106:15732–15737
60. Maillot G, Lacroix-Triki M, Pierredon S, Gratadou L, Schmidt S, Bénès V, Roché H et al (2009) Widespread estrogen-dependent repression of micrornas involved in breast tumor cell growth. Cancer Res 69:8332–8340
61. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL et al (2009) An oestrogen-receptor-alpha bound human chromatin interactome. Nature 462:58–64
62. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10:126–139
63. Li X, Mertens-Talcott SU, Shu Z, KyoungHyun K, Judith B, Stephen S (2010) MicroRNA-27a indirectly regulates estrogen receptor alpha expression and hormone responsiveness in MCF-7 breast cancer cells. Endocrinology 151:2462–2473
64. Nagel R, le Sage C, Diosdado B, van der Waal M, Oude Vrielink JA, Bolijn A, Meijer GA et al (2008) Regulation of the adenomatous polyposis coli gene by the miR-135 family in colorectal cancer. Cancer Res 68:5795–5802
65. Navarro A, Diaz T, Martinez A, Gaya A, Pons A, Gel B, Codony C et al (2009) Regulation of JAK2 by miR-135a: prognostic impact in classic Hodgkin lymphoma. Blood 114:2945–2951
66. Suzuki HI, Yamagata K, Sugimoto K, Iwamoto T, Kato S, Miyazono K (2009) Modulation of microRNA processing by p53. Nature 460:529–533
67. Davis BN, Hilyard AC, Lagna G, Hata A (2008) SMAD proteins control DROSHA-mediated microRNA maturation. Nature 454:56–61
68. Yamagata K, Fujiyama S, Ito S, Ueda T, Murata T, Naitou M, Takeyama K et al (2009) Maturation of microRNA is hormonally regulated by a nuclear receptor. Mol Cell 36:340–347
69. Buffa FM, Camps C, Winchester L, Snell CE, Gee HE, Sheldon H, Taylor M et al (2011) microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. Cancer Res 71:5635–5645
70. Qian B, Katsaros D, Lu L, Preti M, Durando A, Arisio R, Mu L et al (2009) High miR-21 expression in breast cancer associated with poor disease-free survival in early stage disease and highTGF-beta1. Breast Cancer Rea Treat 117:131–140
71. Weisz A, Basile W, Scafoglio C, Altucci L, Bresciani F, Facchiano A, Sismondi P (2004) Molecular identification of ERalpha-positive breast cancer cells by the expression profile of an intrinsic set of estrogen regulated genes. J Cell Physiol 200:440–450

Oncogene (2012) 1–11

www.nature.com/onc

ORIGINAL ARTICLE

# Direct regulation of microRNA biogenesis and expression by estrogen receptor beta in hormone-responsive breast cancer

O Paris[1], L Ferraro[1], OMV Grober[1], M Ravo[2], MR De Filippo[1,2], G Giurato[1,2], G Nassa[2], R Tarallo[2], C Cantarella[2], F Rizzo[2], A Di Benedetto[3], M Mottolese[3], V Benes[4], C Ambrosino[5], E Nola[1] and A Weisz[1,2,6]

*[1]Department of General Pathology, Second University of Naples, Napoli, Italy; [2]Laboratory of Molecular Medicine and Genomics, University of Salerno, Baronissi, Italy; [3]Department of Pathology, Regina Elena Cancer Institute, Rome, Italy; [4]Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany; [5]Department of Biological and Environmental Sciences, University of Sannio, Benevento, Italy and [6]Division of Molecular Pathology and Medical Genomics, 'SS. Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Salerno, Italy*

**Estrogen effects on mammary epithelial and breast cancer (BC) cells are mediated by the nuclear receptors ERα and ERβ, transcription factors that display functional antagonism with each other, with ERβ acting as oncosuppressor and interfering with the effects of ERα on cell proliferation, tumor promotion and progression. Indeed, hormone-responsive, ERα+ BC cells often lack ERβ, which when present associates with a less aggressive clinical phenotype of the disease. Recent evidences point to a significant role of microRNAs (miRNAs) in BC, where specific miRNA expression profiles associate with distinct clinical and biological phenotypes of the lesion. Considering the possibility that ERβ might influence BC cell behavior via miRNAs, we compared miRNome expression in ERβ+ vs ERβ− hormone-responsive BC cells and found a widespread effect of this ER subtype on the expression pattern of these non-coding RNAs. More importantly, the expression pattern of 67 miRNAs, including 10 regulated by ERβ in BC cells, clearly distinguishes ERβ+, node-negative, from ERβ−, metastatic, mammary tumors. Molecular dissection of miRNA biogenesis revealed multiple mechanisms for direct regulation of this process by ERβ+ in BC cell nuclei. In particular, ERβ downregulates miR-30a by binding to two specific sites proximal to the gene and thereby inhibiting pri-miR synthesis. On the other hand, the receptor promotes miR-23b, -27b and 24-1 accumulation in the cell by binding in close proximity of the corresponding gene cluster and preventing *in situ* the inhibitory effects of ERα on pri-miR maturation by the p68/DDX5-Drosha microprocessor complex. These results indicate that cell autonomous regulation of miRNA expression is part of the mechanism of action of ERβ in BC cells and could contribute to establishment or maintenance of a less aggressive tumor phenotype mediated by this nuclear receptor.**

Correspondence: Dr A Weisz, Laboratorio di Medicina Molecolare e Genomica, Università degli Studi di Salerno, via Allende, Baronissi SA 84081, Italy.
E-mail: aweisz@unisa.it
Received 5 September 2011; revised 29 October 2011; accepted 3 November 2011

*Oncogene* advance online publication, 9 January 2012; doi:10.1038/onc.2011.583

**Keywords:** estrogen receptor beta; microRNA; breast cancer; hormones; gene transcription

## Introduction

Estrogens have a role in breast cancer (BC) pathogenesis and progression by controlling mammary cell proliferation and key cellular functions via the estrogen receptors (ERα and ERβ: Heldring *et al*., 2007). ERs are members of the nuclear receptors superfamily of ligand-dependent transcription factors that both regulate gene expression controlling the estrogen signal transduction cascade with distinct and even antagonistic roles. In hormone-responsive, ERα-positive BC cells ERβ inhibits estrogen-mediated cell proliferation by increasing the expression of growth-inhibitory genes and by interfering with activation of cell cycle and anti-apoptotic genes by ERα in response to 17β-estradiol (E2: Chang *et al*., 2006; Grober *et al*., 2011). ERβ is frequently lost in BC, where its presence generally correlates with a better prognosis of the disease (Sugiura *et al*., 2007), is a biomarker of a less aggressive clinical phenotype (Novelli *et al*., 2008; Shaaban *et al*., 2008) and its downregulation has been postulated to represent a critical stage in estrogen-dependent tumor progression (Roger *et al*., 2001; Bardin *et al*., 2004). Despite the direct relationships between estrogen and breast carcinogenesis, the divergent roles of the two ER subtypes in BC are not fully understood, mostly because they are complex, involving genomic and non-genomic actions, regulation of gene transcription and control of mRNA stability and translation efficiency.

MicroRNAs (miRNAs) are small (20–25 nt) non-coding RNAs that can regulate gene activity in a posttranscriptional manner. These molecules, frequently transcribed as polycistronic RNAs, are synthesized in the nucleus by RNA polymerase II or III as long primary transcripts (pri-miRNAs), that are then

processed by the class-2 RNase-III Drosha (Han *et al*., 2004) in ~70-nucleotide stem-loop RNAs (pre-miRNAs), that in turn are exported from nucleus to cytoplasm by exportin 5 and Ran-GTP (Kim *et al*., 2009) and cleaved by Dicer/TRBP endoribonuclease into an imperfect miRNA/miRNA* duplex (Chendrimada *et al*., 2005). Only one strand of the duplex is finally selected to function as a mature miRNA, whereas the other (passenger) strand is typically degraded (Okamura *et al*., 2008; Newman and Hammond, 2010). Mature miRNAs are then incorporated into an RNA-induced silencing complex, which binds to target mRNAs, determining gene silencing by either inhibition of translation or mRNA degradation (Newman and Hammond, 2010). miRNAs have been shown to regulate a wide variety of cellular phenotypes, including neoplastic transformation, cell proliferation, differentiation and homeostasis (Garzon *et al*., 2009) and altered expression of these small RNAs contributes to tumorigenesis, as some of them can function as either tumor suppressors or oncogenes (Zhang *et al*., 2007; Croce, 2009). Interestingly, in solid tumors, such as prostate, colon, stomach, pancreas, lung and breast, the spectrum of miRNAs expressed (miRNome) is different from that of the corresponding normal tissues (Volinia *et al*., 2006), suggesting the involvement of miRNAs in transformed cell biology. Differential expression of miRNA genes was found associated with specific pathological features of BC, where distinct miRNA expression profiles in normal vs cancer tissue or between different molecular and clinical tumor subtypes appears to be the rule (Iorio *et al*., 2005; Lu *et al*., 2005; Blenkiron *et al*., 2007; Tavazoie *et al*., 2008). There is increasing evidence, in fact, that specific miRNAs may be responsible at large for disease heterogeneity, functioning as regulators of tumorigenicity, invasion and metastasis (Tavazoie *et al*., 2008). Moreover, genetic defects in key components of the miRNA biosynthetic pathway have been described in tumors (Hill *et al*., 2009; Melo *et al*., 2009, 2010), and several genes involved in BC progression have been identified as targets of miRNAs that, in turn, are found deregulated in BC cells (Garzon *et al*., 2009).

Several evidences indicate that ERα is among the transcription factors regulating miRNA biogenesis in hormone-responsive BC cells (Bhat-Nakshatri *et al*., 2009; Castellano *et al*., 2009; Maillot *et al*., 2009; Yamagata *et al*., 2009; Cicatiello *et al*., 2010; Ferraro *et al*., 2010, 2011). More recently, global mapping of ERβ binding to ERα-positive, hormone-responsive BC cells chromatin *in vivo* showed ERβ interaction with several miRNA genes, suggesting the possible involvement of this receptor in hormonal control of small non-coding RNA biogenesis in this cell type (Grober *et al*., 2011). Starting from this observation, we investigated here miRNA expression pattern in estrogen-responsive BC cell lines engineered to express full-length ERβ and in primary-tumor samples selected according to the presence or absence of this nuclear receptor. Results indicate a role of ERβ in the control of miRNA biogenesis and expression pattern in BC cells.

## Results

### *ERβ induces widespread changes in miRNA expression in hormone-responsive cells*

In order to investigate the role of ERβ in BC, we generated MCF-7 cells stably expressing full-length human ERβ (ERβ-1) fused at the N- (N-TAP-ERβ) or C- (C-TAP-ERβ) terminus to a TAP tag in pTRE2pur-HA expression vector (Puig *et al*., 2001). As shown in Figure 1a, the expression levels of C-TAP-ERβ (two independent clones: lanes 2–3), N-TAP-ERβ (lane 4) or C-TAP-ERα (used as control: lane 5) are comparable to those relative to endogenous ERα, as detected by WB under comparable test conditions, to avoid toxic and artifactual events consequent to overexpression of the exogenous protein. The functional integrity of tagged ERβ was assessed by measuring their ability to counteract induction of ERE-TK-luciferase reporter-gene transcription by ligand-activated endogenous ERα. As shown in Figure 1b, cell expressing TAP-ERβ show a marked reduction in E2-mediated activation of reporter-gene transcription compared with *wt* cells, a phenotype that could be almost completely recovered by stimulation with the ERα-selective ligand 4,4′,4″-(4-propyl-[1H]-pyrazole-1,3,5-triyl)trisphenol (PPT). TAP-ERβ effects on E2-induced MCF-7 cell proliferation and cell cycle progression were also investigated and the results, reported in Figures 1c and d, show that cells expressing exogenous ERβ grow much slower in response to estrogen than *wt* or C-TAP-ERα cells, consequent to reduced G1–S transition (Figure 1d). It is worth mentioning that the cell cycle inhibitory effects of ERβ are well known (Heldring *et al*., 2007; Grober *et al*., 2011, and references therein) and are more evident at relatively higher concentrations of E2 ($\geqslant 10^{-10}$), compatible with the lower affinity of this ER subtype for the hormone (compare, for each cell clone, the S + G2 fraction in hormone-stimulated vs -starved cells). The efficiency of PPT in promoting cell cycle progression (Figure 1d) relates to its ability to promote ERα-mediated gene transcription (Figure 1b), confirming the direct link between transcriptional activity of this receptor subtype and the mitogenic effects of estrogen (Cicatiello *et al*., 2010). Gene-expression profiling of asynchronously growing cells showed no major differences between N- and C-TAP-ERβ cells, whereas their transcriptomes were significantly different from that of C-TAP-ERα cells (Supplementary Figure S1), confirming previous results obtained in E2-stimulated cells (Grober *et al*., 2011). Based on these results, expression of the TAP-ERβ fusion proteins appears to significantly affect ERα-mediated estrogen signal transduction to target genes and the cell cycle, confirming previous observations indicating that they are fully functional *in vivo* (Grober *et al*., 2011; Nassa *et al*., 2011).

Multiple roles have been proposed for miRNAs in hormone-responsive BC, where the presence of ERβ has been shown to associate with less aggressive disease forms. We decided to use our ERβ-expressing cells to investigate potential links between ERβ and miRNA activity in hormone-responsive BC cells, as these

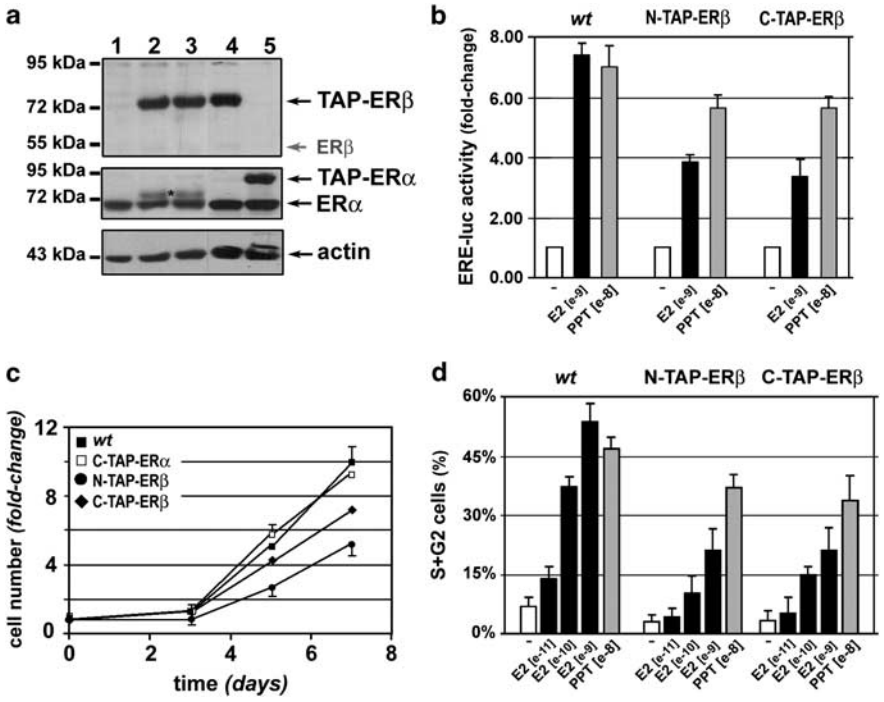



**Figure 1** Functional characterization of ERβ-expressing MCF-7 cell clones. (**a**) Western blotting analysis of protein extracts from control (*wt*, lane 1) and TAP-ERs (N-TAP-ERβ: lanes 2–3, C-TAP-ERβ: lane 4, C-TAP-ERα: lane 5) expressing cells. Asterisks mark non-specific bands. (**b**) The ability of tagged ERβ to interfere with ERα activity was assessed by comparing estrogen effects in *wt*, N-TAP-ERβ and C-TAP-ERβ cells by ERE-TK-luciferase reporter-gene activation mediated by E2 or PPT (selective ERα agonist). (**c**) Proliferation rate of *wt*, N-TAP-ERβ and C-TAP-ERβ cells was measured in hormone-starved cells stimulated with $10^{-8}$ M E2, respect to untreated cells. Cell counting was performed with a colorimetric assay at the indicated times. (**d**) Analysis of cell cycle progression after estrogen stimulation of *wt* or TAP-ERβ-expressing cells. The percent of S + G2 phase cells was determined by flow cytometry in estrogen-starved cultures 27 h after treatment with either vehicle alone (EtOH) or the indicated concentrations of E2 or PPT.

represent a useful *in vitro* model to investigate the molecular mechanisms underlying the biological effects of this receptor subtype in hormone-responsive tumors. To this aim, total RNA was extracted from *wt* MCF-7, N-TAP-ERβ, C-TAP-ERβ (2 independent clones) and C-TAP-ERα cells. Global analysis of miRNome expression was peformed with microarrays detecting the vast majority of known and characterized miRNAs (Illumina MicroRNA Expression Beadchip, Illumina Italia, Milano, Italy) as described in Material and methods. Results indicates that expression of ERβ has a deep impact on BC cell miRNome, as 84 miRNAs were found differentially expressed in three ERβ+ vs two ERβ− cell lines, whereas no significant differences could be detected among cells expressing the different tagged forms of ERβ, or between C-TAP-ERα, *wt* and MCF7-TAP cells (not shown), that express only the TAP peptide and show no differences in ERα signaling with respect to *wt* cells (Ambrosino *et al*., 2010; Grober *et al*., 2011). To validate this result, we performed miRNA expression profiling with a different microarray platform (Agilent Human microRNA Microarrays 18 × 15 K v3, Agilent Technologies Italia, Milano, Italy) and compared the results obtained in the two experimental settings. As expected, we observed some differences between the two data sets, likely due to technical differences between the two microarrays platforms (in particular sensitivity and quality of the probes) and the two probe sets (Supplementary Materials and methods and data not shown). Nevertheless, 73 among the differentially expressed miRNAs identified with the Illumina platform were either fully confirmed with the Agilent array or, in some instances, could not be detected here due to a lower sensitivity of this platform. For this reason, we performed a further validation of the results obtained with the Illumina arrays analyzing by real-time RT–PCR (reverse transcriptase–PCR) the expression levels of 10 miRNAs selected according to their relative expression level, ranging from very low to high, and including also miRNA undetectable with Agilent arrays or differentially expressed between cell lines (except for miR-181c, that was not differentially expressed and is included as negative control). Results (reported in Supplementary Figure S2) show a very high correlation between rtPCR and Illumina array data (correlation coefficient: 0.76), indicating reliability of this microarray platform. The 73 differentially expressed miRNAs listed in Figure 2a and Table 1 were thus considered validated. To gather insights on the molecular mechanisms for ERβ effects on miRNAs, expression profiling was carried out in both cell types after estrogen starvation. Under these conditions, no differences could be detected between ERβ+ and ERβ− cells (left panel in Figure 2b, referring to average values measured in ERβ− vs ERβ+ cell lines), indicating a key role of the liganded in determining the observed differences. For this reason, we next investigated whether expression of the 73 miRNAs identified in the

