Applying microarray-based techniques to study gene expression patterns: a bio-computational approach

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Yevhen Vainshtein
aus Heidelberg

Würzburg 2010

Eingereicht am: ....

Mitglieder der Promotionskommission:

Vorsitzender: ....
Gutachter: Professor Dr. Thomas Dandekar
Gutachter: Professor Dr Martina Muckenthaler

Tag des Promotionskolloquiums: ....

Doktorurkunde ausgehändigt am: ....

For my whole family and friends who were very supportive all the time

Contents

Summary 1

Zusammenfassung 3

Introduction 6

  Iron homeostasis and the IRE/IRP regulatory system 6

  Identification of mRNAs containing IRE-like motifs 9

  Microarray analysis 10

    Principle 10

    Experimental system 11

    Variability of the data 11

    Microarray data evaluation 12

    IronChip 13

    Expression analysis of IRE/IRP regulatory network 14

  Microarray data evaluation 14

Personal contributions 15

Materials and Methods 16

  System requirements 16

  IronChip data sets 16

  Strategy to Identify Novel IRE-containing mRNAs 17

  Biocomputational Identification of Novel IRE-like Motifs 21

  IronChip Analysis of IRP/IRE mRNPs 21

Results 22

  Identification of novel IRE (Cdc14a) 22

  Iron Regulation of CDC14A mRNA 25

  IronChip Evaluation Package (ICEP) 26

  Implementation 26

  Data formats 27

  Performance 28

User interface 28
Application of the analysis pipeline 29
Using IronChip and ICEP for gene expression studies 37
Iron metabolism and gene expression analysis 43
Discussion 46
Identification of a novel IRE-containing genes 46
Cdc14a – possible role in iron homeostasis 46
Developing of new IRE motifs identification methods. 49
IronChip Evaluation Package (ICEP) – novel microarray analysis tool 50
Challenges to understand iron network 51
Microarrays-based techniques to improve a disease diagnosis 52
Conclusions 52
References 53
Abbreviations used 59
Supplementary materials 60
List of Publications 60
Conferences presentations 60
Curriculum vitae 61
Acknowledgements 62

Summary

Background

The regulation and maintenance of iron homeostasis is critical to human health. As a constituent of hemoglobin, iron is essential for oxygen transport and significant iron deficiency leads to anemia. Eukaryotic cells require iron for survival and proliferation. Iron is part of hemoproteins, iron-sulfur (Fe-S) proteins, and other proteins with functional groups that require iron as a cofactor.

At the cellular level, iron uptake, utilization, storage, and export are regulated at different molecular levels (transcriptional, mRNA stability, translational, and posttranslational). Iron regulatory proteins (IRPs) 1 and 2 post-transcriptionally control mammalian iron homeostasis by binding to iron-responsive elements (IREs), conserved RNA stem-loop structures located in the 5’- or 3‘- untranslated regions of genes involved in iron metabolism (e.g. FTH1, FTL, and TFRC). To identify novel IRE-containing mRNAs, we integrated biochemical, biocomputational, and microarray-based experimental approaches.

Gene expression studies greatly contribute to our understanding of complex relationships in gene regulatory networks. However, the complexity of array design, production and manipulations are limiting factors, affecting data quality. The use of customized DNA microarrays improves overall data quality in many situations, however, only if for these specifically designed microarrays analysis tools are available.

Methods

In this project response to the iron treatment was examined under different conditions using bioinformatical methods. This would improve our understanding of an iron regulatory network. For these purposes we used microarray gene expression data.

To identify novel IRE-containing mRNAs biochemical, biocomputational, and microarray-based experimental approaches were integrated. IRP/IRE messenger ribonucleoproteins were immunoselected and their mRNA composition was analysed using an IronChip microarray enriched for genes predicted computationally to contain IRE-like motifs.

Analysis of IronChip microarray data requires specialized tool which can use all advantages of a customized microarray platform. Novel decision-tree based algorithm was implemented using Perl in IronChip Evaluation Package (ICEP).

Results

IRE-like motifs were identified from genomic nucleic acid databases by an algorithm combining primary nucleic acid sequence and RNA structural criteria. Depending on the choice of constraining criteria, such computational screens tend to generate a large number of false positives. To refine the search and reduce the number of false positive hits, additional constraints were introduced. The refined screen yielded 15 IRE-like motifs. A second approach made use of a reported list of 230 IRE-like sequences obtained from screening UTR databases. We selected 6 out of these 230 entries based on the ability of the lower IRE stem to form at least 6 out of 7 bp. Corresponding ESTs were spotted onto the human or mouse versions of the IronChip and the results were analysed using ICEP. Our data show that the immunoselection/microarray strategy is a feasible approach for screening bioinformatically predicted IRE genes and the detection of novel IRE-containing mRNAs. In addition, we identified a novel IRE-containing gene CDC14A (Sanchez M, et al. 2006).

The IronChip Evaluation Package (ICEP) is a collection of Perl utilities and an easy to use data evaluation pipeline for the analysis of microarray data with a focus on data quality of custom-designed microarrays. The package has been developed for the statistical and bioinformatical analysis of the custom cDNA microarray IronChip, but can be easily adapted for other cDNA or oligonucleotide-based designed microarray platforms. ICEP uses decision tree-based algorithms to assign quality flags and performs robust analysis based on chip design properties regarding multiple repetitions, ratio cut-off, background and negative controls (Vainshtein Y, et al., 2010).

References

Vainshtein Y, Sanchez M, Brazma A, Hentze MW, Dandekar T, Muckenthaler MU. The IronChip evaluation package: a package of perl modules for robust analysis of custom microarrays. BMC Bioinformatics. 2010 Mar 1;11:112.

Sanchez M, Galy B, Dandekar T, Bengert P, **Vainshtein Y**, Stolte J, Muckenthaler MU, Hentze MW. Iron regulation and the cell cycle: identification of an iron-responsive element in the 3'-untranslated region of human cell division cycle 14A mRNA by a refined microarray-based screening strategy. J Biol Chem. 2006 Aug 11;281(32):22865-7

Zusammenfassung

Hintergrund

Die Regulierung und Aufrechterhaltung der Eisen-Homeostase ist bedeutend für die menschliche Gesundheit. Als Bestandteil des Hämoglobins ist es wichtig für den Transport von Sauerstoff, ein Mangel führt zu Blutarmut. Eukaryotische Zellen benötigen Eisen zum Überleben und zum Proliferieren. Eisen ist am Aufbau von Hämo- und Eisenschwefelproteinen (Fe-S) beteiligt und kann als Kofaktor dienen.

Die Aufnahme, Nutzung, Speicherung und der Export von Eisen ist zellulär auf verschiedenen molekularen Ebenen reguliert (Transkription, mRNA-Level, Translation, Protein-Level). Die iron regulatory proteins (IRPs) 1 und 2 kontrollieren die Eisen-Homeostase in Säugetieren posttranslational durch die Bindung an Iron-responsive elements (IREs). IREs sind konservierte RNA stem-loop Strukturen in den 5' oder 3' untranslatierten Bereichen von Genen, die im Eisenmetabolismus involviert sind (z.B. FTH1, FTL und TFRC). In dieser Arbeit wurden biochemische und bioinformatische Methoden mit Microarray-Experimenten kombiniert, um neue mRNAs mit IREs zu identifizieren.

Genexpressionsstudien verbessern unser Verständnis über die komplexen Zusammenhänge in genregulatorischen Netzwerken. Das komplexe Design von Microarrays, deren Produktion und Manipulation sind dabei die limitierenden Faktoren bezüglich der Datenqualität. Die Verwendung von angepassten DNA Microarrays verbessert häufig die Datenqualität, falls entsprechende Analysemöglichkeiten für diese Arrays existieren.

Methoden

Um unser Verständnis von eisenregulierten Netzwerken zu verbessern, wurde im Rahmen dieses Projektes die Auswirkung einer Behandlung mit Eisen bzw. von Knockout Mutation unter verschiedenen Bedingungen mittels bioinformatischer Methoden untersucht. Hierfür nutzen wir Expressionsdaten aus Microarray-Experimenten.

Durch die Verknüpfung von biochemischen, bioinformatischen und Microarray Ansätzen können neue Proteine mit IREs identifiziert werden. IRP/IRE messenger Ribonucleoproteine wurden immunpräzipitiert. Die Zusammensetzung der enthaltenen mRNAs wurde mittels einem IronChip Microarray analysiert: Für diesen Chip wurden bioinformatisch Gene vorhergesagt, die

IRE-like Motive aufweisen. Der Chip wurde mit solchen Oligonucleotiden beschichtet und durch Hybridisierung überprüft, ob die präzipitierten mRNA sich hieran binden.

Die Analyse der erhaltenen Daten erfordert ein spezialisiertes Werkzeug um von allen Vorteilen der angepassten Microarrays zu profitieren. Ein neuer Entscheidungsbaum-basierter Algorithmus wurde in Perl im IronChip Evaluation Package (ICEP) implementiert.

Ergebnisse

Aus großen Sequenz-Datenbanken wurden IRE-like Motive identifiziert. Dazu kombiniert der Algorithmus, insbesondere RNA-Primärsequenz und RNA-Strukturdaten.

Solche Datenbankanalysen tendieren dazu, eine große Anzahl falsch positiver Treffer zu generieren. Daher wurden zusätzliche Bedingungen formuliert, um die Suche zu verfeinern und die Anzahl an falsch positiven Treffer zu reduzieren. Die angepassten Suchkriterien ergaben 15 IRE-like Motive. In einem weiteren Ansatz verwendeten wir eine Liste von 230 IRE-like Sequenzen aus UTR-Datenbanken. Daraus wurden 6 Sequenzen ausgewählt, die auch im unteren Teil stabil sind (untere Helix über 6 bp stabil). Die korrespondierenden Expressed Sequence Tags (ESTs) wurden auf die humane oder murine Version des IronChips aufgetragen. Die Microarray Ergebnisse wurden mit dem ICEP Programm ausgewertet. Unsere Ergebnisse zeigen, dass die Immunpräzipitation mit anschließender Microarrayanalyse ein nützlicher Ansatz ist, um bioinformatisch vorhergesagte IRE-Gene zu identifizieren. Darüber hinaus ermöglicht uns dieser Ansatz die Detektion neuer mRNAs, die IREs enthalten, wie das von uns gefundene Gen CDC14A (Sanchez et al., 2006).

ICEP ist ein optimiertes Programmpaket aus Perl Programmen (Vainshtein et al., BMC Bioinformatics, 2010). Es ermöglicht die einfache Auswertung von Microarray Daten mit dem Fokus auf selbst entwickelten Microarray Designs. ICEP diente für die statistische und bioinformatische Analyse von selbst entwickelten IronChips, kann aber auch leicht an die Analyse von oligonucleotidbasierten oder cDNA Microarrays adaptiert werden. ICEP nutzt einen Entscheidungsbaum-basierten Algorithmus um die Qualität zu bewerten und führt eine robuste Analyse basierend auf Chipeigenschaften, wie mehrfachen Wiederholungen, Signal/Rausch Verhältnis, Hintergrund und Negativkontrollen durch.

Literaturliste

Vainshtein Y, Sanchez M, Brazma A, Hentze MW, Dandekar T, Muckenthaler MU. The IronChip evaluation package: a package of Perl modules for robust analysis of custom microarrays. BMC Bioinformatics. 2010 Mar 1;11:112.