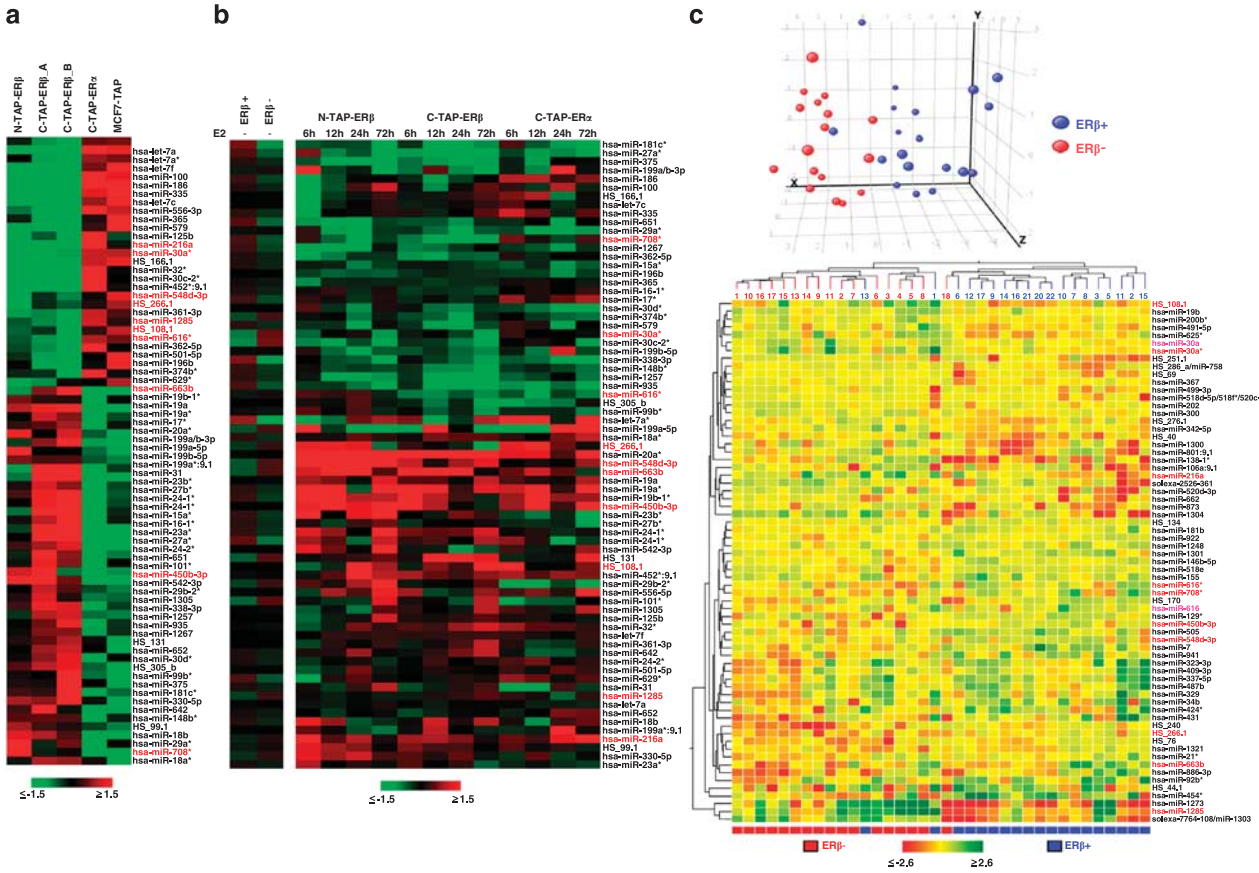



**Figure 2** Correlations between ERβ and miRNome expression in hormone-responsive BC cells and primary breast carcinomas. (**a**) Heatmap showing 73 miRNAs differentially expressed between ERβ+ and ERβ− cells maintained in standard culture conditions. Data displayed represent the ratio between the fluorescence intensity value of each miRNA in a given array (cell line) vs the average of the fluorescence intensity value of the same miRNA in all arrays. (**b**) Heatmaps showing relative expression of 73 ERβ-responsive miRNAs in ERβ+ and ERβ− cells treated with vehicle alone (EtOH; left panel) or with E2 for the indicated times (right panel). Data displayed represent the ratio between the fluorescence intensity value of each miRNA at the indicated time after E2 stimulation (+E2) vs the same in hormone-starved cells (−E2, control). (**c**) Top: Principal component analysis (PCA) relative to differential miRNA expression in 17 ERβ+ and 19 ERβ− primary BC samples. Bottom: Cluster analysis of 67 miRNAs discriminating between ERβ+ and ERβ− BC samples. Data displayed represent the ratio between the fluorescence intensity value of each miRNA in a given array (tumor sample) vs the average of the fluorescence intensity value of the same miRNA in all arrays. miRNA marked in red were differentially expressed both in ERβ+ cell lines and BC samples, whereas those marked in purple in C derive from the same pre-miR of those differentially expressed in ERβ+ cell lines and tumors.

previous experiment was affected by estrogen. To this aim, N-TAP-ERβ, C-TAP-ERβ and C-TAP-ERα were E2-deprived and subsequently stimulated with $10^{-8}$ M E2 for 6–72 hrs before miRNA analysis. Results displayed in Figure 2b (right panel) show that all investigated miRNAs respond to the hormone in a time-dependent manner. Although kinetics and extent of miRNA response to the stimulus were comparable between the two ERβ+ cell lines, they were significantly different in ERβ+ vs ERβ− (C-TAP-ERα) cells. Direct comparison of the data from the two cell types indicates that the differences in steady-state miRNA levels consequent to ERβ expression are due to ERβ antagonism upon ERα activity or to a specific effect of ligand-activated ERβ. As shown in Supplementary Figure S3, for example, expression of hsa-miR30a* and, to a lesser extent, hsa-miR30a shows a time-dependent decrease following E2 stimulation only in ERβ+ cells, whereas it is unaffected by the stimulus in the absence of ERβ. On the contrary, hsa-miR-23b and -23b*, hsa-miR-27b and -27b* and hsa-miR-24 and -24-1* levels decrease in the presence of E2 in ERβ− whereas they increase in ERβ+ cells. The putative mRNA targets of the miRNAs regulated by ERβ were searched with TargetScan and, subsequently, analyzed for Gene Ontology term over-representation, in order to identify biological processes likely to be influenced by this ER subtype via miRNAs. In this way, several cellular processes were found downstream of ERβ-responsive miRNAs, including those known to be affected by ERβ, such as response to hormonal stimuli, regulation of transcription and cell proliferation and others that represent key cellular processes in malignant cells, including cell motility, migration, adhesion, differentiation and fate determination, and are targeted by regulatory cascades in cancer cells (Supplementary Figure S4).

The data described above were obtained *in vitro* in a BC cell model that, although it has been shown to reflect

**Table 1** Seventy-three miRNAs differentially expressed following ERβ expression in hormone-responsive human breast cancer cells

| *miRNA* | *Fold-change (ERβ+/ERβ−)* | P-*value* | *miRNA* | *Fold-change (ERβ+/ERβ−)* | P-*value* |
|---|---|---|---|---|---|
| **HS_108.1** | −1.58 | *0.03780* | hsa-miR-24-2* | 1.73 | *0.00002* |
| HS_131 | 1.46 | *0.00782* | hsa-miR-23b* | 2.07 | *0.00016* |
| HS_166.1 | −2.00 | *0.00001* | hsa-miR-27b* | 1.60 | *0.00024* |
| **HS_266.1** | −1.62 | *0.01176* | hsa-miR-24-1* | 2.06 | *0.00000* |
| HS_305_b | 1.84 | *0.00000* | hsa-miR-24-1*(miR-189:9.1) | 2.51 | *0.00000* |
| HS_99.1 | 1.39 | *0.00505* | hsa-miR-29a* | 1.50 | *0.00857* |
| hsa-let-7a | −1.50 | *0.00648* | hsa-miR-29b-2* | 1.41 | *0.00509* |
| hsa-let-7a* | −13.48 | *0.00000* | **hsa-miR-30a*** | −2.97 | *0.00012* |
| hsa-let-7c | −4.02 | *0.00000* | hsa-miR-30c-2* | −2.05 | *0.00003* |
| hsa-let-7f | −1.62 | *0.00510* | hsa-miR-30d* | 1.62 | *0.00046* |
| hsa-miR-100 | −4.57 | *0.00000* | hsa-miR-31 | 1.65 | *0.00435* |
| hsa-miR-101* | 1.53 | *0.00495* | hsa-miR-32* | −2.29 | *0.00000* |
| hsa-miR-1257 | 1.46 | *0.00331* | hsa-miR-330-5p | 2.04 | *0.00003* |
| hsa-miR-125b | −2.37 | *0.00000* | hsa-miR-335 | −4.61 | *0.00000* |
| hsa-miR-1267 | 1.40 | *0.00514* | hsa-miR-338-3p | 1.94 | *0.00000* |
| **hsa-miR-1285** | −1.64 | *0.00090* | hsa-miR-361-3p | −1.44 | *0.00998* |
| hsa-miR-1305 | 1.44 | *0.00254* | hsa-miR-362-5p | −1.67 | *0.00074* |
| hsa-miR-148b* | 1.41 | *0.00600* | hsa-miR-365 | −2.92 | *0.00000* |
| hsa-miR-15a* | 2.05 | *0.00097* | hsa-miR-374b* | −1.53 | *0.00680* |
| hsa-miR-16-1* | 1.98 | *0.00004* | hsa-miR-375 | 1.65 | *0.00274* |
| hsa-miR-17* | 1.51 | *0.00113* | **hsa-miR-450b-3p** | 1.69 | *0.00645* |
| hsa-miR-181c* | 1.70 | *0.00005* | hsa-miR-452*:9.1 | −1.98 | *0.00024* |
| hsa-miR-186 | −7.14 | *0.00000* | hsa-miR-501-5p | −1.79 | *0.00033* |
| hsa-miR-18a* | 1.40 | *0.01173* | hsa-miR-542-3p | 3.22 | *0.00000* |
| hsa-miR-18b | 1.78 | *0.00000* | **hsa-miR-548d-3p** | −1.97 | *0.00001* |
| hsa-miR-196b | −1.65 | *0.00101* | hsa-miR-556-5p | −4.07 | *0.00000* |
| hsa-miR-199a-5p | 1.78 | *0.00412* | hsa-miR-579 | −2.31 | *0.00079* |
| hsa-miR-199b-5p | 1.52 | *0.00574* | **hsa-miR-616*** | −1.56 | *0.00366* |
| hsa-miR-199a*:9.1 | 1.43 | *0.00283* | hsa-miR-629* | −1.67 | *0.00746* |
| hsa-miR-199a-3p/199b-3p | 1.72 | *0.00001* | hsa-miR-642 | 1.79 | *0.00144* |
| hsa-miR-19a | 1.29 | *0.04889* | hsa-miR-651 | 1.56 | *0.00148* |
| hsa-miR-19a* | 2.86 | *0.00000* | hsa-miR-652 | 1.54 | *0.00179* |
| hsa-miR-19b-1* | 1.95 | *0.00800* | **hsa-miR-663b** | −1.37 | *0.02954* |
| hsa-miR-20a* | 1.46 | *0.00291* | **hsa-miR-708*** | 2.05 | *0.00478* |
| **hsa-miR-216a** | −2.47 | *0.02759* | hsa-miR-935 | 1.53 | *0.00068* |
| hsa-miR-23a* | 1.93 | *0.00007* | hsa-miR-99b* | 1.56 | *0.00327* |
| hsa-miR-27a* | 2.10 | *0.00000* | | | |

Bold entries denote miRNAs differentially expressed also in ERβ-positive vs ERβ-negative breast tumor biopsies.

only in part the complexity of the hormone-responsive phenotype, in several cases provided molecular insights that could be validated and find application in the clinical setting. For this reason, we considered these evidences as an indication that ERβ might indeed influence miRNome activity also in primary breast tumors. To this end, BC samples were selected, among those originally included in the study reported by Novelli *et al*. (2008), for presence or absence of ERβ expression according to immuno-histochemistry (Supplementary Figure S5). Tumors were divided in two groups of 22 ERβ+ and 18 ERβ− tumors, respectively, that did not show significant differences from each other with respect to key clinical and molecular parameters, summarized in Supplementary Table S1, with the notable exception of the presence of lymphnodal metastases and a worst tumor grading for ERβ− tumors. RNA was extracted from formalin-fixed, paraffin-embedded tissues and that from 17 ERβ+ and 19 ERβ− tumors was of quality and concentration apt to perform miRNA expression profiling as described (Ravo *et al*., 2008). This led to the identification of 67 miRNAs, whose expression level discriminates ERβ+ from the ERβ− breast tumors, including 10 miRNAs that were found differentially expressed also in ERβ+ vs ERβ− BC cells *in vitro* (Figure 2c and Supplementary Table S2). These results confirm those obtained in cell lines (Figures 2a and b), pointing to a role of ERβ in the control of BC miRNome and thereby indicating that miRNAs are integral components of the gene regulation cascade mediating the effects of this nuclear receptor in tumor cells.

### *Direct regulation of miRNA biogenesis by hormone-activated ERβ in BC cells*

Mature miRNA expression can be regulated through control of either transcription or one of the key steps of primary transcript (pri-miR) maturation. We analyzed by chromatin immunoprecipitation sequencing (ChIP-Seq) the entire ERα and ERβ cistromes in the ERβ+ (Grober *et al*., 2011) and ERβ− cells (Cicatiello *et al*., 2010) upon E2 stimulation. Aligning ER-binding sites and miRNA gene positioning in the genome we observed that several miRNA-encoding genes differentially expressed in ERβ+ vs ERβ− cell lines (Supplementary Table S3A) and/or mammary tumors

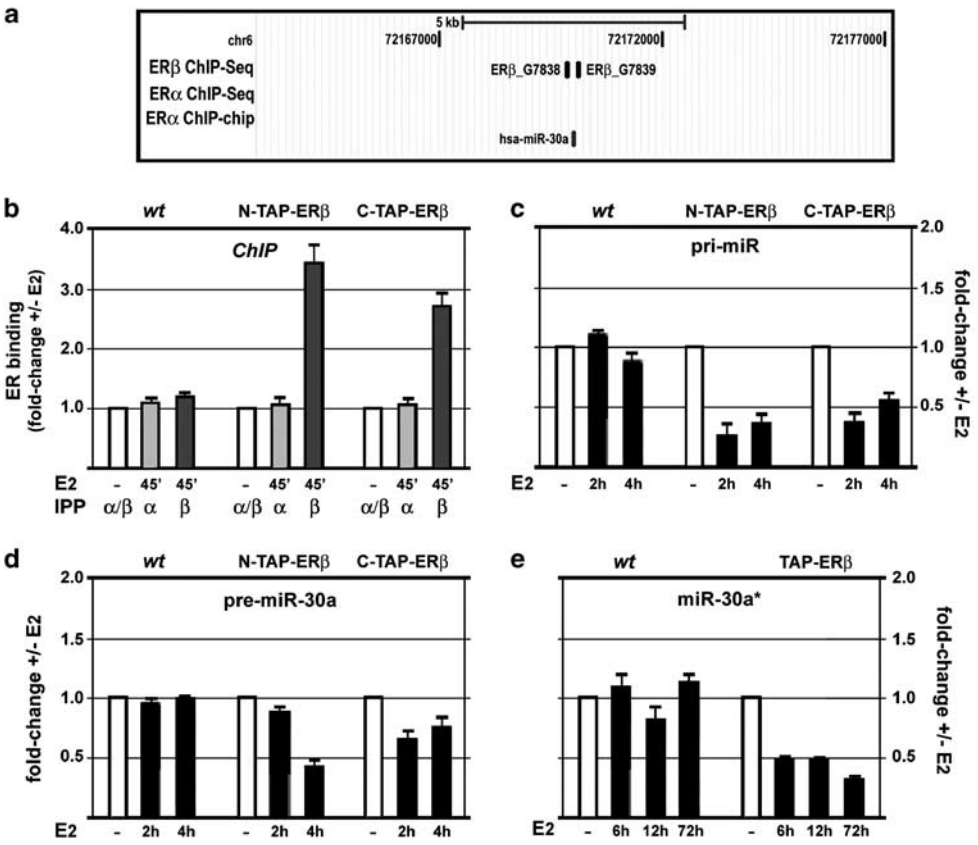



**Figure 3** Analysis of ERβ regulation of miR-30a and precursor biogenesis. (**a**) Genome browser view of the two ERβ-binding sites within 10 kb upstream or downstream from hsa-miR-30a locus on chromosome 6. (**b**) Validation of ERβ-binding site by ChIP and real-time PCR in *wt* or ERβ+ cells before (-) and after stimulation with E2 for 45 min. (**c**, **d**) Real-time rtPCR analysis of pri-miR-30a (**c**) and pre-miR-30a (**d**) in *wt* or ERβ+ cells before (-) and after stimulation with E2 for the indicated times. (**e**) Real-time rtPCR analysis of mature hsa-miR-30a* in *wt* or ERβ+ cells (C- and N-TAP-ERβ cell RNA combined) before (-) and after stimulation with E2 for the indicated times.

(Supplementary Table S3B) display ER-binding sites within 10 kb of the transcription unit, including sites where both ERs can be found together, likely associated in heterodimers. This finding suggested us the possibility that miRNA gene activity could be modulated in BC cells by an interplay of the two ER subtypes bound to chromatin, with ERβ antagonizing ERα-mediated regulation of pri-miR biosynthesis and/or maturation rate. To verify this possibility, we choose to investigate in detail differences in miRNA precursor levels in ERβ+ vs ERβ− cells following stimulation with E2, focusing on miR-30a gene and the miR-23b/27b/24-1 chromosomal cluster. The first was selected as it encodes two miRNAs (miR-30a and -30a*) that are downregulated by estrogen in ERβ+ cells only (Supplementary Figure S3) and it shows two binding sites for ERβ in close proximity —one upstream and one downstream— of the transcription unit, but no ERα sites (Figures 3a and b). The second caught our attention, instead, as it shows sites for both receptors (Figure 4a and Supplementary Figure S6) and it encodes three distinct couples of miRNAs, all accumulating in ERβ+ cells and decreasing in ERβ− cells in response to the hormone (Supplementary Figure S3). Interestingly, in both cases the effect of the hormone was more evident on the 'star' strand that, for this reason, led us first to their identification (Figure 2a) and was routinely used here to monitor ERβ effects.

The results relative to the miR-30a locus are reported in Figure 3 and show that ERβ binding results in a significant reduction of pri-, pre- and mature miR-30a levels following E2 stimulation, detectable already after 2 h (Figures 3c–e), to indicate that the predominant effect of ligand-activated ERβ is to *trans*-repress basal gene transcription by direct binding to this transcription unit. Noteworthy, activation of ERα alone (*wt* cells) did not affect miR-30a biogenesis, in agreement with the lack of binding of this receptor to the locus (Figure 3a). When combined, these results indicate a specific and direct role of ERβ in repression of miR-30a expression in BC cells, possibly mediated by promoter *trans*-repression. This could be due to direct transcriptional repression, via recruitment of a repressor complex to the chromatin by ligand-activated ERβ, or, alternatively, to inhibition of gene *trans*-activation caused by tethering of ERβ to a transcription factor constitutively bound to the locus, resulting in displacement or inhibition of an activator complex. The latter possibility, that could explain also lack of ERα binding to such regulatory site, is worth investigating further, extending the analysis

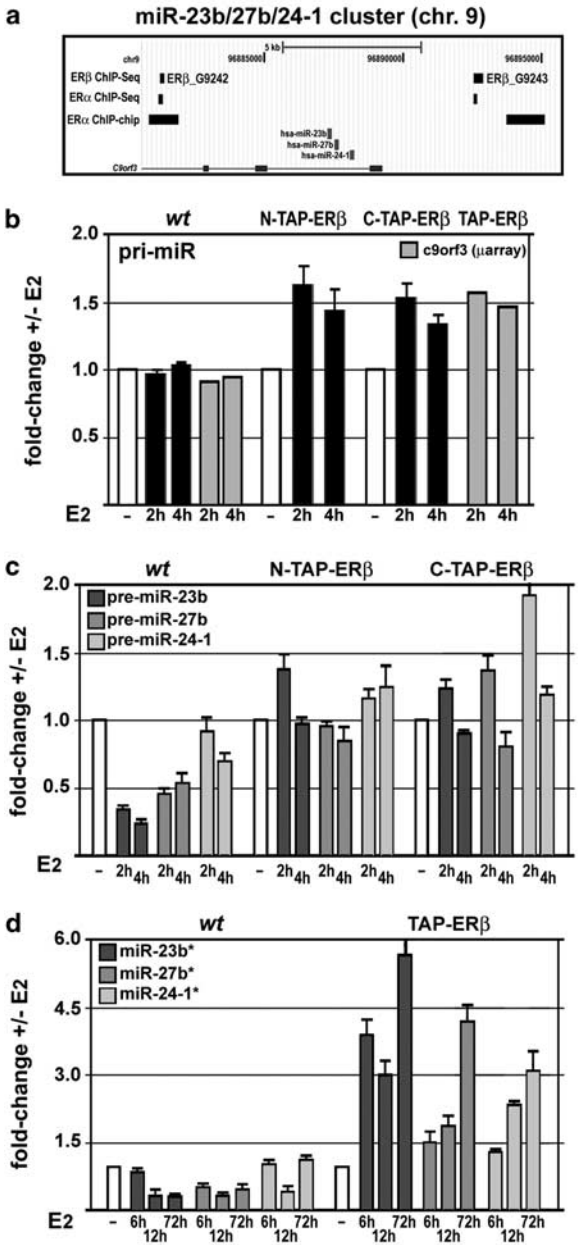



**Figure 4** Analysis of ERβ regulation of miR-23b/27b/24-1 and precursor biogenesis. (**a**) Genome browser view of ERβ- and ERα-binding sites within 10 kb upstream or downstream from miR-23b/27b/24-1 cluster within the c9orf3 locus on chromosome 9. (**b–d**) Real-time rtPCR analysis of the 23b/27B724-1 pri-miR (**b**), pre-miR (**c**) and mature miRNA* (**d**) in *wt* or ERβ+ cells (C- and N-TAP-ERβ cell RNA combined) before (-) and after stimulation with E2 for the indicated times. The right columns in (**b**) show the relative expression of c9orf3 RNA in ERβ+ cells following E2 stimulation, measured by mRNA expression profiling (Grober *et al*., 2011).

also to other genetic loci selectively regulated by ERβ in BC cells under the same conditions.