Sanchez M, Galy B, Dandekar T, Bengert P, **Vainshtein Y**, Stolte J, Muckenthaler MU, Hentze MW. Iron regulation and the cell cycle: identification of an iron-responsive element in the 3'-untranslated region of human cell division cycle 14A mRNA by a refined microarray-based screening strategy. J Biol Chem. 2006 Aug 11;281(32):22865-7

Introduction

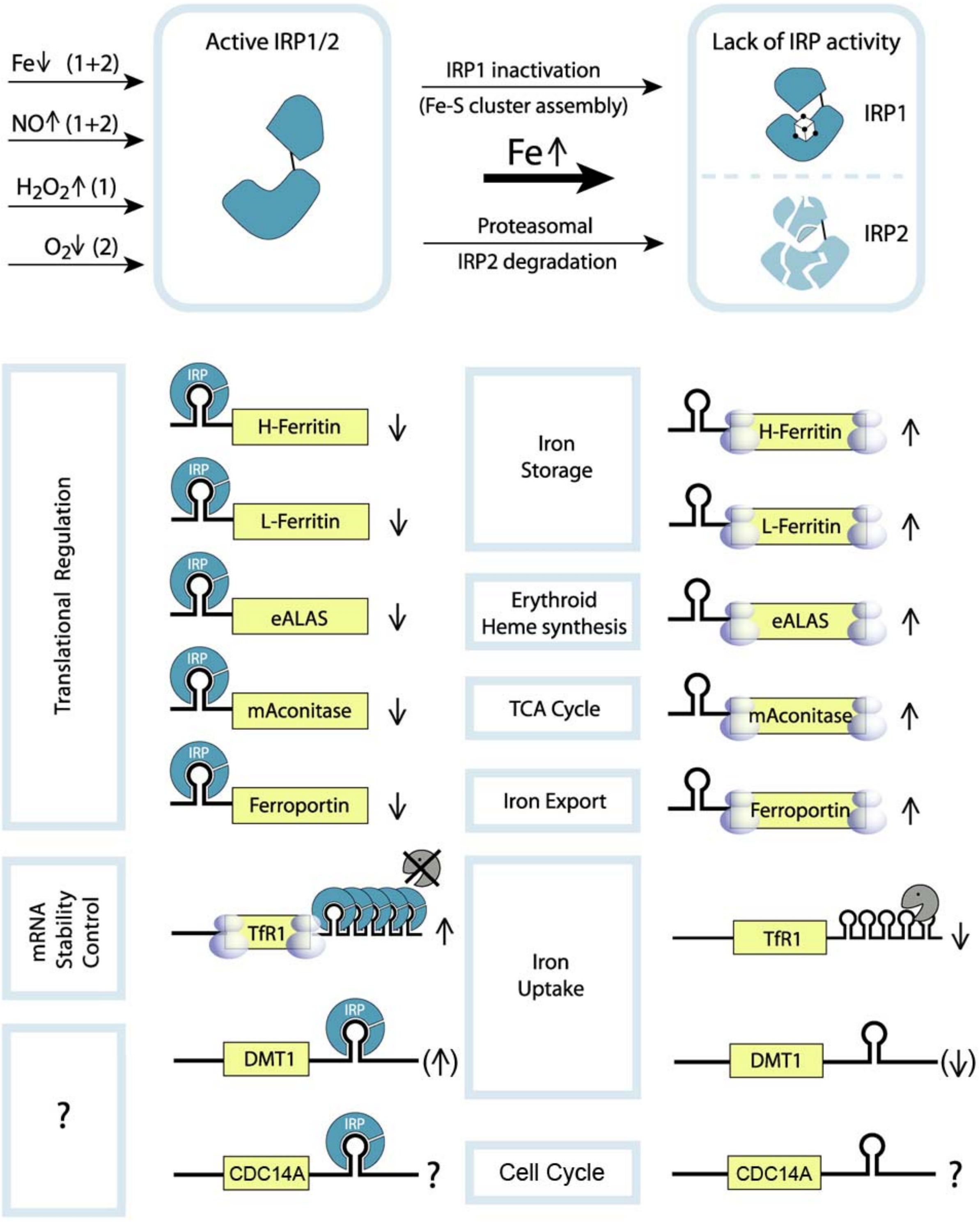
Iron homeostasis and the IRE/IRP regulatory system

The regulation and maintenance of iron homeostasis is critical to human health. As a constituent of hemoglobin, iron is essential for oxygen transport and significant iron deficiency leads to anemia. Eukaryotic cells require iron for survival and proliferation. Iron is part of hemoproteins, iron-sulfur (Fe-S) proteins, and other proteins with functional groups that require iron as a cofactor. These proteins carry out essential housekeeping functions for cellular metabolism; hence, iron deficiency leads to cell arrest and death. Conversely, iron excess is toxic because a fraction of “free” and reactive ferrous iron reacts with hydrogen- or lipid- peroxides to generate hydroxyl or lipid radicals that damage lipid membranes, proteins and nucleic acids. Since both iron deficiency and iron overload are detrimental to the cell, the levels of reactive iron must be carefully controlled and limited to prevent clinical disorders due to anomalies in iron metabolism.

Systemic iron homeostasis is largely dependent on the iron hormone hepcidin that controls iron fluxes through tissues responsible for dietary iron absorption (the proximal intestine), iron recycling (tissue macrophages) or storage (hepatocytes).

At the cellular level, iron uptake, utilization, storage, and export are regulated at different molecular levels (transcriptional, mRNA stability, translational, and posttranslational). Posttranscriptional regulation is the best characterized mechanism (Figure 1). Trans-acting iron regulatory proteins (IRP)-1 and -2 interact with iron-responsive elements (IREs), which are conserved hairpin structures (Figure 2) found in untranslated regions (UTRs) of mRNAs encoding iron-related proteins. The binding of IRP1 and IRP2 to IREs is regulated by the cellular labile iron pool through distinct mechanisms. When iron levels are high, a cubane [4Fe-4S] cluster assembles in IRP1, inhibiting IRE binding activity and converting IRP1 to an aconitase. When cellular iron stores are low, IRP1 binds to IRE targets as an apoprotein. In contrast, IRP2 does not contain a Fe-S cluster. It accumulates in iron-deficient cells and is targeted for degradation when iron levels are high. Single IREs are located in the 5′UTRs of mRNAs encoding ferritin H and L chains (iron storage proteins), erythroid 5-aminolevulinic acid synthase (the first enzyme of heme biosynthesis), mitochondrial aconitase (a citrate cycle enzyme), and ferroportin (iron export). The formation of IRE/IRP complexes on the 5′ UTR inhibits early steps of translation. On the other hand, the binding of IRPs to the multiple IREs present in the 3’UTR of the transferrin receptor 1 (TfR1, iron uptake protein) mRNA increases its stability.

Figure 1. The iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network



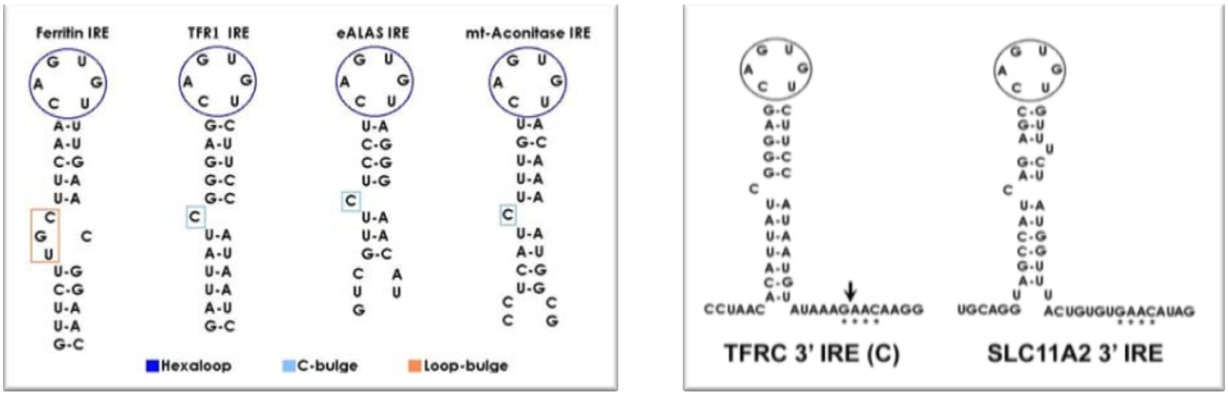
IRPs-1 and -2 interact with IREs to coordinate the expression of proteins involved in iron uptake, export, and storage, as well as in erythroid proliferation and hemoglobinization, the tricarboxylic acid (TCA) cycle, and cell cycle control. IRP binding to single IREs located in 5'-untranslated regions (UTRs) inhibits translation, whereas IRP binding to the multiple 3'-UTR IREs of the transferrin receptor 1 (TfR1) mRNA increases its stability. Cellular iron loading switches IRP1 from its IRE-binding form to a Fe-S cluster containing cytoplasmic aconitase and triggers proteasomal degradation of IRP2. Low iron levels promote accumulation of active IRP1 in its apo form and stabilize IRP2. eALAS, erythroid 5-aminolevulinate synthase; CDC14a, human cell division cycle 14A; DMT1, divalent metal transporter 1; HIF2 alpha, hypoxia inducible-factor 2 alpha (Muckenthaler, Galy et al. 2008)

The mechanisms by which the formation of an IRE/IRP complex on a single 3'UTR IRE affects the fate of the DMT1 (divalent metal transporter 1, involved in iron uptake) or the CDC14A (phosphatase important for cell cycle progression) mRNA remain unclear.

The IRPs (1 and 2) play a central role in cellular iron metabolism by co-ordinately regulating the post-transcriptional expression of genes containing iron-responsive elements (IREs) in their 5'- or 3'-untranslated regions (UTRs). IRE-containing mRNAs encode proteins of iron acquisition (transferrin receptor 1 (TFRC) and divalent metal transporter 1 (SLC11A2-DMT1-DCT1-NRAMP2)), storage (FTH1, or ferritin heavy polypeptide 1; FTL, or ferritin light polypeptide), utilization (erythroid 5'-aminolevulinic acid synthase (ALAS2 or eALAS), mitochondrial aconitase, Drosophila succinate dehydrogenase (SDH)), and export (SLC40A1-FPN1-IREG1-MTP1). Independently, both IRPs inhibit translation initiation when bound to 5'-UTR IREs (e.g. FTH1 and FTL mRNAs), whereas their association with the 3'-UTR IREs of the TFRC mRNA decreases its turnover (Schneider and Leibold 2000; Eisenstein and Ross 2003); (Muckenthaler, Gray et al. 1998). Although IRP1-deficient mice present no steady-state phenotypic abnormalities (Meyron-Holtz, Ghosh et al. 2004; Galy, Ferring et al. 2005), Irp2$^{-/-}$ animals display microcytosis associated with abnormal body iron distribution (Galy, Ferring et al. 2005) and have been reported to suffer from an overt, late onset, neurodegenerative disease (Rouault 2001). Early embryonic lethality in mice lacking both proteins indicates that the IRP/IRE regulatory network is essential (Smith, Ghosh et al. 2006). In humans, failure to coordinate the expression of IRE-containing genes is associated with pathological conditions, as illustrated by the autosomal dominant hyperferritinemia-cataract syndrome observed in patients carrying mutations in the FTL IRE (Beaumont, Leneuve et al. 1995) or by an autosomally dominant iron overload syndrome associated with a mutation in the FTH1 (Kato, Fujikawa et al. 2001).

Iron regulatory proteins (IRPs)-1 and -2 interact with iron responsive elements (IREs) to coordinate the expression of proteins involved in iron uptake, export, and storage, as well as in the TCA cycle and cell cycle control. IRP binding to single IREs located in 5'-untranslated regions (UTRs) inhibits translation, while their binding to multiple 3'-UTR IREs of the TfR1 mRNA increases its stability. Cellular iron loading switches IRP1 from its IRE-binding form to a Fe-S cluster containing cytoplasmic aconitase, and triggers proteasomal degradation of IRP2. Low iron levels and NO promote accumulation of active IRP1 in its apo-form and stabilize IRP2.