Our attention focused next on the miR-23b/27b/24-1 cluster on chromosome 9, whose organization is showed in Figure 4a. In this case, both ERβ- and ERα-binding sites are detected. Noteworthy, the two ERα-binding sites identified by ChIP-Seq were also found by ChIP-on-chip in an independent study (Hurtado *et al*., 2008) and binding of the two ERs to both sites identified here in ERβ+ cells was confirmed by ChIP (Supplementary Figure S6). The effects of ERβ in regulation of the first step in miRNA biogenesis were investigated by measuring changes in pri-miR expression in control (*wt*), N-TAP- and C-TAP-ERβ cells before and after E2 stimulation. Results show that in the absence of ERβ, estrogen stimulation did not influence primary-transcript levels, assessed by both quantitative real-time rtPCR and RNA-expression profiling (*c9orf3* RNA; Figure 4b and Cicatiello *et al*., 2010). On the other hand, a slight but reproducible accumulation of pri-miR-23b/-27b/-24-1 was detectable in ERβ+ cells already 2 h after E2 (Figure 4b). We next measured the intracellular concentration of the individual pre-miR deriving from this primary transcript (pre-miR-23b, -27b and -24-1) in both cell types and the results obtained were surprisingly very different. Indeed, as shown in Figure 4c, whereas stimulation with E2 of ERβ− cells caused a substantial loss of pre-miR (ranging from −20 to −75%), the same treatment caused instead accumulation of these pre-miRs in ERβ+ cells. This was reflected in comparable changes in expression of the corresponding mature miRNAs for up to 72 h after E2 stimulation (Figure 4d and Supplementary Figure S3). These results indicate that the presence of ERβ in ERα-expressing, estrogen-responsive BC cells can modify substantially the response of miRNA genes to hormonal stimulus. In the case of the miR-23b/27b/24-1 gene cluster, this results from changes in pri-miR maturation, rather than synthesis, leading to increase in pre-miR biosynthesis in the presence of chromatin-bound ERβ.

*ERβ interferes with ERα-mediated recruitment of Drosha in inactive chromatin-bound complexes*

Yamagata *et al*. (2009) reported ERα-mediated regulation of miRNA maturation by direct interaction in the nucleus of ERα with a protein complex comprising Drosha and the DEAD box RNA helicase p68/DDX5, resulting in inhibition of pri- to pre-miRNA conversion by Drosha. We thus considered the possibility that the enhancing effect of ERβ on pri-miR-23b/-27b/-24-1 maturation shown in Figure 4 could result from competition for binding of the ERα-p68-Drosha complex to this locus by ERβ, as recently described for other target genes (Grober *et al*., 2011), thereby preventing the inhibitory effect of ERα on nascent pri-miR maturation. In both cases, we should expect inhibition of ERα-mediated p68/DDX5-Drosha recruitment to miR-23b/-27b/-24-1 chromatin by ERβ. Indeed, this appears to be the case, as E2-induced p68/DDX5 and Drosha binding to ERβ-G9242 and ERβ-G9242 chromatin sites was strongly reduced in C-TAP-ERβ compared with *wt* cells, concomitant with a reduction of ERα and appearance of ERβ (upper panel of Figure 5a and data not shown). It is worth mentioning that binding of ERα to chromatin in ERβ+ cells occurs mainly via heterodimerization with ERβ (Grober *et al*., 2011). ERβ-mediated inhibition of p68/DDX5 binding could be observed also at the ERβ-G5984 site of the TFF1/pS2 gene promoter,

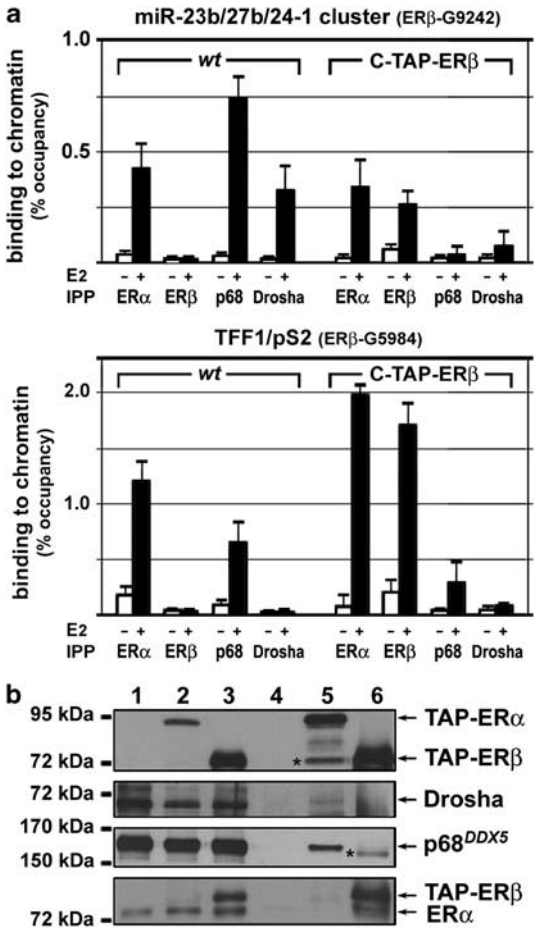



**Figure 5** ER/Drosha interaction in MCF-7 cell nuclei. ChIP real-time PCR results showing binding of ERα, ERβ, p68 and Drosha to miRNA 23b/27b/24-1 cluster and the TFF1/pS2 loci (**a**) in the TFF1/pS2 loci (**b**) in *wt* and C-TAP-ERβ cells; data are expressed as % occupancy respect to input chromatin. (**b**) Western blot analysis of whole nuclear extracts (lanes 1–3) and IgG-Sepharose-affinity-purified nuclear extracts (lanes 4–6) from *wt* (lanes 1 and 4), C-TAP-ERα (lanes 2 and 5) or C-TAP-ERβ (lanes 3 and 6) cells, probed with the indicated antibodies. Asterisks mark non-specific bands.

although Drosha could not be detected tethered to this site under any condition (lower panel of Figure 5b), suggesting that association of this enzyme to chromatin may be promoted by ERα only at sites of pri-miR synthesis, where Drosha could be 'locked' in an inactive complex comprising ERα and the hairpin structure of the nascent pri-miR. Concerning the nature of the physical interaction between ERα and Drosha, it was suggested that this is mediated by p68/p72 RNA helicases (Yamagata *et al*., 2009). Interestingly, a systematic analysis of the ERβ interactome of MCF-7 cell nuclei (Nassa *et al*., 2011) failed to identify p68/DDX5 binding to this receptor subtype as well as to ERα/ERβ heterodimers, suggesting that the presence of ERβ could determine inhibition of p68/DDX5-mediated sequestering of Drosha to the chromatin in an inhibitory complex. This possibility would provide a rationale for the ChIP results obtained in ERβ+ cells, where we failed to detect these two proteins in the presence of both ERs (Figure 5a). To verify this possibility, we performed co-purification analysis of all these proteins in nuclear extracts from *wt*, C-TAP-ERα or C-TAP-ERβ cells. The two ERs were adsorbed to Sepharose-bound IgG via their TAP tag, as described (Ambrosino *et al*., 2010; Nassa *et al*., 2011). As shown in Figure 5b, Drosha and p68/DDX5 could be co-purified with ERα but not with ERβ, demonstrating that ERβ is unable to bind these proteins. It is worth mentioning that as under these experimental conditions ERα co-purifies with C-TAP-ERβ (Nassa *et al*., 2011 and lower section of Figure 5b), ERα/ERβ heterodimers do not bind Drosha and p68/DDX5.

## Discussion

The results described here demonstrates that ERβ controls synthesis, maturation and steady-state levels of a significant number of miRNAs in BC cells by interfering with ERα activity or acting autonomously, as demonstrated here for the miR-23b/-27b/-24-1 cluster and the miR-30a gene, respectively. This, in turn, determines a profound effect on miRNome expression and activity in tumors expressing ERβ, which could help explain their less aggressive clinical phenotype (Novelli *et al*., 2008; Shaaban *et al*., 2008). Identification of the intracellular targets of these ERβ-regulated miRNAs, and the effects they exert on key cellular functions of BC cells, will now provide a new venue to understand the pleiotropic role of this oncosuppressive factor in breast carcinogenesis and tumor progression. Furthermore, it is reasonable to conceive that proteins encoded by the mRNAs targeted by these miRNA may represent molecular markers exploitable for prognostic evaluation of primary breast tumors or for prediction of the disease responsiveness to hormonal therapy.

## Materials and methods

*Cell Culture, transient transfection and cell cycle analyses*
Human hormone-responsive BC cells MCF-7 Tet-Off (Clontech-Takara) expressing TAP (control cells), C-TAP-ERα, C-TAP-ERβ or N-TAP-ERβ were described previously (Ambrosino *et al*., 2010; Nassa *et al*., 2011). They were propagated, hormone starved and analyzed for estrogen signaling, cell cycle progression and cell proliferation as described earlier (Cicatiello *et al*., 2000; Grober *et al*., 2011).

*RNA purification*
Total RNA was extracted from hormone-starved (+EtOH, −E2) or stimulated ($+E_2$) cell cultures as described previously (Cicatiello *et al*., 2004). FFPE tumor samples were cut in 5-μm-thick sections on a microtome with a disposable blade. RNA was extracted from three and eight sequential sections as described (Ravo *et al*., 2008). RNA concentration in each sample was determined with a NanoDrop-1000 spectrophotometer (Thermo Fisher Scientific Italy, Cinisello Balsamo, Italia) and quality assessed with the Agilent 2100 Bioanalyzer and Agilent RNA 6000 cartridges (Agilent Technologies). For microarray analysis, RNAs extracted from replicate samples of the same tumor were pooled.

*Microarray analyses*
See Supplementary Materials and methods.

*Protein-complex immunoprecipitations and analysis*
Cells were hormone starved for 5 days and following stimulation with $10^{-8}$ M E2 for 2 h, nuclear proteins were extracted and incubated with IgG-Sepharose beads (GE Healthcare, Milano, Italy) for 4 h at 4 °C, as described earlier (Ambrosino *et al*., 2010). Affinity-purified complexes were resuspended in SDS sample buffer (Invitrogen Life Technologies Italia, Milano, Italy) and analyzed by SDS–PAGE and western blotting by using anti-TAP (CAB1001, Open Biosystems, Euroclone Spa, Milano, Italy), anti-ERα (sc-543, Santa Cruz Biotechnology), anti-Drosha (ab12286, Abcam, Cambridge, UK) and anti-DDX5 (ab21696, Abcam) antibodies. The primary antibodies were detected with a horseradish peroxidase-conjugated anti-rabbit antibody (GE Healthcare) and revealed by chemiluminescence and autoradiography.

*Chromatin immunoprecipitation*
Cells were hormone deprived for 4 days and chromatin was extracted from replicate samples before (−E2) or 45 min after stimulation with E2 as described previously (Cicatiello *et al*., 2010; Grober *et al*., 2011). Chromatin samples were incubated at 4 °C overnight with Abs against the C- (HC-20, from Santa Cruz Biotechnology, Europe) or N- (18–32, Sigma Aldrich Italia, Milano, Italy) terminus of human ERα, anti-Drosha (ab12286, Abcam, used as described by Nakamura *et al*. (2007), anti-DDX5 (ab21696, Abcam) or, for TAP-ERβ, with IgG Sepharose 6 fast Flow (GE Healthcare) as described earlier (Grober *et al*., 2011). As control, aliquots of the same chromatin were processed in the same way but Abs were omitted from the incubation mixtures (+E2/-Abs) or, where required, underivatized Sepahrose was used.

*Quantitative real-time rtPCR*
Total RNA was extracted from cell lines (as described before) after stimulation for 2 h and 4 h with$10^{-8}$ M E2. For miRNA analysis, mature miRNA was reverse transcribed using a miRNA-specific stem-loop reverse transcriptase and real-time PCR was performed using Taqman microRNA assays (Assay ID: 2822, 416, 2439, 2445, 2441, 2126, 2174, 2440, 2333, 482; Applied Biosystems Italia, Monza, Italy) according to the manufacturer's instruction. RNU49 was used as an internal control to normalize all data using the Taqman RNU49 assay (Applied Biosystems Italia). RNU49 was unaffected by hormone treatment. For pre-miRNA and pri-miRNA analysis, RNA was reverse transcribed using Quantitect Rev. Transcription kit (Qiagen Italy, Milano, Italia) and real-time PCR was performed in triplicates in three independent experiments using Power Syber Green PCR Master Mix (Applied Biosystems Italia) and normalized to U6 snRNA. All the real-time PCR were performed on a MJ Research PTC-200 Opticon Instrument (MJ Research, Waltham, MA, USA). Primers used are listed in Supplementary Table 4.

*ChIP-Seq data analysis*
For ER-binding-site mapping in genome, ChIP-Seq data relative to ERβ (Grober *et al*., 2011; accession number E-MTAB-345) and ERα (Cicatiello *et al*., 2010; accession number E-MTAB-131) were analyzed as follows. Enriched ChIP-Seq peaks were identified using FindPeaks (Fejes *et al*., 2008), with a subpeaks value of 0.5. To select only highly relevant sites, the statistical cut-off of the first quartile was applied. The binding sites supported by a number of tags lower than 25% of the range of the values was discarded. This led to re-mapping of ERβ-binding sites (renumbered here from ERβ_G1 to ERβ_G12430); for ERα-binding sites, numbering was the same as previously described (Cicatiello *et al*., 2010).

*miRNA target prediction and functional analysis of their putative mRNA targets*
For comprehensive prediction of miRNA-target genes, we used TargetScan, release 5.1 (http://www.targetscan.org). To identify statistically overrepresented 'biological process' gene ontology terms among sets of selected mRNA target, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) functional annotation tool (Dennis *et al*., 2003; Huang *et al*., 2009). To this aim, we used as background data coming from gene expression-profiling experiments performed on the same cell line and under the same experimental conditions used in this study.

*Immunohistochemistry*
See: Supplementary Materials and methods.

*BC samples clinical hallmarks*
For the purpose of this study, 40 breast carcinomas were selected from a series of 936 cases with a median follow up (FU) of 50 months (min 1–max 108) subjected to breast surgery at the Regina Elena Cancer Institute between 2001 and 2005 (Novelli *et al*., 2008). Of these, 22 were ERβ+ without any recurrence, whereas 18 were ERβ− and presented local or distant metastasis. In these patients, ERβ expression was routinely determined at the time of surgical treatment along with other conventional biological factors namely ERα and progesterone receptors (PgR), HER2 and Ki-67, before any adjuvant therapy was planned. As showed in Supplementary Table S1, the group included 37 (92.5%) invasive ductal carcinomas and 3 (7.5%) invasive lobular carcinomas. Among these, 28 (70%) were pT1, 9 (22.5%) pT2 and 3 (7.5%) pT3-4, 27 (67.5%) were node negative and 13 (32.5%) were node positive, 29 (72.5%) G1-2 and 11 (27.5%) G3. ERα was positive in 37 tumors (92.5%) and negative in 3 (7.5%), PgR was positive in 31 tumors (77.5%) and negative in 9 (22.5), HER2 was positive in 12 tumors (30%) and negative in 28 (70%) and Ki-67 was positive in 16 tumors (40%) and negative in 24 (60%). Tumors were graded according to Bloom and Richardson and staged according to the Unione Internationale Contre le Cancer tumor-node-metastasis system criteria, and histologically classified according to the World Health Organization (Tavassoli and Devilee, 2003). In the selected group, ERβ+ was significantly associated to negative lymphnodes ($P<0.0001$) and low tumor grade (G1-2) ($P=0.03$) whereas, as already described on a large series of BC patients, (Novelli *et al*., 2008) no significant correlation was observed between ERβ expression and the other parameters analyzed. Follow-up data were obtained from hospital charts and by corresponding with the referring physicians.

## Conflict of interest

The authors declare no conflict of interest.

## Acknowledgements

We thank Manuela Ferracin and Massimo Negrini for help with analysis of miRNA expression data from primary tumor

samples and useful suggestions, and Luigi Cicatiello, Margherita Mutarelli and Maria Francesca Papa for technical assistance. Work supported by: Italian Association for Cancer Research (grants IG-8586 to AW and IG-8706 to MM), European Union (CRESCENDO IP, contract number LSHM-CT2005-018652), Italian Ministry for Education, University and Research (grant PRIN 2008CJ4SYW_004), University of Salerno (Fondi FARB 2011) and Fondazione con il Sud (grant 2009-PdP-22). CC, FR, GN, MR and RT are fellows of Fondazione con il Sud.

## References

Ambrosino C, Tarallo R, Bamundo A, Cuomo D, Franci G, Nassa G *et al*. (2010). Identification of a hormone-regulated dynamic nuclear actin network associated with estrogen receptor alpha in human breast cancer cell nuclei. *Mol Cell Proteomics* **9**: 1352–1367.

Bardin A, Boulle N, Lazennec G, Vignon F, Pujol P. (2004). Loss of ERbeta expression as a common step in estrogen-dependent tumor progression. *Endocr Relat Cancer* **11**: 537–551.

Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS *et al*. (2009). Estradiol-regulated microRNAs control estradiol response in breast cancer cells. *Nucleic Acids Res* **37**: 4850–4861.

Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin SF, Dunning MJ *et al*. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol* **8**: R214.

Castellano L, Giamas G, Jacob J, Coombes RC, Lucchesi W, Thiruchelvam P *et al*. (2009). The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response. *Proc Natl Acad Sci USA* **106**: 15732–15737.

Chang EC, Frasor J, Komm B, Katzenellenbogen BS. (2006). Impact of estrogen receptor beta on gene networks regulated by estrogen receptor alpha in breast cancer cells. *Endocrinology* **147**: 4831–4842.

Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K *et al*. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**: 740–744.

Cicatiello L, Addeo R, Altucci L, Belsito Petrizzi V, Boccia V, Cancemi M *et al*. (2000). The antiestrogen ICI 182 780 inhibits proliferation of human breast cancer cells by interfering with multiple, sequential estrogen-regulated processes required for cell cycle completion. *Mol Cell Endocrinol* **165**: 199–209.

Cicatiello L, Mutarelli M, Grober OM, Paris O, Ferraro L, Ravo M *et al*. (2010). Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. *Am J Pathol* **176**: 2113–2130.

Cicatiello L, Scafoglio C, Altucci L, Cancemi M, Natoli G, Facchiano A *et al*. (2004). A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. *J Mol Endocrinol* **32**: 719–775.

Croce CM. (2009). Causes and consequences of microRNA dysregulation in cancer. *Nat Rev Genet* **10**: 704–714.

Dennis Jr G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC *et al*. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3.

Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJ. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**: 1729–1730.

Ferraro L, Ravo M, Nassa G, Tarallo R, De Filippo MR, Giurato G *et al*. (2011). Effects of estrogen on microRNA expression in hormone-responsive breast cancer cells. *Horm Cancer* (in press).

Ferraro L, Ravo M, Paris O, Giurato G, De Filippo MR, Stellato C *et al*. (2010). Identification of a gene network controlled by estrogen receptor α in human breast cancer cells by miRNA expression profiling and ChIP-Seq. *Am J Pathol* **177**: S34, Abs. STE 12.

Garzon R, Calin GA, Croce CM. (2009). MicroRNAs in cancer. *Annu Rev Med* **60**: 167–179.

Grober OMV, Mutarelli M, Giurato G, Ravo M, Cicatiello L, De Filippo MR *et al*. (2011). Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals extensive interplay with estrogen receptor alpha for target gene regulation. *BMC Genomics* **12**: 36.

Han J, Lee Y, Yeom KH, Kim YK, Jin H, Kim VN. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18**: 3016–3027.

Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Hartman J *et al*. (2007). Estrogen receptors: how do they signal and what are their targets. *Physiol Rev* **87**: 905–931.

Hill DA, Ivanovich J, Priest JR, Gurnett CA, Dehner LP, Desruisseau D *et al*. (2009). DICER1 mutations in familial pleuropulmonary blastoma. *Science* **325**: 965.

Huang W, Sherman BT, Lempicki RA. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* **4**: 44–57.

Hurtado A, Holmes KA, Geistlinger TR, Hutcheson IR, Nicholson RI, Brown M *et al*. (2008). Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen. *Nature* **45**: 663–666.

Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S *et al*. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res* **65**: 7065–7070.

Kim VN, Han J, Siomi MC. (2009). Biogenesis of small RNAs in animals. *Nat Rev Mol Cell Biol* **10**: 126–139.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D *et al*. (2005). MicroRNA expression profiles classify human cancers. *Nature* **435**: 834–838.

Maillot G, Lacroix-Triki M, Pierredon S, Gratadou L, Schmidt S, Bénès V *et al*. (2009). Widespread estrogen-dependent repression of micrornas involved in breast tumor cell growth. *Cancer Res* **69**: 8332–8340.

Melo SA, Moutinho C, Ropero S, Calin GA, Rossi S, Spizzo R *et al*. (2010). A genetic defect in exportin-5 traps precursor microRNAs in the nucleus of cancer cells. *Cancer Cell* **18**: 303–315.

Melo SA, Ropero S, Moutinho C, Aaltonen LA, Yamamoto H, Calin GA *et al*. (2009). A TARBP2 mutation in human cancer impairs microRNA processing and DICER1 function. *Nat Genet* **41**: 365–370.

Nakamura T, Canaani E, Croce CM. (2007). Oncogenic All1 fusion proteins target Drosha-mediated microRNA processing. *Proc Natl Acad Sci USA* **104**: 10980–10985.