Figure 2. Representation of some known IRE structures



5' IREs **3' IREs**

Conserved secondary structure, among many members of the IRE family, is a hairpin with a terminal hexaloop of the sequence CAGUGU/C (blue circle) (Theil, McKenzie et al. 1994). Another feature shared by IREs is a C-bulge in the lower 5' side of a stem-loop. (light blue box). IREs located in the 5'- or 3'- untranslated regions (UTRs) of genes involved in iron metabolism. The binding of IRPs to single IREs in the 5'- UTRs of mRNAs blocks their translation, while IRP binding to multiple IREs in the 3'- UTR stabilizes the TfR-1 mRNA*. Several known 5' IREs and 3' IREs are represented on the left and right panel respectively.*

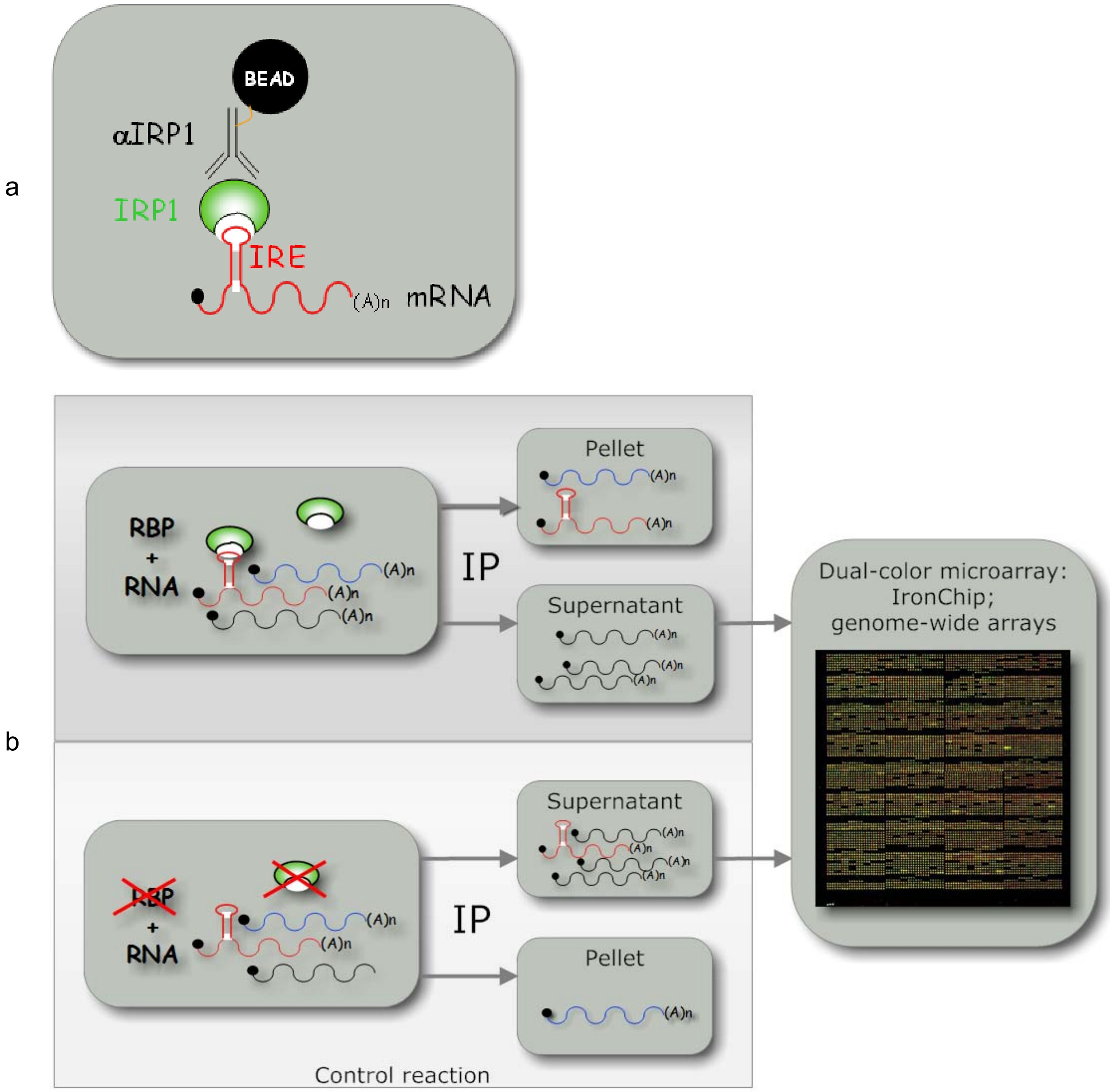
Identification of mRNAs containing IRE-like motifs

A relatively restricted number of IRE-containing mRNAs has been discovered so far (Figure 1). In order to identify all possible IRE-containing genes present in the human or in the mouse genome, purified recombinant IRP1 has been used as a “bait” to “fish” IRP-interacting mRNAs from total RNA extracted from various mouse tissues or human cell lines (Figure 3). The IRP1 ribonucleoparticles (RNPs) have been isolated using a specific anti-IRP1 antibody linked to agarose beads. In a control reaction, recombinant IRP1 was omitted to assess the background levels of the system.

Subsequently, the mRNA composition of the two mRNPs (± the IRP1 “bait”) was compared. For that purpose, RNA was extracted from the two RNPs and the resulting samples were co-hybridized onto cDNA microarrays. The mRNAs that were selectively enriched in the IRP1-mRNP may contain putative IRE motifs.

This biochemical approach was combined with a computational search of IRE-like motifs in nucleic-acid databases. This combined strategy led to the identification of a novel IRE-like motif in the 3' UTR of the CDC14A mRNA. *In vitro* experiments (competitive electromobility-shift assay) confirmed the specific interaction between the CDC14A IRE and the IRPs.

Figure 3. Identification of novel IRE-containing mRNAs



In this experiment, mRNPs are assembled by incubating total RNA with the recombinant RNA-binding protein (RBP) IRP1and the mRNPs immunoselected using a specific antibody raised against the RBP. Dual-colour microarrays were used to identify those mRNAs that specifically interact with RBP the mRNA composition of supernatant and/or immuneprecipitated fraction (Sanchez, Galy et al.)

a. representation of the IRP-mRNP bound to the beads/anti-IRP1 antibody complex.

b. comparison of RNP mRNA composition using dual-color microarrays.

RBP: RNA binding protein, here IRP1; SN: supernatant; SNc: control SN; IP: immunoprecipitate; IPc: control IP.

Microarray analysis

Principle

The principle technique underlying gene arrays is the reversal of the Southern blotting technique. Southern blotting was named after Edward M. Southern who developed this procedure at Edinburgh University in the 1970s (Southern and Mitchell 1971; Southern 1974;

Southern 1975; Staudt and Brown 2000). Southern blotting is designed to locate a particular sequence of DNA within a complex mixture. DNA molecules are transferred from an agarose gel onto a membrane and hybridised with the labelled probe. For example, Southern blotting can be used to locate a particular gene within a mixture of entire genomic DNA.