Nassa G, Tarallo R, Ambrosino C, Bamundo A, Ferraro L, Paris O *et al*. (2011). A large set of estrogen receptor β-interacting proteins identified by tandem affinity purification in hormone-responsive human breast cancer cell nuclei. *Proteomics* **11**: 159–165.

Newman MA, Hammond SM. (2010). Emerging paradigms of regulated microRNA processing. *Genes Dev* **24**: 1086–1092.

Novelli F, Milella M, Melucci E, Di Benedetto A, Sperduti I, Perrone-Donnorso R *et al*. (2008). A divergent role for estrogen receptor-beta in node-positive and node-negative breast cancer classified according to molecular subtypes: an observational prospective study. *Breast Cancer Res* **10**: R74.

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC. (2008). The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nat Struct Mol Biol* **15**: 354–363.

Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, Bragado-Nilsson E *et al*. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**: 218–229.

Ravo M, Mutarelli M, Ferraro L, Grober OMV, Paris O, Tarallo R *et al*. (2008). Quantitative expression profiling of highly degraded

RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays. *Lab Invest* **88**: 430–440.

Roger P, Sahla ME, Mäkelä S, Gustafsson JA, Baldet P, Rochefort H. (2001). Decreased expression of estrogen receptor beta protein in proliferative preinvasive mammary tumors. *Cancer Res* **61**: 2537–2541.

Shaaban AM, Green AR, Karthik S, Alizadeh Y, Hughes TA, Harkins L *et al*. (2008). Nuclear and cytoplasmic expression of ERbeta1, ERbeta2, and ERbeta5 identifies distinct prognostic outcome for breast cancer patients. *Clin Cancer Res* **14**: 5228–5235.

Sugiura H, Toyama T, Hara Y, Zhang Z, Kobayashi S, Fujii Y *et al*. (2007). Expression of estrogen receptor beta wild-type and its variant ERbetacx/beta2 is correlated with better prognosis in breast cancer. *Jpn J Clin Oncol* **37**: 820–828.

Tavassoli FA, Devilee P. (2003). *World Health Organization Classification of Tumors. Pathology and Genetics of Tumours of the Breast and Female Genital Organs*. IARC Press: Lyon.

Tavazoie SF, Alarcón C, Oskarsson T, Padua D, Wang Q, Bos PD *et al*. (2008). Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* **451**: 147–152.

Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F *et al*. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci USA* **103**: 2257–2261.

Yamagata K, Fujiyama S, Ito S, Ueda T, Murata T, Naitou M *et al*. (2009). Maturation of microRNA is hormonally regulated by a nuclear receptor. *Mol Cell* **36**: 340–347.

Zhang B, Pan X, Cobb GP, Anderson TA. (2007). microRNAs as oncogenes and tumor suppressors. *Dev Biol* **302**: 1–12.

Supplementary Information accompanies the paper on the Oncogene website (http://www.nature.com/onc)

BMC Genomics

RESEARCH ARTICLE Open Access

# Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation

Oli MV Grober[1†], Margherita Mutarelli[1†], Giorgio Giurato[1], Maria Ravo[1,2], Luigi Cicatiello[1], Maria Rosaria De Filippo[1], Lorenzo Ferraro[1], Giovanni Nassa[1,2], Maria Francesca Papa[1], Ornella Paris[1], Roberta Tarallo[1,2], Shujun Luo[3], Gary P Schroth[3], Vladimir Benes[4], Alessandro Weisz[1,2*]

## Abstract

**Background:** Estrogen receptors alpha (ERα) and beta (ERβ) are transcription factors (TFs) that mediate estrogen signaling and define the hormone-responsive phenotype of breast cancer (BC). The two receptors can be found co-expressed and play specific, often opposite, roles, with ERβ being able to modulate the effects of ERα on gene transcription and cell proliferation. ERβ is frequently lost in BC, where its presence generally correlates with a better prognosis of the disease. The identification of the genomic targets of ERβ in hormone-responsive BC cells is thus a critical step to elucidate the roles of this receptor in estrogen signaling and tumor cell biology.

**Results:** Expression of full-length ERβ in hormone-responsive, ERα-positive MCF-7 cells resulted in a marked reduction in cell proliferation in response to estrogen and marked effects on the cell transcriptome. By ChIP-Seq we identified 9702 ERβ and 6024 ERα binding sites in estrogen-stimulated cells, comprising sites occupied by either ERβ, ERα or both ER subtypes. A search for TF binding matrices revealed that the majority of the binding sites identified comprise one or more Estrogen Response Element and the remaining show binding matrixes for other TFs known to mediate ER interaction with chromatin by tethering, including AP2, E2F and SP1. Of 921 genes differentially regulated by estrogen in ERβ+ *vs* ERβ- cells, 424 showed one or more ERβ site within 10 kb. These putative primary ERβ target genes control cell proliferation, death, differentiation, motility and adhesion, signal transduction and transcription, key cellular processes that might explain the biological and clinical phenotype of tumors expressing this ER subtype. ERβ binding in close proximity of several miRNA genes and in the mitochondrial genome, suggests the possible involvement of this receptor in small non-coding RNA biogenesis and mitochondrial genome functions.

**Conclusions:** Results indicate that the vast majority of the genomic targets of ERβ can bind also ERα, suggesting that the overall action of ERβ on the genome of hormone-responsive BC cells depends mainly on the relative concentration of both ERs in the cell.

## Background

Estrogens are key regulators of cell growth and differentiation in the mammary gland [1,2], where they are involved in the pathogenesis and clinical outcome of breast cancer (BC) [3]. These steroid hormones exert their effects in normal and transformed mammary epithelial cells by binding to specific receptors, ERα and ERβ, that mediate estrogen signaling by functioning as ligand-dependent transcription factors. Ligand-activated ERs drive gene cascades comprising primary genes, whose transcription is directly controlled by the hormone through physical interaction of ERs with regulatory sites in the genome (genomic pathway) and/or with signal transduction effectors (non genomic pathway), as well as

* Correspondence: alessandro.weisz@unina2.it
† Contributed equally
[1]Department of General Pathology, Second University of Naples, vico L. De Crecchio 7, 80138 Napoli, Italy
Full list of author information is available at the end of the article

downstream genes whose activity depends upon the functions encoded by the primary responders [1,4].

ERs are able to bind DNA at specific sites in the genome and thereby control gene activity by recruiting transcriptional mediators and co-regulators, as well as a host of other nuclear proteins with different roles in ER-mediated control of gene activity [5,6]. The two ERs show 55% identity in their estrogen-binding domains (LBDs) and approximately 97% similarity in the DNA-binding domains (DBDs) [7]. Reflecting the high degree of similarity in their DBDs, both receptors interact with the same conserved estrogen response element (ERE) (5'-GGTCAnnnTGACC-3') as either homodimers or alpha/beta heterodimers [8,9]. ERβ, however, holds low trans-acting capability on ERE-containing estrogen target genes and alpha/beta heterodimers are less efficient than ERα homodimers in promoting target genes activity [10]. The different behaviour of ERα/ERβ heterodimers respect to ERα homodimers on transcriptional regulation of ERE-containing genes might be explained by different co-factor recruitment, as ERβ could prevent efficient co-activator binding to the ERα moiety of the heterodimer, conversely inducing recruitment of co-repressors and/or driving assembly of co-regulatory complexes other than those involving ERα only. [8,11,12].

Although the two receptors are quite similar in sequence and structure, in BC ERβ has considerably different biological effects than ERα [1,13,14]. Furthermore, the two ERs show a remarkably different expression pattern in BCs, with higher ERα and lower ERβ levels observed in malignant cells compared to normal mammary epithelial or benign tumor cells [15,16]. Furthermore, while ERα induces a mitogenic response to estrogen, when expressed alone the β subtype is not only unable to induce the same mitogenic response, but it reduces basal, hormone-independent cell proliferation [17-18 and R. Tarallo *et al.*, unpublished]. Finally, ERβ was shown to change dramatically ERα-positive BC cell behaviour *in vivo*, as its expression in the cell prevents tumorigenicity in mouse xenograft models by reducing tumor growth and angiogenesis [19,20].

Gene expression studies performed in BC cell lines expressing endogenous ERα and recombinant ERβ [21-23] revealed multiple signaling pathways involving the α and/or β receptor subtypes [1]. The two ERs appear thus to share many target genes, although each of them may affect specific downstream targets. For this reason, inhibition of hormone-responsive BC cell growth by ERβ might be due to direct interference with ERα activity on growth-promoting pathways as well as to the activity of ERβ-specific target genes [24].

Recently, next-generation sequencing technologies combining chromatin immunoprecipitation (ChIP) either with genomic DNA hybridization to microarrays (ChIP-on-chip) or massively parallel sequencing (ChIP-Seq, ChIP-PET), opened new venues for our understanding of physical and functional associations between transcription factors and chromatin *in vivo*. These analytical strategies led to genome-wide mapping of ERα-binding regions in intact chromatin of cultured cell lines [25-28], revealing important new information relative to ERα interaction with the genome. Carroll *et al.* [25], for example, using ChIP-on-chip demonstrated that the Forkhead factor FoxA1 plays an important role as pioneering factor for ERα binding to chromatin in BC cells, while Cicatiello *et al.* [26] identified novel gene regulation cascades mediating estrogen actions in hormone-responsive BC cells. In contrast, although several studies focused on ERβ interaction with the genome [29-32], a thorough characterization of this important aspect of ERβ biology in BC cells, essential to clarify the mechanisms mediating its control of estrogen-dependent gene pathways and the hormone-responsive phenotype, is still missing. For this reason, we performed a comprehensive analysis of ERβ and ERα target sites in the genome of MCF-7 cells engineered to express both receptors to comparable levels, by integrating global mapping of *in vivo* ER binding to the genome by ChIP-Seq with comparative gene expression profiling in ERβ+/ERα+ *vs* ERβ-/ERα+ cells during early stimulation with 17-beta-estradiol (E2), followed by *in silico* analyses of the ERβ binding regions and responsive genes identified.

## Results and Discussion

### Establishment and characterization of ERβ-expressing MCF-7 cells

Stabilized human BC cell lines expressing endogenous ERα and ERβ at comparable levels are not available. For this reason, we first generated and characterized cell lines derived from ERα-positive MCF-7 cells expressing full-length human ERβ (ERβ1) at levels similar to those of endogenous ERα. This strategy was adopted to prevent artefacts due to ERβ over-expression in the cell and to mimic what observed in primary breast tumors, where very high expression of this receptor has never been observed. As suitable antibodies for efficient immunoprecipitation of chromatin-bound ERβ are not available, the expressed proteins were tagged on either their C- (Ct-ERβ) or N-terminus (Nt-ERβ) with the TAP epitope. This approach allows to track tagged ERs in different cell compartments and to efficiently immunoprecipitate and purify them *in vitro* by Tandem Affinity Purification, to identify their molecular partners [5,33], and *ex vivo* in chromatin immunoprecipitation assays (see below). Preliminary tests were performed to verify whether the presence of the TAP moiety could influence intracellular redistribution of ERβ in response to 17β-estradiol (E2) and its ability either to *trans*-activate an

estrogen-responsive reporter gene or to interfere with ERα activity on reporter gene transcription and BC cell proliferation. To this end, ER expression and nuclear translocation in response to E2 was determined in *wt* MCF-7, Ct-ERβ and Nt-ERβ cells by subcellular fractionation followed by SDS-PAGE and immunoblotting (Figure 1A). In absence of hormone a larger fraction of both ERs was found in the cytosol in all cases. Following estrogen stimulation, both receptors migrated to the nucleus, a crucial event to trigger target gene transcription *via* the genomic pathway of the estrogen signaling cascade. An antibody against α-tubulin was used as control, and the absence of this protein in the nuclear fractions indicates that they were indeed free from cytosolic contaminants. The ERβ-expressing clones selected for this study showed a ERβ/ERα ratio <2, as verified by immunoblotting analysis of the proteins in whole cell extracts and quantitative rtPCR of the corresponding RNAs [5, and data not shown]. To control that the presence of the TAP tag did not interfere with ERβ activity, ER-negative SKBR-3 BC cells were transiently transfected with expression vectors encoding *wt* ERβ, Ct-ERβ, Nt-ERβ, ERα (HEG0) or 'empty' vector (pSG5), as controls, and ERE-*tk*-luc [34], a reporter gene where luciferase expression is driven by an estrogen-responsive minimal promoter. Exposure of transiently transfected cells to E2 induced reporter gene activation in the presence of ERα, ERβ, Ct-ERβ or Nt-ERβ, with the activity of both tagged ERβ proteins slightly (15-20%) lower than that of *wt* ERβ (Figure 1B, left). We then tested whether the two recombinant forms of ERβ were able to interfere with target gene activation by the endogenous ERα resident in MCF-7 cells. To this end, *wt*, Ct-ERβ+ and Nt-ERβ+ cells were transfected with ERE-*tk*-luc and the response of the reporter gene to E2 was determined. As shown in Figure 1B (right), ERβ-expressing cells showed in all cases a marked (50-60%) reduction in reporter gene response to the hormone, when compared to *wt* MCF-7 cells, indicating that both tagged ERβs are able to interfere with the activity of endogenous ERα. Results show that cell lines stably expressing Ct-ERβ and Nt-ERβ display a marked reduction in proliferative response to the hormone, when compared to *wt* MCF-7 cells (Figure 1C), in agreement with the known effects of ERβ in ERα-positive cells [23,35-37]. Furthermore, comparative RNA expression profiling in exponentially growing Ct-ERβ and Nt-ERβ *vs wt* MCF-7 cells revealed extensive overlapping effects of the two tagged ERβ proteins on the activity of estrogen target genes [O. Paris *et al.*, manuscript in preparation and data not shown]. Taken together, these

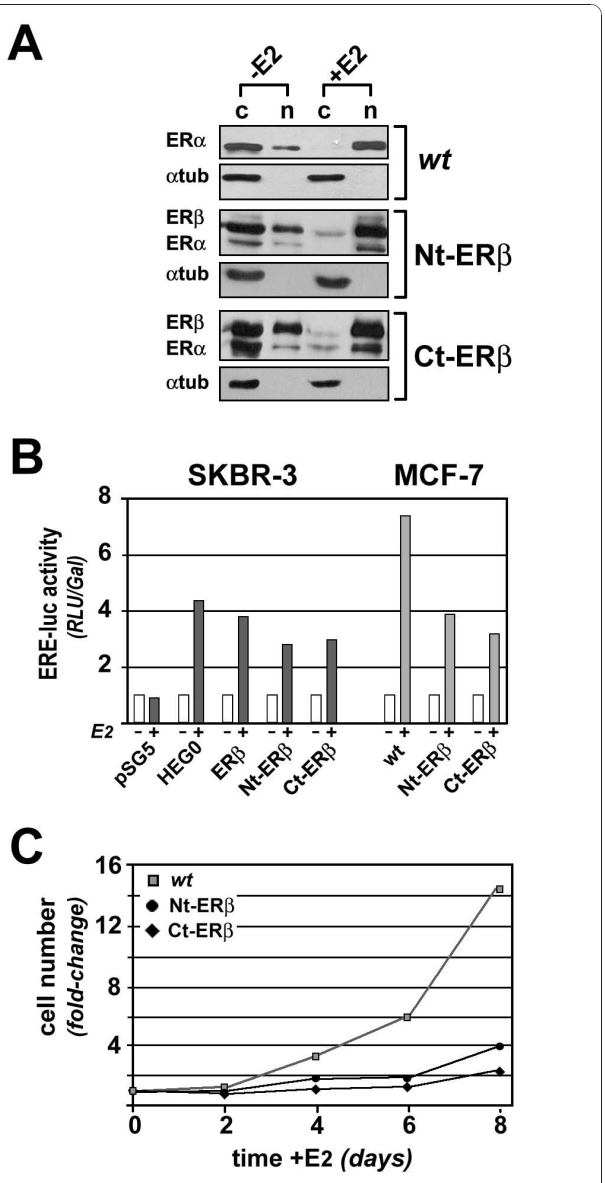



**Figure 1 Functional characterization of ERβ-expressing MCF-7 cells**. (A) Nuclear translocation of ERα and ERβ shown by western blot analysis on cytosolic (c) and nuclear (n) protein extracts, prepared from *wt* MCF-7, Nt-ERβ and Ct-ERβ cells after treatment with either 17β-estradiol ($10^{-8}$M, +E2) or vehicle alone (-E2) for 45 minutes. The amount of α-tubulin was also analyzed to verify the absence of cytosolic contaminants in the nuclear fractions. (B) The transcriptional activity of ERα, Nt-ERβ and Ct-ERβ was measured by transient transfection in SKBR3 cells (*left*) and the ability of tagged ERβ to interfere with ERα activity was assessed by comparing estrogen effects in wt, in Nt-ERβ and Ct-ERβ MCF-7 cells (*right*); in all cases transiently transfected ERE-*tk*-luc was used as reporter gene. (C) Proliferation of *wt* MCF-7, Nt-ERβ and Ct-ERβ cells was measured by stimulating hormone-starved cells with $10^{-8}$M E2, followed by cell counting with a colorimetric assay at the indicated times.

observations confirmed that both tag-ERβ expressing cell lines generated for this study show a well defined phenotype, with respect to the known activities of this ER subtype in BC cells, and are thus suitable to investigate the genomic bases of ERβ actions in this cell type. As we could not exclude that the presence of the TAP tag at either the N- or C- term of ERβ may specifically influence its activity on cellular targets or pathways different from those investigated above, all experiments reported in this study were performed in both Ct-ERβ and Nt-ERβ cells and the data were combined for analysis, with the aim to focus on the most significant and reproducible actions of ERβ independently from position of the tag in the receptor moiety.

### Effects of ERβ on the estrogen-responsive MCF-7 cell transcriptome

Expression of ERβ is known to cause significant changes in the genomic response to estrogen in target cells. To identify the genes whose estrogen regulation in hormone-responsive BC cells is perturbed by ERβ, we performed gene expression profiling with microarrays in estrogen-starved, quiescent *wt* and TAP-ERβ+ (Ct-ERβ and Nt-ERβ) MCF-7 cells following stimulation with $10^{-8}$M E2. Total RNA was extracted from the three cell lines either before or after 2, 4 or 8 hrs stimulation, fluorescently labelled and analyzed on whole-genome microarrays. Samples extracted from the two ERβ-expressing cell lines were pooled before analysis, to reduce the impact of clone-specific differences and to focus on the most significant effects of ERβ, independent from tag position in the protein. Results obtained in these samples were then compared with those obtained under the same conditions in *wt* MCF-7 cells. This analysis yielded 921 transcripts differentially regulated by the hormone in ERβ+ *vs* ERβ- cells (Figure 2), including 234 mRNA whose regulation was detected only in *wt* MCF-7 cells, 516 regulated only in ERβ+ cells (see Venn diagram in Figure 2), 154 showing a similar pattern of response in both cell types (up- or down-regulated in all cases, although at different levels) and 17 showing opposite responses to the hormone in ERβ+ *vs* ERβ- cells (14 transcripts repressed in *wt* cells but activated in ERβ+ cells and 3 showing an opposite behaviour). The full list of these differentially regulated transcripts is reported, with relevant information, in Additional Table S1. Taken together, these results indicate that the presence of ERβ greatly influences the response of the MCF-7 cell genome to estrogen, by interfering with ERα-mediated hormonal regulation of 405 genes (Figure 2, left and central panels) and promoting *de novo* regulation of 516 genes (Figure 2, right panel). It should be noted that these analyses were performed with data obtained after 8 hrs of hormonal stimulation, a timing that allowed us to focus on early

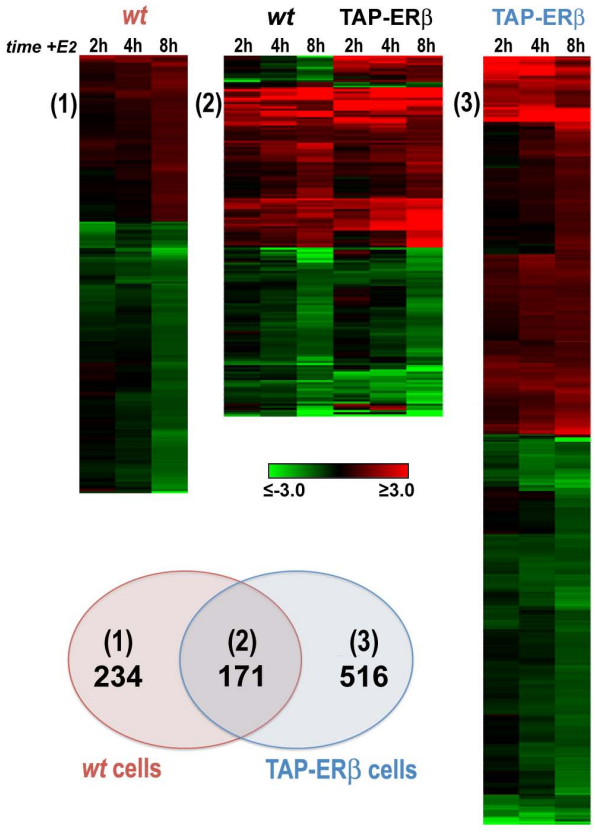



**Figure 2 Gene expression differences in absence or presence of ERβ**. *Top*: Heatmap summarizing the effects of ERβ expression of the estrogen responsive transcriptome of MCF-7 cells, showing changes in expression ($log_2$ of the fold-change) of 921 transcripts after cell exposure to $10^{-8}$M E2 for the indicated times. Transcripts are grouped as follows: regulated only in *wt* MCF-7 cells (1), in both cell lines (2) and only in TAP-ERβ cells (3). *Bottom*: Venn diagram showing the numbers of differentially regulated by E2 in *wt* MCF-7 only (1), both cell lines (2) or ERβ expressing cells only (3).

response genes, positioned upstream in the composite transcriptional cascade set in motion by the hormone in this cell type and more likely to include primary genomic targets of ERs [26]. It is thus possible that this analysis missed ERβ-responsive genes showing significant changes in expression only at later times after hormonal stimulation. However, analysis of the global effects of ERβ on gene expression in these same cells, performed as described above in cultures exponentially growing under continuous hormonal stimulus, suggests that the number of regulated transcripts identified here is rather close to the total number of genes targeted by this ER subtype in MCF-7 cells [O. Paris *et al.*, manuscript in preparation].

### Global mapping of ERα and ERβ binding to MCF-7 cell genome

The widespread effects of ERβ on MCF-7 cell transcriptome are likely to result from multiple effects of this

receptor in the cells, including direct regulation of primary response genes *via* genomic or non genomic mechanisms, and indirect gene regulation events mediated by the products of primary genes. The primary ERβ target genes are most likely to comprise also master regulators of complex cellular responses to the receptor, mediating its effects on the biological and clinical phenotype of BC cells. To identify such primary genomic targets and investigate the mechanisms that allow their regulation by ERβ, a global analysis of *in vivo* binding of this receptor to the genome was carried out in TAP-ERβ cells by chromatin immunoprecipitation coupled to massively parallel sequencing (ChIP-Seq) [38], that allows detailed mapping of *in vivo* TF binding to the genome. In parallel, we studied ERα binding to the genome under the same conditions, to allow comparative analyses between the two ER subtypes. Replicate chromatin samples were prepared from both Ct-ERβ and Nt-ERβ cells before and after stimulation for 45 minutes with $10^{-8}$M E2 and DNA-bound proteins were immunoprecipitated either with antibodies against the N- and C-terminus of ERα, or with IgGs binding with high affinity the TAP moiety of tagged ERβ (see Methods). Preliminary testing on several known ERβ binding sites, including the promoter-near region of pS2/TFF1 gene [26], confirmed that the method selected to immunoprecipitate chromatin-bound Ct-ERβ and Nt-ERβ was efficient and specific (data not shown). The resulting DNAs were used to generate ChIP-Seq libraries for ERα and ERβ, respectively, that were then sequenced with the Illumina Genome Analyzer. The sequence tags obtained were then aligned to the human genome sequence and peaks enriched in the libraries generated after E2-treatment were identified using MACS (Model-based Analysis of ChIP-Seq) [39]. This led to the identification of 9702 binding sites for ERβ and 6024 sites for ERα, of which 4506 were shared by both receptors (Figure 3A), with an average False Discovery Rate (FDR) of 3%. The full list of these binding sites is available, with relevant information, in Additional Files 1 and 2. Interestingly, about half (4862) ERβ binding sites identified map within transcription units, mainly (3942 sites) in intronic regions. This distribution is maintained also in 424 ERβ-regulated transcription units (see below), where 966 ERβ binding sites located in the gene or within 10 kbps from it are distributed as follows: 154 in promoter regions, 51 in exons, 471 within one or more introns and the remaining either upstream of promoters (156) or downstream of the gene (134). In both cases the ERβ binding sites within genes did not show any preference with respect to exon or intron position nor for know intragenic regulatory elements (splice sites, polyadenylation sites, etc). It should be mentioned that the number of ERβ binding sites identified is significantly higher that

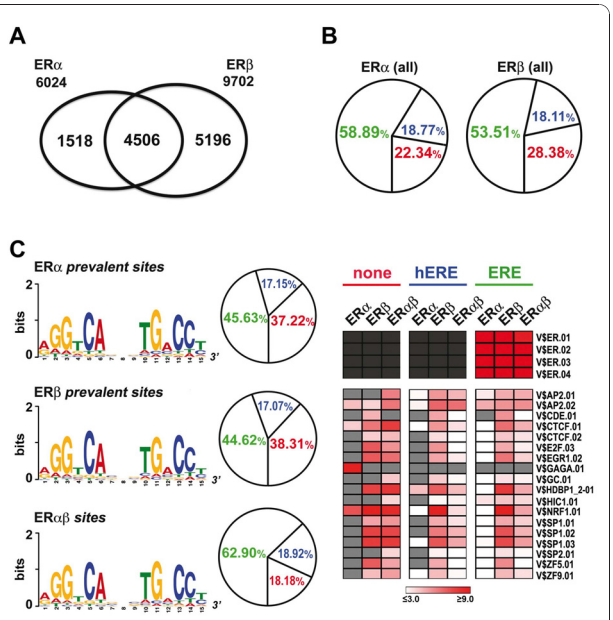



**Figure 3 Sequence analysis of ERα, ERβ and ERα+ERβ binding sites**. (A) Venn diagram showing a summary of ERα and ERβ binding sites identified in TAP-ERβ cells by ChIP-Seq. (B) Classification of ERα and β binding sites based on the presence of a perfectly or imperfectly palindromic Estrogen Response Element (ERE, green), an ERE hemipalindrome (hERE, blue) or no ERE (none, red). (C) ERE motif matrices identified in each of the three ER binding regions indicated (*left*), classification of the binding sites belonging to each region according to the presence of ERE (*center*) and grid summarizing the results of TFBS matrix enrichment (overrepresentation) analyses performed on the binding sites groups indicated (*right*). Z-Score cut-off was 3.0 and only TFBSs showing an over-representation score ≥4.0 in at least one of the ERE- (none) binding site groups. Light grey cells indicate a Z-Score <3.0 while dark grey cells indicate absence of the matrix.

those mapped in MCF-7 cells by ChIP-on-chips [30,31], possibly for technical differences due to ERβ expression levels in the different MCF-7 cell-derived clones used, in immunoprecipitation efficiency and/or in DNA analysis. Since only Zhao et al. [31] performed an unbiased search for ERβ binding with whole-genome chips, we could confront our results only with those reported in that study. This showed that 86% of high confidence ERβ sites described in that study appear also in our dataset. The binding sites identified here were then subjected to sequence analysis, searching first for the presence of EREs (Estrogen Receptor Elements), the characteristic ER binding signature (Figure 3B). This analysis revealed that in all three cases (i.e. ERβ, ERα and ERβ+ERα) a high percentage of sites displayed one or more imperfectly palindromic ERE (ERE+), with a slightly higher positivity in ERα sites (58.89 *vs* 53.51%). As ERs have been shown to bind both *in vitro* and *in vivo* to PuGGTCA hemi-palindromes (hEREs), we

searched the sequence of the remaining (ERE-) sites for perfect matches to this sequence. Results showed that almost half of them indeed contained one or more hEREs. The percentages of sites not carrying a known ER-binding element (ERE- and hERE-) were similar for both receptors (ERα: 22.34%, ERβ: 28.38%). We observed that ERα and ERβ binding sites were often found in close proximity to each other, a confounding factor when attempting to discern and analyze separately ER subtype-specific sites and target genes. This could be due to the limits of the ChIP-Seq technology or of the algorithm used for peak selection. To overcome this problem, and allow the identification of potential ER subtype-specific sites, we used a cartographic approach to group nearby binding sites that might be the result, at least in part, of shortfalls of the mapping methods applied. Each binding peak was thus elongated in both directions by 1000 bp and the overlapping ones obtained were merged into 8536 ERβ and 5371 ERα 'extended' binding regions. These regions were intersected to define ERα only, ERβ only or ERα+ERβ binding regions. In this way we could identify 1271 ERα-only and 4541 ERβ-only binding sites, comprised in these regions, none of which showed nearby binding of the other receptor. These were named: ER subtype 'prevalent' sites. The binding peak sequences included in each of the three regions obtained (ERα only, ERβ only or ERα+ERβ) were then re-analyzed for the presence of ERE or hERE elements. In this way we could observe that sites within the ERα+ERβ regions showed now a much higher percentage of ERE+ sequences (62.90%), respect to those present in the ERα-only or ERβ-only regions (45.63% and 44.62%, respectively, Figure 3C). Since all three types of sites showed almost identical proportions of hERE+, this result suggests that perfectly or imperfectly palindromic EREs are preferential binding sequences for ERα-ERβ heterodimers, while ERα and ERβ homodimers appear to be more flexible in DNA recognition. ERE+ sequences were then analyzed in more detail with MEME [40], to investigate if the three classes of sites identified showed any difference in the relative base composition of their respective ERE signatures. For each list of sequences, the most significant position-specific probability matrix generated by MEME was compared to the matrices present in the JASPAR transcription factor binding profile database [41], using the STAMP tool-kit for DNA motif comparison [42]. As shown in Figure 3C (left panels), this analysis revealed that the ERE matrices derived from the three types of binding regions identified (ERα selective, ERβ selective and ERα+ERβ) are identical and, as a consequence, that ERβ does not appear to display ERE variant selectivity.

We then examined the ERE- sequences to search for enriched binding motifs for other transcription factors that might play a role in ER binding to chromatin in the absence of canonical EREs. ERs are known to be able to bind chromatin indirectly, by physically interacting with DNA-bound TFs (tethering), including SP1 [43,44] or the AP1complex [31,45,46], for gene *trans*-regulation. TFBS enrichment respect to the whole genome was calculated thus in ERE- sites with using the RegionMiner tool [47] and only statistically significant (Z-score ≥3.0) and highly enriched (over-representation ≥4-fold) matrices were further considered. The results are summarized in the right panel of Figure 3C, showing for comparison the over-representation values scored in hERE+ and ERE+ sites by the same matrices selected in the ERE- sites. These numerical values, together with relevant information, are reported in Additional Table S2 [Additional file 3]. The enriched matrices found in the ERE- set of ERβ prevalent sites include V$SP1.01, V $SP1.02, V$SP1.03, V$SP2.01 and V$GC.01, all belonging to the GC-Box family targeted by SP1 and GCFC1 (GC-rich sequence DNA-binding) factors, V$CTCF.01 and V$CTCF.02, binding site matrices for the CCCTC-binding factor CTCF, that is a known transcriptional repressor of c-myc [48], V$NRF1.01, binding NRF1 (nuclear respiratory factor 1), a transcription factor that regulates the expression of nuclear-encoded mitochondrial genes [49], V$ZF5.01, for the POZ domain zinc finger protein ZF5, and V$ZNF9.01, recognized by the zinc finger proteins ZNF148, 202, 219 and 281. The majority of these TFs bind GC- and C-rich sequences that are structurally related to each other, suggesting the possibility that a significant portion of the sequence elements listed above might indeed be recognized by one or a limited number of TFs. On the other hand, the V$GAGA.01 matrix was specifically enriched only in the ERα prevalent ERE- binding sites. This sequence binds a transcription factor known to influence chromatin structure in *Drosophila* [50] and to bind throughout the genome [51], but nothing is known about physical or functional interactions of this factor with ERs or other nuclear receptors. The results of this analysis point to TFs that could act as partners of ERβ for binding to chromatin in the absence of canonical EREs. Interestingly, the majority of these same matrices were found enriched also in the ERβ binding sites comprising hEREs or EREs (Figure 3C), suggesting that one or more such DNA matrices might cooperate with ERβ for either DNA binding or gene *trans*-regulation. We performed a direct search for conserved motifs in the ERE- binding sites of ERβ with MEME [40], and the most significant position-specific probability matrices were compared to those present in the JASPAR TF binding profile database [41]. The results failed to provide any conclusive information, as each of several sequence motifs obtained with this analysis was found only in a small fraction of the binding sites analyzed.

### Identification of primary ERβ target genes

To identify the genes directly controlled by ERβ binding to the genome, and analyze the interplay between the two ERs in gene regulation, we combined the ChIP-Seq data with those relative to estrogen responsive genes differentially regulated by E2 in ERβ+ *vs* ERβ- MCF-7 cells under the same experimental conditions (Figure 2). Three sets of ER binding regions (defined as described in Methods) were used for this analysis: ERβ (8872, Set 1) and ERα (5558, Set 2) binding regions identified in TAP-ERβ cells, and all ERα binding regions identified so far in *wt* MCF-7 (17888, Set 3). The regions from Set 1 (ERβ in ERβ+ cells) were intersected with those from Set 2 (ERα in ERβ+ cells), to define which of them was binding both receptor subtypes ('heterodimer ERα+ERβ': 4186), only ERβ ('homodimer ERβ': 4686) or only ERα ('homodimer ERα': 1372) in TAP-ERβ cells. The 'homodimer ERβ' and the 'heterodimer ERα+ERβ' groups were further filtered against Set 3 binding regions (ERα sites detected in ERβ- cells), to identify the genomic sites recognized by ERβ, with or without ERα, but never by ERα alone. This allowed us to classify the sites comprised in these regions as follows: Class 1, including 2126 sites occupied by ERα in *wt* MCF-7 cells and by ERβ in TAP-ERβ cells (ERβ *vs* ERα competition); Class 2, showing 4340 sites where ERα can bind in *wt* MCF-7 cells and both receptors are detected in TAP-ERβ cells (ERβ+ERα); Class 3, with 2707 sites binding only ERβ and never, under any condition, ERα (ERβ specific); Class 4, comprising 529 sites where both receptors are found in TAP-ERβ cells but none in *wt* MCF-7 cells (ERβ+ERα specific); Class 5, including 617 sites where ERα binds only in TAP-ERβ but never in *wt* cells (ERα displacement); Class 6, composed of 773 sites that bind only ERα both in *wt* and TAP-ERβ cells (ERα specific). When combined with the results of the sequence analyses described above, this classification reveals that ERβ-specific *cis*-acting regulatory elements are unlikely to exist in the genome, as all evidence point to the fact that the two ER subtypes can recognize identical sequence features.

To identify among the genes differentially regulated by estrogen in ERβ+ *vs* ERβ- cells those representing direct targets for transcriptional regulation by DNA-bound ERβ in our cell model, we extracted from the list in Additional Table S1 [Additional file 3] the genes bearing one or more ERβ binding sites inside or within 10 kb of the TU, and termed them 'primary', to indicate that they are most likely to respond directly to the signal conveyed by the E2-activated receptor [26]. Of these 424 genes - listed in Additional Table S3 [Additional file 3], whose kinetics of response to E2 in *wt* and TAP-ERβ cells is shown in Figure 4, 52 show one ERβ site of Class 1 (ERβ *vs* ERα competition), 90 a site of Class 2 (ERβ+ERα), 71 a site of Class 3 (ERβ specific) and only 9 a Class 4 site (ERβ+ERα specific), while 202 showed multiple ERβ sites belonging to different classes and were thus classified accordingly (grey in Figure 4). In the right panels of Figure 4 are reported examples for each of the primary gene classes described above, showing the location of the receptor binding sites respect to the promoter and structural gene. It is worth mentioning that when the gene expression data from *wt* MCF-7 cells stimulated with E2 for 8hrs (Figure 2) were combined with information concerning ERα binding regions identified in *wt* MCF-7 cells under comparable conditions (Set 3 described above), 228 putative primary ERα target genes were identified -listed in Additional Table S4 [Additional file 3], 71% of which (163 genes) showed ERβ binding in hormone-stimulated ERβ+ cells. This result supports the view that the two ER subtypes tend to interact with the same targets in BC cells genome.

A functional analysis of the primary ERβ target genes identified here with Ontologizer [52] showed that most primary ERβ responsive genes are involved in key cellular processes, including control of cell proliferation, survival and differentiation status as well as cell motility, migration and adhesion, and can all greatly influence BC cell phenotype and response to estrogen - listed in Additional Table S5 [Additional file 4]. When GO analysis was performed on the ERα target gene set from *wt* MCF-7 cells (228 genes), results showed that the genes controlled directly by this ER subtype appear to be involved in the same cellular processes described above for ERβ -compare results reported in Additional Tables S5 and S6 [Additional files 4 and 5], providing a further indication of the significant overlapping between gene pathways targeted by ERβ and ERα in BC cells. Focusing on the genes differentially regulated by E2 in ERβ- *vs* ERβ+ cells known for their involvement in cell proliferation, we observed that many of them encode for transcription factors and other key proteins controlling large gene networks of cell division cycle and cell survival and, in general, cell proliferation and differentiation pathways. These include, in particular, CDK-6 (cyclin-dependent kinase 6), CEBPA (CCAAT/enhancer binding protein, alpha), DAB2 (Disabled homolog 2, mitogen-responsive phosphoprotein), HES-1 (Hairy and enhancer of split homologue 1), IGFBP-4 (Insulin-like growth factor binding protein 4), IRS-1 and -2 (Insulin receptor substrates 1 and 2), JAK-2 (Janus kinase 2), JunB, MITF (Microphthalmia-associated transcription factor), MYC, SLIT-2 (Slit homolog 2, *Drosophila*), SMARCA-2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2), SOX-9 (Sex determining region Y-box 9), TGFB-2 (Transforming growth factor beta 2), TGFBI/LCD-1 (transforming growth factor, beta-induced, 68 kDa), TGM-2 (Transglutaminase 2),

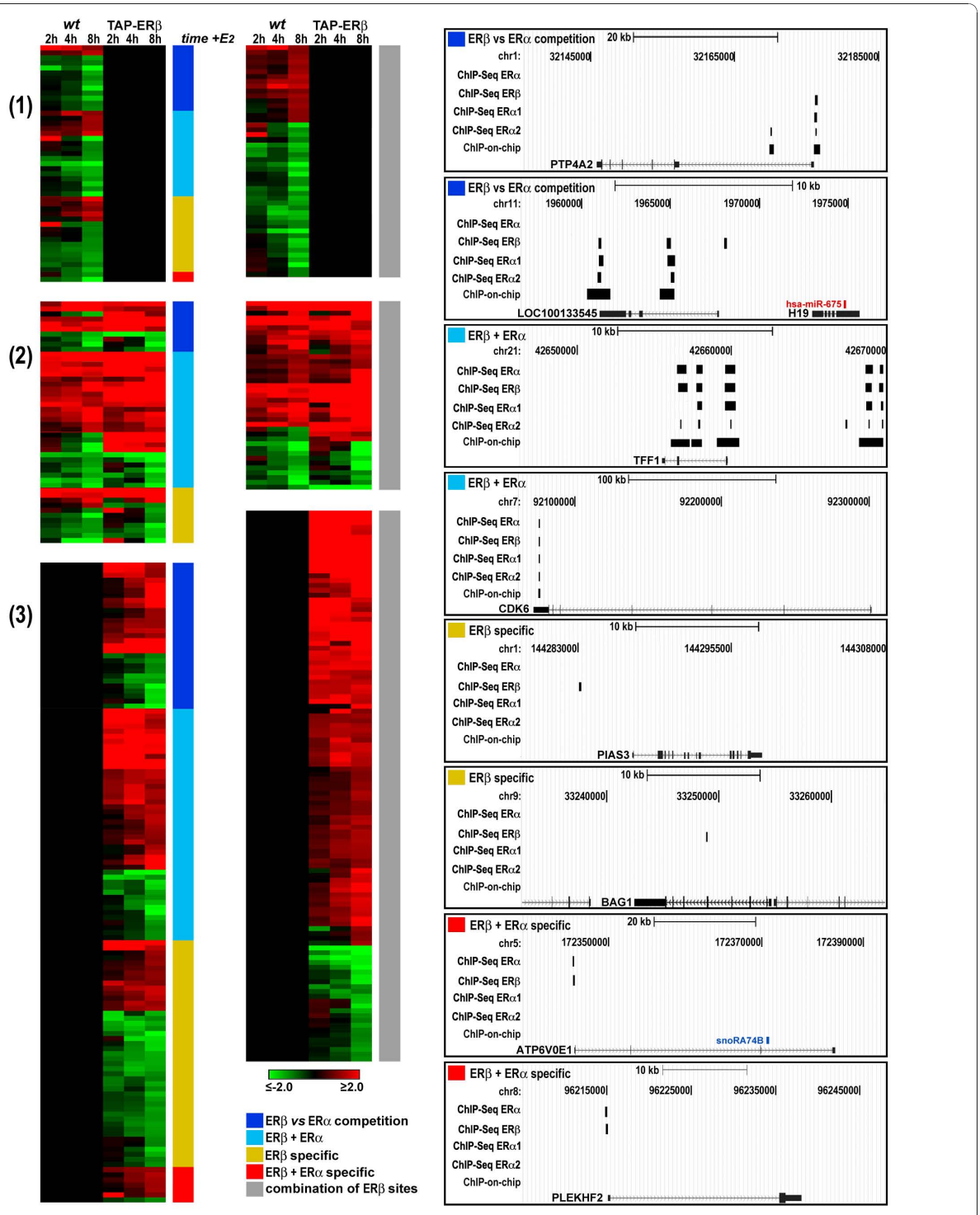



**Figure 4 Putative ERβ primary targets**. *Left*: Heatmap summarizing the effects of estrogen stimulation on 424 mRNAs encoded by genes showing one (*left*) or more (*right*) ERβ binding sites within 10 kb of the TU (primary response genes) transcriptome of MCF-7 cells, as changes in expression ($log_2$ of the fold-change) after cell exposure to $10^{-8}$M E2 for the indicated times. Transcripts are grouped as follows: regulated only in *wt* MCF-7 cells (1), in both cell lines (2) and only in TAP-ERβ cells (3). Vertical bars to the right of each heatmap indicate the class of ERβ binding site present, as indicated in the legend. When a regulated gene showed multiple ERβ binding sites belonging to different classes it was included in a separate group, classified as 'combination of ERβ sites' (grey bar).*Right*: Genome Browser view of genomic loci representative of the different ERα and ERβ binding site categories identified. ChIP-Seq ERα and ChIP-Seq ERβ indicate sites identified in this study, ChIP-Seq ERα1, ChIP-Seq ERα2 and ChIP-on-chip indicates sites identified in MCF-7 cells by Cicatiello et al. [26], Fullwood et al. [27], and Hurtado et al. [28], respectively.