In contrast, for microarray analysis gene-specific probes are immobilized on a membrane and then hybridised to the labelled target population of cDNAs (Kurian, Watson et al. 1999). However, unlike dot blots, which are typically prepared on membranes and rarely contain more than 700 individual gene spots, gene arrays can be prepared on glass or silicon substrates called chips (Chetverin and Kramer 1994; Southern 1995; Lipshutz, Fodor et al. 1999) or positively charged nylon membranes (Gress, Hoheisel et al. 1992; Schena, Shalon et al. 1995; DeRisi, Penland et al. 1996). Hence, a markedly higher number and density of spots can be compared. This approach combines high sensitivity with high throughput because of the possibility of an enormous number of parallel experiments carried out simultaneously on a single high-density DNA array (Poustka, Pohl et al. 1986; Schena, Shalon et al. 1996; Brown and Botstein 1999).

Experimental system

Most microarray expression studies performed to date have used relatively controlled systems that can be manipulated in vitro, such as single-cell organisms (e.g., yeast) and clonal cell lines (Lashkari, DeRisi et al. 1997; Chu, DeRisi et al. 1998; Ross, Scherf et al. 2000). This technology has also been applied to study in vivo mammalian tissues and organs. Many studies have been performed using the mouse as a model organism, in part because of the relative ease of genetic manipulation coupled with the extensive genomic, anatomical, and physiological similarity with humans. Microarrays have been used to analyse gene expression in different mouse tissues, such as liver, kidney, brain, bone marrow, spleen, pancreas, placenta, skeletal muscle, and heart.

Variability of the data

Differential analysis of microarray data is difficult because of the high variability of the data. The variability results from a large number of factors operating at different times and levels during the course of an experiment. Such factors, represented In the following table include fluctuations in sample, target and array preparation, in the hybridisation process, background and overshining effects and effects resulting from the image processing (Table 1) (Schuchhardt, Beule et al. 2000; Herzel, Beule et al. 2001).

All these factors are often interrelated in complex ways but, for the purpose of comprehensibility, they can be split into two major categories: biological and experimental variability.

Table 1. Factors that can induce variability in microarray data

Source of the variability	Description
mRNA preparation	Depending on tissue and sensitivity of RNA degradation sample may vary from sample to sample
Transcription	reverse transcription to cDNA will result in DNA species of varying lengths
Labelling	radioactive labelling may fluctuate randomly and systematically depending on nucleotide composition
Amplification	PCR amplification is difficult to quantify and may fail completely
Systematic variations in pin geometry	pins have different characteristics and surface properties and therefore carry different amounts of target cDNA
Random fluctuations in target volume	the amount of transported target fluctuates stochastically even for the same pin
Target fixation	the fraction of target cDNA that is chemically linked to the slide surface or to the nylon membrane from the droplet is unknown
Hybridization parameters	efficiency of the hybridization reaction is influenced by a number of experimental parameters, notably temperature, time, buffering conditions and the overall amount of sample molecules used for hybridization
Slide inhomogeneities	the sample may be distributed unequally over the slide or the hybridization reaction may perform differently in different parts of the slide
Non-specific hybridization	a typical source of error that cannot be completely excluded
Non-specific background and overshining	non-specific radiation and signals from neighbouring spots
Image analysis	non-linear transmission characteristics and saturation effects and variations in spot shape

The table contains most common factors that can cause data variability during microarray experiments. The list of factors is not complete, since different experimental conditions can introduce their sources of variability.

Microarray data evaluation

DNA microarrays are a popular high throughput technique to perform genome-wide molecular and genetic experiments. This requires computational post-processing of the resulting data. The computational analysis of microarrays is challenging due to the high data variability. Variability results from a large number of factors operating at different steps and levels during the course of an experiment (see Table 1) (Eickhoff, Schuchhardt et al. 2000). Series of computational steps for processing the data have to be applied: image processing, extraction of raw data, storage and normalization of the raw data, feature extraction, final data analysis and biological interpretation of the results. Several packages are available to perform the described tasks, but it is often necessary to develop custom-made solutions to fulfil individual requirements and good statistical evaluation of gene array data (Dudoit, Gentleman et al. 2003; Gentleman, Carey et al. 2004). For this purpose many customized microarrays platforms are available.

Whole-genome arrays, such as Affymetrix GeneChips are very reproducible, but still cannot provide reliable data for all genes, especially for low-expressed genes. Moreover, there are technical limitations in array design. For example, standard glass slide arrays cannot accommodate more than 60,000 spots, including all controls and replicates. This is often not sufficient for complex genomes such as the human genome. The relationship between fluorescent signal intensity and gene expression level is linear only for a certain range of concentrations of spotted material.

Thus, the differences in linearity range can become more pronounced for larger whole-genome arrays. In contrast, customized microarrays may contain fewer genes. This allows including more replicates and controls. Smaller number of genes, in comparison with whole-genome arrays, and higher number of repetitions will increase reliability of the data and allow detecting low-expressed genes.

IronChip

IronChip is a cDNA microarray platform specifically designed to investigate regulatory networks in iron metabolism. We have developed two versions of this platform to analyse both human and mouse genes (Muckenthaler, Roy et al. 2003). The design of this microarray enables detection of small, but physiologically significant changes in gene expression due to the high number of repetitive features. The current version of the IronChip contains 520 genes involved in iron homeostasis and related pathways. To improve array sensitivity and data robustness, each gene on the array is represented by several ESTs. Each EST, in turn, is represented by a minimum of six spots. Some of the most relevant iron-related genes are represented by up to 24 spots. This microarray further contains a collection of negative controls, specificity controls and positive (spike-in) controls (Muckenthaler, Richter et al. 2003).

Expression analysis of IRE/IRP regulatory network

To explore the IRE/IRP regulatory network, we immunopurified IRP1/mRNA complexes formed with RNA isolated from various cell lines and tissues and identified novel IRP binding mRNAs using the IronChip cDNA microarray platform (Muckenthaler, Richter et al. 2003) herein this study, the IronChip was complemented with genes bearing IRE-like motifs that we identified by bioinformatic screens of nucleic acid databases. We demonstrate that this integrated experimental strategy reliably identifies known IRE-containing genes. We also identify a conserved IRE in the 3'-UTR of mRNA encoding the CDC14A tumor suppressor gene, pointing to a previously unrecognized regulatory link between iron metabolism and the cell cycle.