TNS-3 (Tensin 3) and WISP-2 (WNT1 inducible signaling pathway protein 2). Interestingly, the role of all these genes in tumor cell proliferation and differentiation is known and an involvement in hormone-mediated BC cell responses to ERα has been reported for most of them, suggesting that discovery of an ERβ direct effect on transcription of these genes provides a new molecular framework to elucidate the anti-proliferative and differentiative effects of this receptor subtype in hormone-responsive cells.

Among the RNAs encoded in the genome, microRNAs (miRNAs) have emerged as master regulators of gene expression for their ability to influence mRNA concentration and activity by post-transcriptional mechanisms. Recent results highlighted the role of miRNA in BC cells response to estrogen [26,53-58] and, in addition, several lines of evidence indicate extensive miRNA deregulation in BC, including differential expression of miRNAs in normal *vs* transformed mammary epithelial cells [59-61]. For these reasons, we focused our attention on the TUs encoding pre-miRNAs, to test the possibility that ERβ binding sites might be located in close proximity of these genes. Results show that 52 miRNA-encoding loci (isolated or in genomic clusters) show one or more ERβ site within 10 kb from the pre-miR sequence of the host gene - listed in Additional Table S7 [Additional file 6]. Distribution of these sites among the ER binding Classes described above was comparable to what observed for primary genes. Interestingly, in several cases ERβ binds both up- and down-stream of the pre-miR locus, further suggesting that receptor docking might exert multiple regulatory actions on miRNA biogenesis.

We tested the hypothesis that the observed distribution of sites near the pre-miRNA loci occurred at random by applying a bootstrap approach. To this end, we repeatedly sampled 1000 times the same number of sites of the real ERβ binding set, with the same length distribution, a similar distribution among chromosomes but randomly selected coordinates. We then counted the number of pre-miRNA loci and the number of randomly generated sites found within 10 kb of each other and compared their distributions with that of the experimentally detected ones. The number of randomly generated sites within 10 kb of a pre-miRNA never reached the value detected experimentally, while the number of miRNA loci scoring an artificial site in close proximity was equal or above what measured only in 7.6% of the cases. These results can be explained also by the observation that in several cases ERβ binds both up- and down-stream of the pre-miR locus, a result that support a functional significance of this observation. In fact, the ratio between the number of ERβ binding sites and the number of pre-miRNA loci within 10 kb of each other is 1.5, while this varied between 0.5 and 1 for the randomly generated sets (data not shown). We thus conclude that although some of the ERβ sites detected in or near pre-miRNA loci may be the result of a random, non functional event, these is likely to represent rare events, as random distribution never reaches the enrichment level observed experimentally. Indeed, preliminary miRNA profiling analyses carried out in *wt* MCF-7 and TAP-ERβ cells indicate that mature miRNA expression undergo extensive deregulation in the presence of ERβ [O. Paris *et al.*, manuscript in preparation and data not shown], suggesting that at least some of the sites identified here might indeed be involved in ERβ-mediated regulation of miRNA biosynthesis in BC cells.

### ERβ binding to the mitochondrial genome

Mitochondrial DNA (mtDNA) is usually overlooked in whole-genome ChIP-seq analyses, since identification of enriched peaks is more difficult here due to a much higher noise, consequence of the high and variable number of mtDNA copies in the cell. ERβ has been shown to localize in the mitochondria in different cell types [62] including human BC cells [63,64], and a role for estrogen in mitochondrial function, with implications on cell growth, has been established [65,66]. We thus analyzed the sequence reads that aligned on the mtDNA sequence with the same method used for whole genome data analysis, but applying a supplementary fold-intensity filter (see Methods) to deal with the higher background noise. This analysis revealed one ERβ binding site in proximity of the mtDNA D-loop, but no ERα binding sites either in *wt* MCF-7 or in TAP-ERβ cells (Figure 5A). Independent ChIP analysis confirmed this results, showing ERβ binding to this same mtDNA region upon activation by E2 or its selective agonist 2,3-bis(4-hydroxyphenyl) propionitrile (DPN) and lack of ERα binding (Figure 5B). Furthermore, the presence of ERβ in mitochondria of TAP-ERβ cells was confirmed biochemically, by western blotting (Figure 5C), an analysis that revealed also the presence of ERα in the organelle. When we analyzed the sequence of the ERβ mitochondrial site with MatInspector, we observed the presence of the matrix V$GATA1.06, bound by GATA1, a factor whose activity is strongly repressed by ERα [67], and V$HMGA.01, bound by HMGA1, a non-histone chromosomal protein that is highly overexpressed in cancer cells [68] and has been shown to interact with ERα and to enhance its binding to DNA [69].

## Conclusions

The results of this study indicate that *in vivo* ERβ can interact with hormone-responsive BC cell chromatin either alone or complexed with ERα, but in all cases the two receptors share the same genomic targets. An

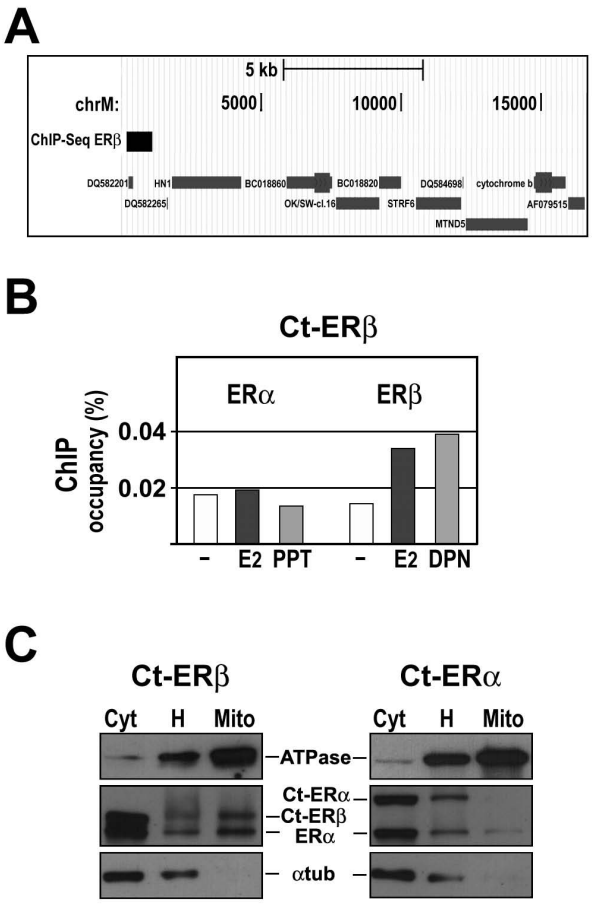



**Figure 5 Mitochondrial ER-beta binding sites**. (A) Genome Browser view of the ERβ binding site identified in mitochondrial genome. (B) Validation of ERβ binding to mitochondrial DNA by ChIP. Results shown are representative of duplicate analyses. E2: $10^{-8}$M 17β-estradiol; PPT: $10^{-8}$M 1,3,5-tris(4-hydroxyphenyl)-4-propyl-1H-pyrazole (selective ERα agonist); DPN: $10^{-8}$M 2,3-bis(4-hydroxyphenyl) propionitrile (ERβ agonist). (C) Western blot analysis of ERβ and/or ERα in purified mitochondria from Ct-ERβ and Ct-ERα [33] cells. Cyt: cytosol depleted of mitochondria, H: whole cell homogenate; Mito: purified mitochondrial fraction. ATPase is a mitochondrial marker and α-tubulin was included to determine the level cytosolic contaminants in 'Mito' samples.

observation that is in agreement with the conclusions of previous studies based on analysis of ERα and ERβ heterodimerization and binding to the ERE [9,70-72]. When both ER subtypes are expressed in the same cell, the main action of ERβ in the genome is thus achieved in combination with ERα, by either heterodimerization or competition for binding to the same target sites in chromatin. Based on these observations, we propose that in hormone-responsive BC the final cellular response to estrogen is likely to depend upon the relative concentration of the two ERs in the cell, their activation status and DNA binding kinetics and the presence of other factors influencing their respective functions.

## Methods

### Plasmid preparation, cell lines, cell culture and stable transfections

Different BC cell lines were used: MCF-7 TET Off (ER-alpha positive; ATCC, Cat No. HTB-22) and SKBR3 (ER-alpha negative; ATCC, Cat No. HTB-30). MCF-7 TET Off cells (described here as *wt* or ERβ-) were used to produce stable clones expressing ERβ tagged with TAP-tag respectively at the C-term and at the N-term (Ct-ERβ and Nt-ERβ) or C-tagged ERα (Ct-ERα), as previously described [5,33]. All were grown in Dulbecco's modified Eagle's medium (DMEM), supplemented with 10% fetal bovine serum (FBS) (both from Invitrogen), 100 U/ml penicillin, 100 mg/ml streptomycin, 250 ng/ml Amfotericin-B, 50 μg/ml G418 (normal growing condition). For estrogen starvation, cells were plated at 40% confluence in steroid-free medium (phenol red-free DMEM medium, with 5% fetal bovine serum pre-treated with dextran-coated charcoal and antibiotics) and maintained for 5 days with replacement of fresh medium before stimulation with $10^{-8}$M 17β-estradiol (E2).

### Preparation of cell extracts, mitochondria isolation and immunoblotting analyses

Cells starved in 100-mm dishes were stimulated for 45 minutes, harvested in cold PBS and collected by centrifugation at 1000 × g. The cell pellets were then resuspended in three volumes of Hypotonic Buffer (HB) (20 mM HEPES pH 7.4, 5 mM Sodium Fluoride, 10 μM Sodium Molybdate, 0.1 mM EDTA, 1 mM dithiothreitol, 1 mM protease inhibitors, 1 mM Phenylmethyl-Sulfonyl Fluoride). Cells were then incubated on ice for 15 min and 0.5% Nonidet P-40 followed by spinning 30 sec at 4°C at 13000 × g. Supernatants were recovered and clarified at 13,000 × g for 15 min at 4°C while pellets were resuspended in hypotonic buffer, stratified on 25% sucrose-HB solution and centrifuged at 6000 × g for 15 min at 4°C. The resulting pellets were then resuspended in one volume of Nuclear Lysis Buffer [73] containing 800 mM NaCl, incubated for 30 min at 4°C with gentle shaking and centrifuged for 15 min at 4°C at 13000 × g. The supernatant fraction was recovered.

Mitochondria were isolated from 20 × $10^6$ Ct-ERβ or Ct-ERα cells (in 150 mm culture dishes) as described [74], with minor modifications. All steps during mitochondria isolation were performed at 4°C, cells were washed twice in PBS, scraped, and centrifuged at 290 × g, 5 minutes. The samples were resuspended in buffer A (250 mM Sucrose, 50 mM Tris-HCl, 2 mM EGTA), homogenized in Glass-Teflon Potter homogenizer and centrifuged at 600 × g for 3 min, then the supernatants were re-centrifuged at the same speed. Mitochondria were pelleted by centrifugation at 10400 × g for 10 min, resuspended in buffer A and centrifuged again at 5300 × g for 10 min, in order to

eliminate microsomal and cytosolic contamination. The samples were dissolved in buffer A and centrifuged at 1500 × g for 4 min and then pelleted again at 9600 × g for 10 minutes. The final pellet was resuspended in Buffer B (50 mM Tris pH 7.5, 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 1% Na Deoxycholate, 1 mM PMSF, 1X Protease Inhibitor) and incubated on ice for 20 min to extract mitochondrial proteins. A small portion of sample was collected after homogenization and processed to obtain the samples corresponding to homogenate and cytosol. An equivalent protein amount was fractionated by SDS-PAGE on Mini Protean Precast polyacrylamide gels 4-20% from Biorad.

Homogenate, cytosol, nuclear or mitochondrial extracts from equivalent cell number were fractionated by SDS-PAGE. Immunoblot analysis was performed using the following primary antibodies: Ct-ERα (HC-20; sc-543) from Santa Cruz Biotechnology, Inc., TAP tag (CAB1001) from Open Biosystems, α-Tubulin (T 6199) from Sigma-Aldrich, ATPase B (ab14730) from Abcam. Peroxidase-labelled anti-rabbit or -mouse immunoglobulin antisera were used according to the manufacturer's instructions (Amersham Italia).

#### Transient Transfections and Luciferase Assay

Wild type MCF-7, Ct-ERβ and Nt-ERβ clones were starved for 5 days in estrogen-free medium. Then 5 × $10^5$ cells/dish were seeded in 60 mm culture dishes and transfected by using 25 μg/dish polyethylenimine (Polysciences, Inc.) with 2.5 μg/dish DNA, including 300 ng ERE-tk-Luc [75], 500 ng pSG-Δ2-NLS-LacZ vector [76], co-transfected as an internal control for transfer efficiency, and carrier DNA (Bluescribe M13+). At 48 hrs after transfection, cells were treated for 24 hrs with either vehicle (EtOH) or E2 ($10^{-8}$M).

SKBR3 cells were grown to 60-70% confluence and transfected with Lipofectamine 2000 reagent (Invitrogen) and OPTI-MEM (Invitrogen) according to the manufacturer's instructions. The plasmids used were pSG5-ERβ, encoding full-length ERβ (ERβ1), pSG-HEGO, encoding the full-length ERα and the corresponding pSG5 empty vector (Stratagene), pUSE-C-TAP-ERβ and pUSE-N-TAP-ERβ, encoding full-length ERβ tagged, respectively, at the C-term and at the N-term [5], ERE-*tk*-Luc and pSG-Δ2-NLS-LacZ. Six hours after transfection, the medium was changed and 24 hrs later cells were stimulated as described above for 24 hrs. At the end of treatment, cells were washed with cold PBS and lysed in 100 μl lysis buffer (Promega). Luciferase activity was measured in extracts using the Luciferase Assay Reagent (Promega), according to the manufacturer's instructions, and values were expressed as relative light units normalized to the β-galactosidase activity or to the protein concentrations measured using the Bradford technique. For each condition, average luciferase activity was calculated from the data obtained from three independent dishes.

#### Cell Cycle Analysis

Estrogen-deprivation was always controlled by cell cycle analysis as follows. Cells (1.5 × $10^5$ cells/dish) were starved in 60 mm culture dishes, stimulated for 27 hrs and collected in PBS containing 50 μg/ml propidium iodide, 0.1% (v/v) sodium citrate, 0.1% (v/v) Nonidet P-40. Cell samples were incubated in the dark for at least 15 min at room temperature, or overnight at 4°C, and analyzed by a FACScalibur flow cytometer using the CellQuest software package (BD Biosciences), according to standard protocols suggested by the manufacturer [77,78]. Data analysis was performed with Modfit software (Verity Software, Topsham). Values were plotted as increasing of S-phase respect to unstimulated controls. Results showed were obtained from two independent experiments.

#### Cell Proliferation Assay

Hormone-starved cells (3000/well) were seeded in 96-well dishes. After 12 hrs, medium was changed to include the indicated compounds. After appropriate stimulation, cells were washed in phosphate-buffered saline (PBS) and fixed with 12.5% glutaraldehyde for 20 min at room temperature, followed by washing with distilled water, incubation with 0.05% methylene blue for 30 min, rinsing and incubation with 0.33 M HCl for 18 hrs. Absorption was measured at 620 nm.

#### RNA purification

Total RNA was extracted from *wt* MCF-7, Ct-ERβ and Nt-ERβ clones, generated as described above, using the standard RNA Extraction method with TRIzol (Invitrogen) method, as described previously [79,80]. Cells were estrogen-deprived and total RNA was extracted before or at the indicated times after stimulation with $10^{-8}$M 17β-estradiol (+E2) or ethanol vehicle.

In each case RNA derived from two independent experiments performed in duplicate was used. Before use, RNA concentration in each sample was assayed with a ND-1000 spectrophotometer (NanoDrop) and its quality assessed with the Agilent 2100 Bioanalyzer with Agilent RNA 6000 nanokit (Agilent Technologies).

#### Microarray analyses

Total RNA extracted from Ct-ERβ and Nt-ERβ cells were pooled. For mRNA expression profiling, 500 ng total RNA were reverse transcribed, as described previously [26,81] and used for synthesis of cDNA and biotinylated cRNA according to the Illumina TotalPrep RNA Amplification Kit (Ambion, Cat. n. IL1791)

protocol. For each sample, 750 ng of cRNA were hybridized for 18 hrs at 55°C on Illumina HumanHT-12 v3.0 BeadChips, containing 48,804 probes (Illumina Inc.), according to the manufacturer's protocol and subsequently scanned with the Illumina BeadArray Reader 500. Data analyses were performed with GenomeStudio software version 2009 (Illumina Inc.), by comparing all values obtained at each time point against the 0 hrs values. Data was normalized with the quantile normalization algorithm, and genes were considered as detected if the detection p-value was lower than 0.01. Statistical significance was calculated with the Illumina DiffScore, a proprietary algorithm that uses the bead standard deviation to build an error model. Only genes with a DiffScore ≤-40 and ≥40, corresponding to a p-value of 0.0001, were considered as statistical significant.

#### Chromatin Immunoprecipitation

Ct-ERβ and Nt-ERβ cells were hormone-deprived for 4 days and chromatin was extracted in several replicates before (-E2) and after stimulation for 45 minutes with $10^{-8}$M 17β-estradiol (+E2) or, where indicated, with the $10^{-8}$M selective ERα agonist 1,3,5-tris(4-hydroxyphenyl)-4-propyl-1H-pyrazole (PPT) or $10^{-8}$M ERβ agonist 2,3-bis (4-hydroxyphenyl) propionitrile (DPN), from Tocris Cookson.

Chromatin was prepared with the Millipore/Upstate Chromatin Immunoprecipitation (ChIP) Assay Kit (Millipore) according to the instruction provided by the producer, using a variation of the protocol described at the Upstate website. For each assay, a total of $5 \times 10^6$ cells were fixed with 1% formaldeyde for 10 min at room temperature, the reaction was then stopped by adding glycine at final concentration of 0.125 M. Fixed cells were washed twice with ice-cold PBS, harvested by scraping, centrifuged and the cell pellets were re-suspended in SDS lysis buffer. Samples were sonicated with a Diagenode Bioruptor (Diagenode) for 12 cycles of 30 sec at high power, centrifuged at 12500 xg for 15 minutes and diluted 8-fold in ChIP dilution buffer. After removing an aliquot (whole-cell extract input), the chromatin sample was divided in three aliquots, that were incubated at 4°C overnight with antibodies against either the C-term (HC-20, from Santa Cruz Biotechnology) or N-term (anti-Estrogen Receptor 18-32, from SigmaAldrich) of human ERα or with IgG Sepharose 6 fast Flow (GE Healthcare Bio-Science AB) for TAP-ERβ [5]. As control, an aliquot of the same chromatins were processed in the same way but Abs or IgGs were omitted form the reaction. The samples were then precipitated by binding to protein-A Agarose/Salmon Sperm DNA beads (Millipore), for ERα, or as such for to ERβ. The beads were washed sequentially with 'low-salt immune complex wash buffer, 'high salt immune complex wash buffer, 'LiCl immune complex wash buffer' and TE buffer, before elution in Elution buffer by ON incubation at 65°C and treatment with Proteinase K. DNA was purified from immunoprecipitated (IPP) chromatin by extraction with phenol:chloroform:isoamyl alcohol (25:24:1) and ethanol precipitation according to standard procedures. DNA pellets were dissolved in nuclease-free water and kept frozen before further use.

Primers for ChIP-QPCR validation of the mitochondrial genome ERβ binding site were the following: 5'-GATCACAGGTCTATCACCCTATTAACC (forward) and 5'-CAGCGTCTCGCAATGCTATC (reverse).

#### Samples preparation for ChIP-Seq

DNAs from Ct-ERβ and Nt-ERβ cells treated with E2 were pooled together to generate an ERβ +E2 sample and the same was done for DNAs from hormone-starved cells (ERβ -E2 sample). Similarly, IPP DNAs obtained with anti-C-term and anti-N-term ERα Abs were pooled together to generate ERα +E2 and ERα -E2 samples. About 20 ng of ChIP DNA was purified using the MinElute PCR Purification Kit (QIAGEN, Italy), with a recovery of 55-70%, as assessed with the Quant-IT DNA Assay Kit-High Sensitivity and a Qubit Fluorometer (Invitrogen). Preparation of IPP DNA libraries for massively parallel sequencing was performed from 10 ng purified DNA according to the Illumina ChIP-Seq sample preparation kit protocol (Illumina Inc.). Libraries were sequenced with the Illumina Cluster Station and Genome Analyzer II according to manufacturer's instructions.

#### ChIP-Seq data analysis

The sequence tags generated by massively parallel sequencing were aligned on the human genome (hg18) with the software ELAND, allowing up to 2 mismatches. Enriched regions from ERα +E2 and ERβ +E2 samples were compared with the same from ERα -E2 and ERβ -E2 samples, respectively. The enriched ChIP-Seq peaks were identified using MACS (Model-based Analysis of ChIP-Seq) version 1.3.7.1 [39], with standard parameters (p-value cut-off of 1e-5, mfold of 32). For mtDNA we further filtered out sites with tag density below 0.5 (N/l; N = number of tags, l = length of site)

#### Computational searches for ERE sequences

The ERE binding motif was searched in binding sites with the MatInspector application [82], a part of GenomatixSuite software (Genomatix Software GmbH, Germany). The matrices ER.01, ER.02, ER.03 and ER.04 (Genomatix Matrix Library 8.2), were used with a core similarity threshold of 0.75 and a matrix similarity threshold of Optimal -0.02. The sequences bearing a match of any of the four matrices were termed ERE+ sequences. On the remaining sequences, the hemi-palindromic ERE

motif was searched with the same application, by defining a custom half ERE matrix PuGGTCA (hERE). The remaining sequences were classified as ERE- and were scanned for other TF binding sites motifs contained in the Genomatix Matrix Library using the standard parameters, as described previously [26].

### TFBS over-representation analysis

ERE- sequences were analyzed to search for known TF binding sites (TFBSs) that were enriched (over-represented) with respect to their relative frequency in the whole human genome. This analysis was performed with the RegionMiner application of the Genomatix software suite [47]. The software automatically searches for all TFBS matches present in the submitted sequence list and calculates the over-representation value of the actual number of matches over the expected value based on its frequency in the reference set (genome or promoters) for each matrix. It reports also the significance of the over-representation ratio, expressed as Z-scores [83]. Enrichment values with a Z-score <3.0 were not considered further. A filter was applied also on the over-representation values, depending upon the range of values set for each analysis, to highlight only the stronger associations. Results are shown as heatmaps representing over-representation values, generated with MeV software [84].

### Classification of ERβ binding sites and identification of primary ERβ responsive genes by combining ChIP-Seq and expression profiling data

Three sets of ER binding sites were taken in consideration. The first and second comprised, respectively, the ERβ and ERα sites mapped in TAP-ERβ cells and the third (named 'MCF-7 ERα') included all ERα binding sites identified in *wt* MCF-7 cells by ChIP-Seq [26,27] and/or by ChIP-on-chip [28]. First of all, the ER binding sites from ChIP-Seq analyses were elongated in both directions by 1000 bp. Subsequently, using UCSC Table Browser [85], for each of the first two sets the extended sites overlapping with each other were merged in ERβ or ERα binding 'regions', respectively. For the third set (*wt* MCF-7 ERα), the extended ChIP-Seq sites and the ChIP-on-chip sites overlapping with each other in the genome were all merged to generate unique MCF-7 ERα binding 'regions'. To identify putative primary ERβ responsive genes, the TUs corresponding to RNAs differentially regulated by E2 in ERβ+ *vs* ERβ- cells that showed one or more ERβ binding region inside or within 10 kb were extracted using UCSC Table Browser, as described previously [26].

### Gene Ontology analysis

To identify Biological Process GO terms statistically overrepresented in our regulated gene lists, we used Ontologizer 2.0, a tool for GO term enrichment analysis of genes derived from an experiment [52]. We identified enriched terms in primary ERβ or ERα target genes against all genes expressed (detected) in the cell lines under study, set as background of the analysis, with a $p$ value threshold of 0.01.

### Microarray and ChIP-Seq data accession numbers

The microarray and ChIP-Seq data have been deposited in the Array Express database ( HYPERLINK "http://www.ebi.ac.uk/arrayexpress" http://www.ebi.ac.uk/arrayexpress) with Accession Numbers E-TABM-1051 and E-MTAB-345, respectively.

## Additional material

**Additional File 1: Grober_ER-alpha_Binding_Sites ERα binding sites**. The UCSC genome BED formatted file lists the chromosome, start coordinate, stop coordinate and identifier of the ERα binding sites.

**Additional File 2: Grober_ER-beta_Binding_Sites ERβ binding sites**. The UCSC genome BED formatted lists the chromosome, start coordinate, stop coordinate and identifier of the ERβ binding sites.

**Additional File 3: Sheet 1: *Additional Table S1* Differentially Regulated Genes**. Overview of the 921 genes differentially regulated by E2 in ERβ+ *vs* ERβ- cells, containing the following additional information: gene set membership, symbol, Entrez ID, gene name and expression values (fold-change). Sheet 2: *Additional Table S2* TFBS enrichment matrix. The worksheet shows the over-representation values for the TF binding matrices from ERE- binding sites. Sheet 3: *Additional Table S3* Primary ERβ target genes, showing ERβ binding sites within 10 kb of the TU. Overview of the 424 putative ERβ primary gene targets containing the following additional information: gene set membership, category of ERβ binding sites, Symbol, Entrez ID, Gene Name, TU Coordinates, ERα Binding Sites Coordinates and ERβ Binding Sites Coordinates. Sheet 4: *Additional Table S4* Primary ERα target genes, showing ERα binding sites within 10 kb of the TU in ERβ- cells. Overview of the 228 putative ERα primary gene targets containing the following additional information: Symbol, Entrez ID, Gene Name, TU Coordinates.

**Additional File 4: GO analysis of primary ERβ target genes**. Containing the following information: Biological process, Gene Ontology term, Name, Count in total GO population, Count in selected genes, % genes and p-value.

**Additional File 5: GO analysis of primary ERα target genes**. Containing the following information: Biological process, Gene Ontology term, Name, Count in total GO population, Count in selected genes, % genes and p-value.

**Additional File 6: ERβ binding sites in proximity of miRNA loci**. Containing the following information: miRNA name, ID of ER-β binding site upstream, Distance from the closest ERβ binding site upstream, ID of ER-β binding site downstream and Distance from the closest ERβ binding site downstream.

#### Acknowledgements

Work supported by the European Union (CRESCENDO I.P., contract n.er LSHM-CT2005-018652), the Italian Association for Cancer Research (grant IG-8586) and the Italian Ministry for Education, University and Research (grant PRIN 2008CJ4SYW_004), Fondazione per il Sud (grant 2009-PdP-22) to AW. MR, GN and RT are recipient of fellowships from Fondazione per il Sud. The authors thank Dr Rosario Casale for technical assistance and Francesca Rizzo for revision of the manuscript.

#### Author details

[1]Department of General Pathology, Second University of Naples, vico L. De Crecchio 7, 80138 Napoli, Italy. [2]Molecular Medicine Laboratory, University of

Salerno, via Allende, 84081 Baronissi, Italy. [3]Illumina, Inc., Hayward, 94545 California, USA. [4]Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg 69117, Germany.

**Authors' contributions**

OMVG, MM, GG, MR, LC, MRDF, LF, GN, MFP, OP and RT participated in the conception and design of the study, performed *in vivo* and *in vitro* experimental work (OP, GN, MFP, RT), statistical and *in silico* data analyses (OMVG, MM, GG, MRDF) and participated in drafting the manuscript. OP, LC and MR prepared and tested the sequencing libraries. LC, LF and MR carried out the microarray experiments and participated in data analyses. SL, GPS and VB participated in the conception and design of the study and performed massively parallel sequencing. AW coordinated the project, participated in conception and design of the study and participated in drafting and finalization of the manuscript. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

Received: 23 August 2010 Accepted: 14 January 2011
Published: 14 January 2011

**References**

1. Heldring N, Pike A, Andersson S, Matthews J, Cheng G, Hartman J, Tujague M, Strom A, Treuter E, Warner M, Gustafsson JA: **Estrogen receptors: how do they signal and what are their targets.** *Physiol Rev* 2007, **87**:905-931.
2. Hall JM, Couse JF, Korach KS: **The multifaceted mechanisms of estradiol and estrogen receptor signaling.** *J Biol Chem* 2001, **276**:36869-36872.
3. Bai Z, Gust R: **Breast cancer, estrogen receptor and ligands.** *Arch Pharm (Weinheim)* 2009, **342**:133-149.
4. Ordonez-Moran P, Munoz A: **Nuclear receptors: genomic and non-genomic effects converge.** *Cell Cycle* 2009, **8**:1675-1680.
5. Nassa G, Tarallo R, Ambrosino C, Bamundo A, Ferraro L, Paris O, Ravo M, Guzzi PH, Cannataro M, Baumann M, Nyman TA, Nola E, Weisz A: **A large set of estrogen receptor beta-interacting proteins identified by tandem affinity purification in hormoneresponsive human breast cancer cell nuclei.** *Proteomics* .
6. Tarallo R, Bamundo A, Nassa G, Nola E, Paris O, Ambrosino C, Facchiano A, Baumann M, Nyman TA, Weisz A: **Identification of proteins associated with ligand-activated estrogen receptor alpha in human breast cancer cell nuclei by tandem affinity purification and nanoLC-MS/MS.** *Proteomics*.
7. Matthews J, Gustafsson JA: **Estrogen signaling: a subtle balance between ER alpha and ER beta.** *Mol Interv* 2003, **3**:281-292.
8. Klinge CM, Jernigan SC, Mattingly KA, Risinger KE, Zhang J: **Estrogen response element-dependent regulation of transcriptional activation of estrogen receptors alpha and beta by coactivators and corepressors.** *J Mol Endocrinol* 2004, **33**:387-410.
9. Cowley SM, Hoare S, Mosselman S, Parker MG: **Estrogen receptors alpha and beta form heterodimers on DNA.** *J Biol Chem* 1997, **272**:19858-19862.
10. Zhao C, Matthews J, Tujague M, Wan J, Strom A, Toresson G, Lam EW, Cheng G, Gustafsson JA, Dahlman-Wright K: **Estrogen receptor beta2 negatively regulates the transactivation of estrogen receptor alpha in human breast cancer cells.** *Cancer Res* 2007, **67**:3955-3962.
11. Nassa G, Tarallo R, Guzzi PH, Ferraro L, Cirillo F, Ravo M, Nola E, Baumann M, Nyman TA, Cannataro M, Ambrosino C, Weisz A: **Comparative analysis of nuclear estrogen receptor alpha and beta interactomes in breast cancer cells.** *Mol BioSyst* 2010.
12. Matthews J, Wihlén B, Tujague M, Wan J, Ström A, Gustafsson JA: **Estrogen receptor (ER) beta modulates ERalpha-mediated transcriptional activation by altering the recruitment of c-Fos and c-Jun to estrogen-responsive promoters.** *Mol Endocrinol* 2006, **20**:534-43.
13. Katzenellenbogen BS, Montano MM, Ediger TR, Sun J, Ekena K, Lazennec G, Martini PG, McInerney EM, Delage-Mourroux R, Weis K, Katzenellenbogen JA: **Estrogen receptors: selective ligands, partners, and distinctive pharmacology.** *Recent Prog Horm Res* 2000, **55**:163-93.
14. Pettersson K, Delaunay F, Gustafsson JA: **Estrogen receptor beta acts as a dominant regulator of estrogen signaling.** *Oncogene* 2000, **19**:4970-4978.
15. Sugiura H, Toyama T, Hara Y, Zhang Z, Kobayashi S, Fujii Y, Iwase H, Yamashita H: **Expression of estrogen receptor beta wild-type and its variant ERbetacx/beta2 is correlated with better prognosis in breast cancer.** *Jpn J Clin Oncol* 2007, **37**:820-828.
16. Saji S, Hirose M, Toi M: **Clinical significance of estrogen receptor β in breast cancer.** *Cancer Chemother Pharmacol* 2005, **56(Suppl 1)**:21-26.
17. Lindberg MK: **Estrogen Receptor (ER)-beta Reduces ERalpha-Regulated Gene Transcription, Supporting a "Ying Yang" Relationship between ERalpha and ERbeta in Mice.** *Mol Endocrinol* 2003, **17**:203-208.
18. Lazennec G, Bresson D, Lucas A, Chauveau C, Vignon F: **ER beta Inhibits Proliferation and Invasion of Breast Cancer Cells.** *Endocrinology* 2001, **142**:4120-4130.
19. Behrens D, Gill JH, Fichtner I: **Loss of tumourigenicity of stably ERbeta-transfected MCF-7 breast cancer cells.** *Mol Cell Endocrinol* 2007, **274**:19-29.
20. Paruthiyil S, Parmar H, Kerekatte V, Cunha GR, Firestone GL, Leitman DC: **Estrogen Receptor Inhibits Human Breast Cancer Cell Proliferation and Tumor Formation by Causing a G2 Cell Cycle Arrest.** *Cancer Res* 2004, **64**:423-428.
21. Chang EC, Frasor J, Komm B, Katzenellenbogen BS: **Impact of estrogen receptor beta on gene networks regulated by estrogen receptor alpha in breast cancer cells.** *Endocrinology* 2006, **147**:4831-4842.
22. Lin CY, Ström A, Li Kong S, Kietz S, Thomsen JS, Tee JB, Vega VB, Miller LD, Smeds J, Bergh J, Gustafsson JA, Liu ET: **Inhibitory effects of estrogen receptor beta on specific hormone-responsive gene expression and association with disease outcome in primary breast cancer.** *Breast Cancer Res* 2007, **9**:R25.
23. Williams C, Edvardsson K, Lewandowski SA, Ström A, Gustafsson J: **A genome-wide study of the repressive effects of estrogen receptor beta on estrogen receptor alpha signaling in breast cancer cells.** *Oncogene* 2008, **27**:1019-1032.
24. Fritzemeier KH, Hillisch A, Elger W, Kaufmann U, Kollenkirchen U, Kosemund D, Lindenthal B, Müller G, Muhn P, Nubbemeyer R, Peters O, Siebel P, Hegele-Hartung C: **Biological effects of ERalpha- and ERbeta-selective estrogens.** *Ernst Schering Res Found Workshop* 2004, **46**:127-150.
25. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M: **Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.** *Cell* 2005, **122**:33-43.
26. Cicatiello L, Mutarelli M, Grober OM, Paris O, Ferraro L, Ravo M, Tarallo R, Luo S, Schroth GP, Seifert M, Zinser C, Chiusano ML, Traini A, De Bortoli M, Weisz A: **Estrogen Receptor alpha Controls a Gene Network in Luminal-Like Breast Cancer Cells Comprising Multiple Transcription Factors and MicroRNAs.** *Am J Pathol* 2010, **176**:2113-2130.
27. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EG, Huang PY, Welboren WJ, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KD, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RK, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung WK, Liu ET, Wei CL, Cheung E, Ruan Y: **An oestrogen-receptor-alpha-bound human chromatin interactome.** *Nature* 2009, **462**:58-64.
28. Hurtado A, Holmes KA, Geistlinger TR, Hutcheson IR, Nicholson RI, Brown M, Jiang J, Howat WJ, Ali S, Carroll JS: **Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen.** *Nature* 2008, **456**:663-666.
29. Vivar OI, Zhao X, Saunier EF, Griffin C, Mayba OS, Tagliaferri M, Cohen I, Speed TP, Leitman DC: **Estrogen receptor beta binds to and regulates three distinct classes of target genes.** *J Biol Chem* 2010, **285**:22059-22066.
30. Charn TH, Liu ET, Chang EC, Lee YK, Katzenellenbogen JA, Katzenellenbogen BS: **Genome-wide dynamics of chromatin binding of estrogen receptors alpha and beta: mutual restriction and competitive site selection.** *Mol Endocrinol* 2010, **24**:47-59.
31. Zhao C, Gao H, Liu Y, Papoutsi Z, Jaffrey S, Gustafsson JA, Dahlman-Wright K: **Genome-wide mapping of estrogen receptor-beta-binding regions reveals extensive cross-talk with transcription factor activator protein-1.** *Cancer Res* 2010, **70**:5174-83.
32. Liu Y, Gao H, Marstrand TT, Ström A, Valen E, Sandelin A, Gustafsson JA, Dahlman-Wright K: **The genome landscape of ERalpha- and ERbeta-binding DNA regions.** *Proc Natl Acad Sci USA* 2008, **105**:2604-2609.
33. Ambrosino C, Tarallo R, Bamundo A, Cuomo D, Franci G, Nassa G, Paris O, Ravo M, Giovane A, Zambrano N, Lepikhova T, Jänne OA, Baumann M, Nyman TA, Cicatiello L, Weisz A: **Identification of a hormone-regulated**