Microarray data evaluation

Sophisticated bioinformatical tools are required to draw valid conclusions from the massive amount of information obtainable from the complex DNA chip data. Improved access to large electronic data sets, reliable and consistent annotation and effective tools for data mining are critical to achieve robust comparative data evaluations (Dhiman, Bonilla et al. 2002). In addition, custom microarrays platforms, such as IronChip, may have a specific design properties, which one may want to use during the analysis. These custom microarray platforms provide more data than required or exploited in standard statistical analysis. To incorporate all the advantages such a chip design offers for data analysis we developed the IronChip Evaluation Package (ICEP). ICEP makes use of the high number of repetitions to improve data quality. The comparison of different ESTs enables reliable detection of transcript-specific regulation (e.g. alternative splicing variants of the same gene). Analyses of the positive and negative controls allow precise calculation of a reliable ratio cut-off as well as to estimate background noise, respectively (Vainshtein, Sanchez et al. 2010).

ICEP supports all available version of the IronChip (version 1.0 to version 8.0 of a murine and version 1.0 to 5.0 of a human IronChip).

Personal contributions

The major goal of this study is to develop a novel method of customized microarrays analysis, to confirm feasibility of the method on the IronChip microarrays platform and to apply analysis to study gene expression pattern under different conditions and to identify novel IREs.

We developed a tool called “IronChip Evaluation Package” (Vainshtein, Sanchez et al. 2010). Y.Vainshtein, M.U.Muckenthaler, M.Sanchez and A.Brazma contributed to the conceptualization of the method; Y. Vainshtein developed and implemented the methods, wrote the code and the manuscript. , M.U.Muckenthaler, M.W.Hentze, T.Dandekar and A.Brazma provided substantial intellectual contribution to the manuscript; M.U.Muckenthaler and M.Sanchez provided the application cases.

ICEP tool, described in this thesis, was used for IronChip analysis in many projects. One of the interesting examples is the identification of novel IREs by combination of biochemical and biocomputational methods (Sanchez, Galy et al. 2006).

M. Sanchez, B. Galy, MU Muckenthaler contributed to the conceptualization of the method; M.Sanchez developed experimental procedure, performed experiments and wrote the manuscript; Y. Vainshtein, T. Dandekar, P. Bengert provided bioinformatical support of the project; M.U.Muckenthaler, M.W.Hentze and T.Dandekar provided substantial intellectual contribution to the manuscript.

All biological examples and application cases in the thesis are provided by M.Sanchez and M.U. Muckenthaler. Development and implementation of the method, writing the code and analysis of IronChip data was done by Y.Vainshtein.

Materials and Methods

System requirements

ICEP runs on any computer with minimum of 256 Mb of RAM and any Windows operating system. Computers operating more RAM will show better performance running ICEP. The recommended RAM size is 1024 Mb

IronChip data sets

To validate the software and our methodology we applied the ICEP software to analyse microarray data sets obtained from IronChip with a previously tested microarray layout (v.4hs) from hemin-or desferrioxamine-treated HeLa cells. This dataset is composed of previously reported microarray experiments carried out in our laboratory (Muckenthaler, Richter et al. 2003; Muckenthaler, Roy et al. 2003). Human HeLa cells are a very well characterized model system to study iron metabolism. We therefore chose treated HeLa cells to examine ICEP analysis reproducibility. Induced iron overload or iron deficiency caused expected (Muckenthaler, Richter et al. 2003; Muckenthaler, Roy et al. 2003) changes in expression levels of different genes.

IronChip is a cDNA microarray spotted on a coated glass slide, which allows simultaneous hybridization of a two fluorescently labelled samples. In this study we cannot describe all experiments performed with IronChip and analysed by ICEP. We are going to focus only on a small subset of whole dataset (Table 2).

Table 2. Experimental settings for the example experiment

Sample 1				**Sample 2**			
Cell line	**Treatment**	**Effect**	**Labelling**	**Cell line**	**Treatment**	**Effect**	**Labelling**
HeLa	Hemin	Iron-overload	Cy3	HeLa	Desferrioxamine	Iron-loss	Cy5
HeLa	Hemin	Iron-overload	Cy5	HeLa	Desferrioxamine	Iron-loss	Cy3

IronChip was used for many studies, involving human/mouse tissues or cell lines. Table 3 contains a complete list of experiments performed with IronChip and evaluated later with the ICEP:

Table 3 Experiments done with the IronChip and analysed using ICEP

Treatment	Genetic Background	Mice	infections	Tissues
Non-treated	c57bl/6j	HFE-/-	Neisseria	Liver
Desferrioxamine	c57bl/10j	C282Y	Lesteria	Spleen
Hemin	SV129	Tissue-specific HFE-/-		Duodenum
Dietary iron deficiency	SWR/j	B2M-/-		Kidney
Iron Dextran	RF/j	HFE-/- B2M-/-		
Copper deficiency	CD1	H-2Kb-/- H-2Db-/-		cell lines
Dietary copper deficiency		H-2Kb-/-		
LPS		Tfr+/-		
Phenylhydrazine		Hpx		
Drug-treated		DMT1-/-		
		H-Fer+/-		
		CP-/-		
		IRP1-/-		
		IRP2-/-		
		Rag1-/-		
		Abc7-/-		
		HNF4a -/-		
		MyD88-/-		