dynamic nuclear actin network associated with estrogen receptor alpha in human breast cancer cell nuclei. *Mol Cell Proteomics* 2010, **9**:1352-1367.
34. Cicatiello L, Addeo R, Sasso A, Altucci L, Petrizzi VB, Borgo R, Cancemi M, Caporali S, Caristi S, Scafoglio C, Teti D, Bresciani F, Perillo B, Weisz A: **Estrogens and progesterone promote persistent CCND1 gene activation during G1 by inducing transcriptional derepression via c-Jun/c-Fos/estrogen receptor (progesterone receptor) complex assembly to a distal regulatory element and recruitment of cyclin D1 to its own gene promoter.** *Mol Cell Biol* 2004, **24**:7260-7274.
35. Frasor J, Stossi F, Danes JM, Komm B, Lyttle CR, Katzenellenbogen BS: **Selective estrogen receptor modulators: discrimination of agonistic versus antagonistic activities by gene expression profiling in breast cancer cells.** *Cancer Res* 2004, **64**:1522-1533.
36. Frasor J, Danes JM, Komm B, Chang KC, Lyttle CR, Katzenellenbogen BS: **Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype.** *Endocrinology* 2003, **144**:4562-4574.
37. Buterin T, Koch C, Naegeli H: **Convergent transcriptional profiles induced by endogenous estrogen and distinct xenoestrogens in breast cancer cells.** *Carcinogenesis* 2006, **27**:1567-1578.
38. Mardis ER: **ChIP-seq: welcome to the new frontier.** *Nature methods* 2007, **4**:613-614.
39. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9(9)**:R137.
40. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.
41. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A: **JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.** *Nucleic Acids Res* 2010, **38**:D105-10.
42. Mahony S, Auron PE, Benos PV: **DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies.** *PLoS Comput Biol* 2007, **3**:e61.
43. Safe S, Kim K: **Non-classical genomic estrogen receptor (ER)/specificity protein and ER/activating protein-1 signaling pathways.** *J Mol Endocrinol* 2008, **41**:263-275.
44. Vyhlidal C, Samudio I, Kladde MP, Safe S: **Transcriptional activation of transforming growth factor alpha by estradiol: requirement for both a GC-rich site and an estrogen response element half-site.** *J Mol Endocrinol* 2000, **24**:329-338.
45. Webb P, Nguyen P, Valentine C, Lopez GN, Kwok GR, McInerney E, Katzenellenbogen BS, Enmark E, Gustafsson JA, Nilsson S, Kushner PJ: **The estrogen receptor enhances AP-1 activity by two distinct mechanisms with different requirements for receptor transactivation functions.** *Mol Endocrinol* 1999, **13**:1672-1685.
46. Leung Y, Mak P, Hassan S, Ho S: **Estrogen receptor (ER)-beta isoforms: a key to understanding ER-beta signaling.** *Proc Natl Acad Sci USA* 2006, **103**:13162-13167.
47. Weigelt K, Moehle C, Stempfl T, Weber B, Langmann T: **An integrated workflow for analysis of ChIP-chip data.** *Biotechniques* 2008, **45**:131-132, 134, 136 passim.
48. Aulmann S, Bläker H, Penzel R, Rieker RJ, Otto HF, Sinn HP: **CTCF gene mutations in invasive ductal breast cancer.** *Breast Cancer Res Treat* 2003, **80**:347-352.
49. Mattingly KA, Ivanova MM, Riggs KA, Wickramasinghe NS, Barch MJ, Klinge CM: **Estradiol stimulates transcription of nuclear respiratory factor-1 and increases mitochondrial biogenesis.** *Mol Endocrinol* 2008, **22**:609-622.
50. Granok H, Leibovitch B, Shaffer C: **Ga-ga over GAGA factor.** *Curr Biol* 1995, **5**:238-241.
51. Van Steensel B, Delrow J, Bussemaker HJ: **Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding.** *Proc Natl Acad Sci USA* 2003, **100**:2580-2585.
52. Bauer S, Grossmann S, Vingron M, Robinson PN: **Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration.** *Bioinformatics* 2008, **24**:1650-1651.
53. Al-Nakhle H, Burns PA, Cummings M, Hanby AM, Hughes TA, Satheesha S, Shaaban AM, Smith L, Speirs V: **Estrogen receptor {beta}1 expression is regulated by miR-92 in breast cancer.** *Cancer Res* 2010, **70**:4778-4784.
54. Bhat-Nakshatri P, Wang G, Collins NR, Thomson MJ, Geistlinger TR, Carroll JS, Brown M, Hammond S, Srour EF, Liu Y, Nakshatri H: **Estradiol-regulated microRNAs control estradiol response in breast cancer cells.** *Nucleic Acids Res* 2009, **37**:4850-4861.
55. Castellano L, Giamas G, Jacob J, Coombes RC, Lucchesi W, Thiruchelvam P, Barton G, Jiao LR, Wait R, Waxman J, Hannon GJ, Stebbing J: **The estrogen receptor-alpha-induced microRNA signature regulates itself and its transcriptional response.** *Proc Natl Acad Sci USA* 2009, **106**:15732-15737.
56. Di Leva G, Gasparini P, Piovan C, Ngankeu A, Garofalo M, Taccioli C, Iorio MV, Li M, Volinia S, Alder H, Nakamura T, Nuovo G, Liu Y, Nephew KP, Croce CM: **MicroRNA Cluster 221-222 and Estrogen Receptor {alpha} Interactions in Breast Cancer.** *J Natl Cancer Inst* 2010, **102**:706-721.
57. Maillot G, Lacroix-Triki M, Pierredon S, Gratadou L, Schmidt S, Bénès V, Roché H, Dalenc F, Auboeuf D, Millevoi S, Vagner S: **Widespread estrogen-dependent repression of micrornas involved in breast tumor cell growth.** *Cancer Res* 2009, **69**:8332-8340.
58. Wickramasinghe NS, Manavalan TT, Dougherty SM, Riggs KA, Li Y, Klinge CM: **Estradiol downregulates miR-21 expression and increases miR-21 target gene expression in MCF-7 breast cancer cells.** *Nucleic Acids Res* 2009, **37**:2584-2595.
59. Adams BD, Guttilla IK, White BA: **Involvement of microRNAs in breast cancer.** *Semin Reprod Med* 2008, **26**:522-536.
60. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Ménard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM: **MicroRNA gene expression deregulation in human breast cancer.** *Cancer Res* 2005, **65**:7065-7070.
61. Iorio MV, Casalini P, Tagliabue E, Ménard S, Croce CM: **MicroRNA profiling as a tool to understand prognosis, therapy response and resistance in breast cancer.** *Eur J Cancer* 2008, **44**:2753-2759.
62. Yang SH, Liu R, Perez EJ, Wen Y, Stevens SM Jr, Valencia T, Brun-Zinkernagel AM, Prokai L, Will Y, Dykens J, Koulen P, Simpkins JW: **Mitochondrial localization of estrogen receptor beta.** *Proc Natl Acad Sci USA* 2004, **101**:4130-4135.
63. Chen JQ, Eshete M, Alworth WL, Yager JD: **Binding of MCF-7 cell mitochondrial proteins and recombinant human estrogen receptors alpha and beta to human mitochondrial DNA estrogen response elements.** *J Cell Biochem* 2004, **93**:358-373.
64. Chen J, Russo PA, Cooke C, Russo IH, Russo J: **ERbeta shifts from mitochondria to nucleus during estrogen-induced neoplastic transformation of human breast epithelial cells and is involved in estrogen-induced synthesis of mitochondrial respiratory chain proteins.** *Biochim Biophys Acta* 2007, **1773**:1732-1746.
65. Felty Q, Singh KP, Roy D: **Estrogen-induced G1/S transition of G0-arrested estrogen-dependent breast cancer cells is regulated by mitochondrial oxidant signaling.** *Oncogene* 2005, **24**:4883-4893.
66. Simpkins JW, Yang S, Sarkar SN, Pearce V: **Estrogen actions on mitochondria–physiological and pathological implications.** *Mol Cell Endocrinol* 2008, **290**:51-59.
67. Blobel GA, Sieff CA, Orkin SH: **Ligand-dependent repression of the erythroid transcription factor GATA-1 by the estrogen receptor.** *Mol Cell Biol* 1995, **6**:3147-3153.
68. Mao L, Wertzler KJ, Maloney SC, Wang Z, Magnuson NS, Reeves R: **HMGA1 levels influence mitochondrial function and mitochondrial DNA repair efficiency.** *Mol Cell Biol* 2009, **29**:5426-5440.
69. Massaad-Massade L, Tacine R, Dulauroy S, Reeves R, Barouki R: **The functional interaction between HMGA1 and the estrogen receptor requires either the N- or the C-terminal domain of the receptor.** *FEBS Lett* 2004, **559**:89-95.
70. Pace P, Taylor J, Suntharalingam S, Coombes RC, Ali S: **Human estrogen receptor beta binds DNA in a manner similar to and dimerizes with estrogen receptor alpha.** *J Biol Chem* 1997, **272**:25832-25838.
71. Pettersson K, Grandien K, Kuiper GG, Gustafsson JA: **Mouse estrogen receptor beta forms estrogen response element-binding heterodimers with estrogen receptor alpha.** *Mol Endocrinol* 1997, **11**:1486-1496.
72. Li X, Huang J, Yi P, Bambara RA, Hilf R, Muyan M: **Single-chain estrogen receptors (ERs) reveal that the ERalpha/beta heterodimer emulates functions of the ERalpha dimer in genomic estrogen signaling pathways.** *Mol Cell Biol* 2004, **24**:7681-7694.

73. Andrews NC, Faller DV: **A rapid micropreparation technique for extraction of DNA-binding proteins from limiting numbers of mammalian cells.** *Nucleic Acids Res* 1991, **19**:2499.
74. Frezza C, Cipolat S, Scorrano L: **Organelle isolation: functional mitochondria from mouse liver, muscle and cultured fibroblasts.** *Nat Protoc* 2007, **2**:287-295.
75. Caporali S, Imai M, Altucci L, Cancemi M, Caristi S, Cicatiello L, Matarese F, Penta R, Sarkar DK, Bresciani F, Weisz A: **Distinct signaling pathways mediate stimulation of cell cycle progression and prevention of apoptotic cell death by estrogen in rat pituitary tumor PR1 cells.** *Mol Biol Cell* 2003, **14**:5051-5059.
76. Bermont L, Lamielle-Musard F, Chezy E, Weisz A, Adessi GL: **17beta-estradiol inhibits forskolin-induced vascular endothelial growth factor promoter in MCF-7 breast adenocarcinoma cells.** *J Steroid Biochem Mol Biol* 2001, **78**:343-349.
77. Licznar A, Caporali S, Lucas A, Weisz A, Vignon F, Lazennec G: **Identification of genes involved in growth inhibition of breast cancer cells transduced with estrogen receptor.** *FEBS Lett* 2003, **553**:445-450.
78. Bonapace IM, Addeo R, Altucci L, Cicatiello L, Bifulco M, Laezza C, Salzano S, Sica V, Bresciani F, Weisz A: **17 beta-Estradiol overcomes a G1 block induced by HMG-CoA reductase inhibitors and fosters cell cycle progression without inducing ERK-1 and -2 MAP kinases activation.** *Oncogene* 1996, **12**:753-763.
79. Cicatiello L, Scafoglio C, Altucci L, Cancemi M, Natoli G, Facchiano A, Iazzetti G, Calogero R, Biglia N, De Bortoli M, Sfiligoi C, Sismondi P, Bresciani F, Weisz A: **A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome.** *J Mol Endocrinol* 2004, **32**:719-775.
80. Scafoglio C, Ambrosino C, Cicatiello L, Altucci L, Ardovino M, Bontempo P, Medici N, Molinari AM, Nebbioso A, Facchiano A, Calogero RA, Elkon R, Menini N, Ponzone R, Biglia N, Sismondi P, De Bortoli M, Weisz A: **Comparative gene expression profiling reveals partially overlapping but distinct genomic actions of different antiestrogens in human breast cancer cells.** *J Cell Biochem* 2006, **98**:1163-1184.
81. Ravo M, Mutarelli M, Ferraro L, Grober OM, Paris O, Tarallo R, Vigilante A, Cimino D, De Bortoli M, Nola E, Cicatiello L, Weisz A: **Quantitative expression profiling of highly degraded RNA from formalin-fixed, paraffin-embedded breast tumor biopsies by oligonucleotide microarrays.** *Lab Invest* 2008, **88**:430-440.
82. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, Klingenhoff A, Frisch M, Bayerlein M, Werner T: **MatInspector and beyond: promoter analysis based on transcription factor binding sites.** *Bioinformatics* 2005, **21**:2933-2942.
83. Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW: **oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes.** *Nucleic Acids Res* 2005, **33**:3154-64.
84. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: **TM4: a free, open-source system for microarray data management and analysis.** *Biotechniques* 2003, **34**:374-378.
85. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37(Database issue)**:D755-761.

doi:10.1186/1471-2164-12-36
**Cite this article as:** Grober *et al.*: Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. *BMC Genomics* 2011 **12**:36.

# Società Italiana di Genetica Umana

## XV Congresso Nazionale SIGU

Sorrento 21-24 Novembre 2012

**Programma**

P313

**DEVELOPMENT OF AN ACCURATE PIPELINE FOR EXOME SEQUENCING DATA ANALYSIS**

M.R. De Filippo[1], G. Giurato[2], F. Rizzo[2], D. Greco[3], R. Katainen[4], L. Aaltonen[4], G. Coppola[5], A. Weisz[2]

[1]*Fondazione IRCCS SDN, Napoli, Italy*

[2]*Laboratory of Molecular Medicine and Genomics and UOC Molecular Pathology and Medical Genomics, University of Salerno, Faculty of Medicine, Salerno, Italy*

[3]*Research Unit of Molecular Medicine, University of Helsinki, Helsinki, Finland and Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden*

[4]*Department of Medical Genetics, Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland*

[5]*Department of Medicine and Surgery, University of Salerno, Salerno, Italy*

Exome sequencing - targeted sequencing of coding regions of the genome - is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable with conventional gene-discovery strategies. Until now many algorithms have been produced, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. The aim of this work is to combine these algorithms into an accurate analysis pipeline to identify high quality variations in the data produced in our laboratory by exome capture using the Agilent SureSelect 50Mb exome capture kit and massively parallel sequencing with an Illumina GAIIx sequencer. After sequencing, paired-end reads were aligned to the hg19 reference genome using BWA [1] allowing more than one mismatch, and processed using SAMtools [2]. Base quality score recalibration and local realignment around indels were performed using the Genome Analysis Toolkit GATK [3] and PCR duplicates were removed using Picard tool [4]. SNPs calling was done using GATK UnifiedGenotyper, that applies a Bayesian model to estimate the most likely genotypes and allele frequency in a population of N samples, giving an annotated VCF file as output. Only variations supported by a number of reads greater than 8, calls greater than 25 and quality score greater than 40, were considered for the next steps. Subsequently, variants have been marked as missense or synonymous and then dbSNP was used to discard SNPs that were already known. Finally, the tendency of each missense mutation to be deleterious for the function of a protein as opposed to neutral was calculated using CONDEL [5], a software that computes a weighted average of the scores of the SIFT and POLYPHEN methods. The pipeline was tested on a sample derived from a human cell line sequenced in our laboratory. Starting from 165.791 raw variants, we applied all filters described above reducing the list to 3.154 variations, ~1.000 of which were marked as deleterious by CONDEL and manually annotated according to the literature to identify variations of potential interest for the disease.

# German-Italian Dialogue 2012

## "Next Generation Sequencing - Application cases and bioinformatics development"

17 – 19 July 2012

Naples, Italy

# AN ACCURATE PIPELINE FOR ANALYSIS OF EXOME SEQUENCING DATA

**De Filippo Maria Rosaria**[1,2], Giurato Giorgio[1], Rizzo Francesca[1], Ravo Maria[1], Stellato Claudia[1], Marchese Giovanna[1] and Weisz Alessandro[1,3]

1) Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, Salerno, Italy, 2) Fondazione IRCCS SDN, Napoli, Italy, 3) Division of Molecular Pathology and Medical Genomics, 'SS Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Salerno, Italy.

Exome sequencing - the targeted sequencing of the subset of the protein coding human genome - is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable with conventional gene-discovery strategies. Until now many algorithms have been produced, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. The aim of this work is to combine these algorithms into an analysis pipeline to identify high quality variations in the data produced in our laboratory by an exome sequencing experiment using the illumina GAIIx sequencer. Exonic sequences were targeted using the Agilent SureSelect 50Mb exome capture kit and sequenced using the Illumina GAIIx. 72 bp paired-end reads were aligned to the hg19 reference genome using BWA and processed using SAMtools. Base quality score recalibration and local realignment around indels were performed using the GATK. Duplicate marking was conducted using Picard. SNP calling was done using GATK UnifiedGenotyper, that applies a Bayesian model to estimate the most likely genotypes and allele frequency in a population of N samples, giving an annotated VCF file as output. Only variations with coverage greater than 8, calls greater than 25 and quality score greater than 40, were considered for the next steps. Subsequently variants have been marked as missense or synonymous and dbSNP was used to discard SNPs that were already known. Finally, the tendency of each missense mutation to be deleterious for the function of a protein as opposed to neutral was calculated using CONDEL, a software that computes a weighted average of the scores of the SIFT and POLYPHEN methods. We tested our pipeline on a sample derived from human breast cancer MCF-7 cells sequenced in our laboratory. Starting from 165791 raw variants, after filtering for coverage, quality and call, and removing known variants from dbSNP together with synonymous ones causing no amino acid change, we obtained a list of 3154 variations (missense and truncating). After CONDEL prediction, 1140 variants were marked as deleterious and about 50 of them were found implicated in breast cancer by literature.

**ACKNOWLEDGEMENTS.**
Research supported by: AIRC (Grant IG-8586), MIUR (PRIN 2008CJ4SYW_004), Regione Campania, University of Salerno (FARB 2011), Fondazione con il Sud, EU COST (Action BM1006 'SeqAhead'), Fondazione Veronesi. Maria Rosaria De Filippo is a student of PhD School in Computational Biology and Bioinformatics, University of Naples 'Federico II' (Italy)

EMBnet.journal
Volume 18
Supplement A
April 2012

# BITS 2012 - Ninth Annual Meeting of the Bioinformatics Italian Society

## Meeting Abstracts

2-4 May 2012, Catania, Italy

# Development of pipeline for exome sequencing data analysis

**M.R. De Filippo[1] ✉, G. Giurato[1], C. Cantarella[1], F. Rizzo[1], F. Cirillo[1], A. Weisz[2]**

[1]Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, IT, Naples, Italy
[2]Division of Molecular Pathology and Medical Genomics, 'SS Giovanni di Dio e Ruggi d'Aragona' Hospital, University of Salerno, Italy

## Motivations

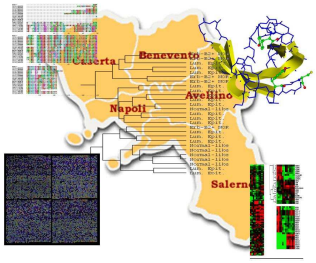
Exome sequencing the targeted sequencing of the subset of the protein coding human genome is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Until now many algorithms have been produced, each of them addressing a different task in the downstream analysis of next-generation sequencing (NGS) data. The aim of this work is to combine these algorithms into an analysis pipeline for the detection of SNP and deletion/insertion polymorphisms within DNA sequences obtained by whole exome sequencing. The pipeline tested with data obtained from SRA (http://www.ncbi.nlm.nih.gov/sra), will then be applied to studies undergoing in our laboratory.

## Methods

Starting from raw sequence data, we first performed quality statistics and filtering of sequence reads and then aligned them to a reference genome. To this end, BWA was used to align both single- and paired-end reads for its computational efficiency and multi-platform compatibility. Post-alignment analysis, including removal of duplicate reads and quality score recalibration, was carried out using GATK, which takes into account several covariates such as machine cycle and dinucleotide context. Next, SNP calling was done using GATK UnifiedGenotyper, that uses a Bayesian model to estimate the most likely genotypes and allele frequency in a population of N samples, giving an annotated VCF file as output. Subsequently, variant quality score was recalibrated to estimate the probability of each variant being a true polymorphism, rather than a sequencer, alignment or data processing artifact, and finally filtered to improve the accuracy of genotype and SNP calling.

## Results

The results obtained support the accuracy of our pipeline to identify SNP and short indels, to provide a global and quantitative catalog of nucleotide variants in the exome. The next step will be to apply this pipeline to samples sequenced in our laboratory.




# BBCC2011

## Sixth Annual Meeting

### Bioinformatica e Biologia Computazionale in Campania

November 4th, 2011
Avellino, Italy
Istituto di Scienze dell'Alimentazione
Consiglio Nazionale delle Ricerche

**http://bioinformatica.isa.cnr.it/BBCC/BBCC2011/**

**Computational approaches for genome-wide mRNA and miRNA expression profiling in human breast cancer cell lines expressing (ERβ+) or lacking (ERβ-) estrogen receptor beta by microarray hybridization and massively parallel sequencing (miRNA-Seq)**

Maria Rosaria De Filippo (1), Giorgio Giurato (1), Roberta Tarallo (1), Maria Ravo (1), Francesca Rizzo (1), Concita Cantarella (1), Giovanni Nassa (1),, Ernesto Nola (2) and Alessandro Weisz (1,2)

*(1) Laboratory of Molecular Medicine and Genomics, Faculty Medicine and Surgery of the University of Salerno, Baronissi (SA), Italy*
*(2) Department of General Pathology of the Second University of Naples, Napoli, Italy*

microRNAs (miRNAs) are evolutionary conserved small non coding RNA that negatively regulate gene expression. Recent studies have demonstrated that mutations or aberrant expression of miRNAs are associated with cancer, suggesting that genes encoding these RNAs may act as oncogenes or tumor suppressors. Estrogen receptor alpha (ERα) and beta (ERβ) are transcriptional factors (TFs) that mediate estrogen signaling and define the hormone responsive phenotype of breast cancer. The two receptors can be found co-expressed, and play specific, often opposite roles, with ERβ being able to modulate the effect of ERα on gene transcription and cell proliferation. ChIP-Seq (Chromatin immunoprecipitation followed by sequencing) analysis of breast cancer cell lines (MCF7) showed that ERα and ERβ bind in close proximity of several miRNA genes, suggesting a direct involvement of these nuclear receptors in biogenesis of these small RNAs[1,2].Starting from these observations, I investigated miRNA expression patterns by miRNA-Seq (direct sequencing of small non coding RNA) and microarray hybridization in two ERα positive breast cancer cell lines: one lacking (ERβ-) and one expressing (ERβ+) estrogen receptor β. At first, I analyzed data obtained from microarray experiments (Agilent Human microRNA microarray). After normalization, student t-test was performed to identify differentially expressed miRNA between the two cell lines[3]. Subsequently, I focused on analyzing miRNA-seq data from the same cell lines. To this end, reads obtained from sequencing were analyzed with a specific bioinformatic tool, miRAnalyzer. The R Bioconductor's package, DeSeq, was then used to perform differential expression analysis for sequence count data between ERβ- and ERβ+ cells. Comparison of data obtained from microarray analyses and miRNA-Seq was carried out to evaluate the reliability, sensitivity and reproducibility of the two analytical tools. Finally, differentially expressed miRNAs was used to search for their putative target mRNAs through the use of dedicated bioinformatics tools. The results of this analysis were exploited to reveal biological functions and molecular processes in which these miRNA targets are involved and are therefore controlled at a post-transcriptional levels by ERβ *via* specific miRNAs.

REFERENCES

[1] Cicatiello L, Mutarelli M, Grober OM, Paris O, Ferraro L, Ravo M, Tarallo R, Luo S, Schroth GP, Seifert M, Zinser C, Chiusano ML, Traini A, De Bortoli M, Weisz A. Estrogen receptor alpha controls a gene network in luminal-like breast cancer cells comprising multiple transcription factors and microRNAs. *Am J Pathol. 2010 May;176(5):2113-30.*
[2] Grober OM, Mutarelli M, Giurato G, Ravo M, Cicatiello L, De Filippo MR, Ferraro L, Nassa G, Papa MF, Paris O, Tarallo R, Luo S, Schroth GP, Benes V, Weisz A. Global analysis of estrogen receptor beta binding to breast cancer cell genome reveals an extensive interplay with estrogen receptor alpha for target gene regulation. *BMC Genomics. 2011 Jan 14;12:36.*
[3] Paris O, Ferraro L, Grober OM, Ravo M, De Filippo MR, Giurato G, Nassa G, Tarallo R, Cantarella C, Rizzo F, Di Benedetto A, Mottolese M, Benes V, Ambrosino C, Nola E, Weisz A. Direct Regulation of microRNA Biogenesis and Expression by Estrogen Receptor Beta in Hormone-responsive Breast Cancer. *Oncogene 2011. In press.*