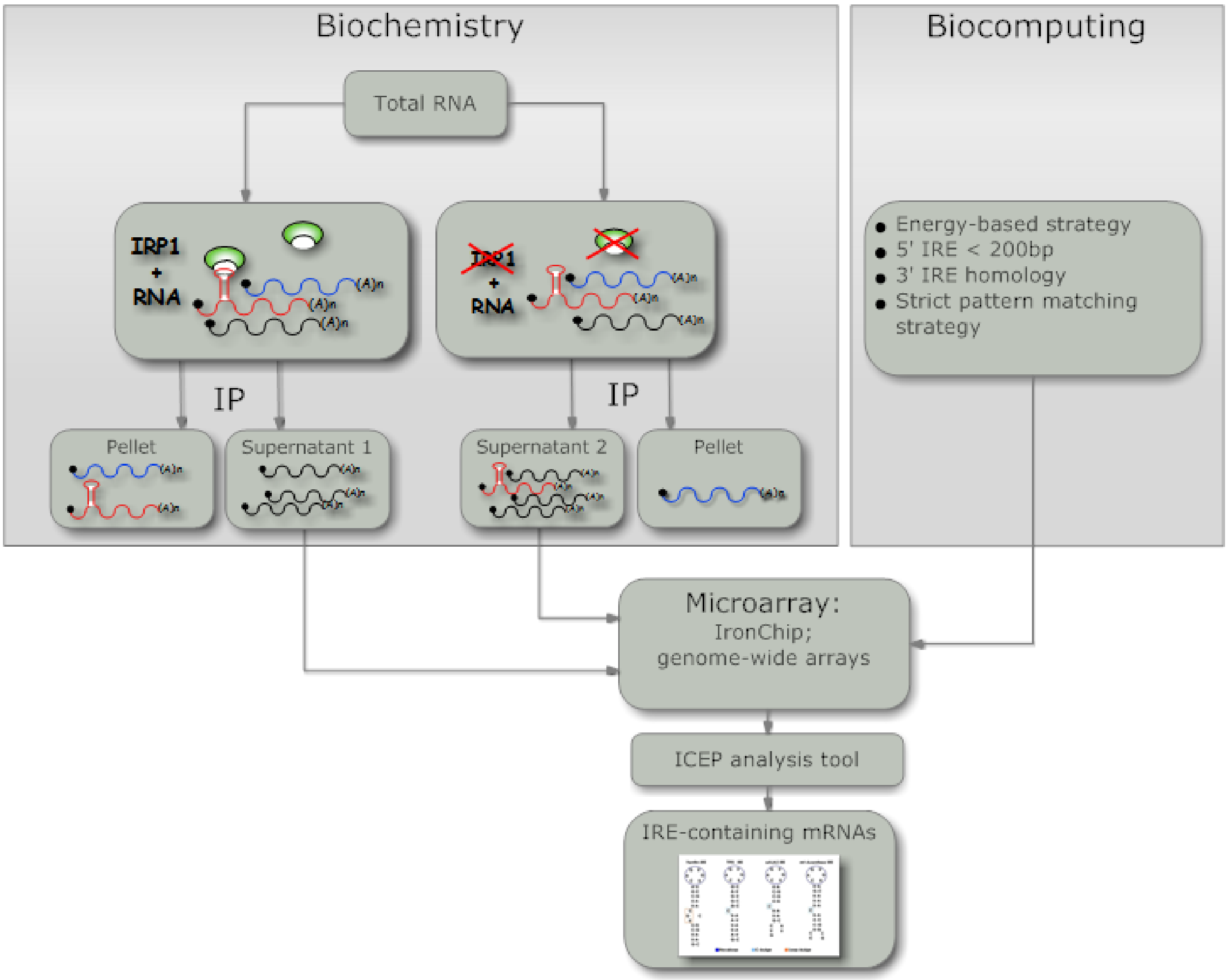
The table contains most mouse knock-out lines and treatment factors, used in combination for a microarray experiment with the IronChip.

Most of the experiments described in the table 3 were carried out using standard control samples and standard conditions and protocols during sample preparation and hybridisation. Usage of an automated analysis tool such as ICEP allows to compare results of different microarray experiments and to build an iron regulatory network.

Strategy to Identify Novel IRE-containing mRNAs

To identify novel IRE-containing mRNAs we combined biochemical and biocomputational approaches coupled with microarray analysis (Figure 4):

Figure 4. Experimental strategy



Biochemical and biocomputational approaches converge in microarray analyses for the identification of novel IRE-containing genes. IRP/IRE mRNPs were isolated using anti-IRP1 antibodies and recombinant IRP1. Immunodepletion of IRE-containing mRNAs from the supernatant was analysed using microarrays. Data evaluation performed using ICEP microarrays analysis package. Nucleic acid databases were screened for IRE-like motifs using several biocomputational approaches and restriction filters. The identified genes were spotted onto the IronChip microarray platform. The specificity and efficiency of IRP/IRE mRNP isolation was tested by Northern blot analysis of supernatant (SN) and immunopurified (IP) fractions (Sanchez, Galy et al. 2006)

In this approach, IRP/IRE mRNPs were immunoselected and their mRNA composition was analysed using microarrays.

The experimental strategy was validated using the IronChip (Gray, Pantopoulos et al. 1996), a specialized cDNA microarray platform established to study iron metabolism-related genes. The IronChip contains all known IRE genes and was complemented here specifically with cDNAs predicted to bear IRE-like sequences (Table 4).

Table 4 Summary of IRE-like motifs identified biocomputationally

Gene name	Ref. sequence	Species	IRE at	IRE sequence	IRE in other species	E IRE[a]
ADAR	NM_001111	H.sapiens	5'	GCCCCGGGGC.C.ACUUC.CAGUGC.GGAGUAGCGGAGGCGUG	D.melanogaster	-12.9
PRKACA	NM_002730	H.sapiens	5'	GCCUUCCCAG.C.CACCG.UAGUGC.CGGUGCCCUGAGAACAG	M.musculus	-10.4
Rnf5	NM_019403	M.musculus	5'	UGUGUGUGCC.C.UGUGU.UAGUGU.AUAUGUGUGUGUGCCUG	H.sapiens	-3.8
S6K[b]	NM_079217	D.melanogaster	5'	GUGCGUG.C.CGUCG.CAGUGU.UGGUGCGUGUGC		-9.0
Sh3gl3	NM_017400	M.musculus	5'	GCGCGCGCGC.C.UGUGC.CAGUGU.GACAGCGCCGUGGCCGU	H.sapiens	-13.1
BRF1[b]	NM_001519	H.sapiens	3'	GCAGGGG.C.CGGUG.CAGAGC.CACUG.UCUGUGU		-9.9
CAV3	NM_001234	H.sapiens	3'	CUUGGGCUGG.C.AGGGG.CAGUGA.CCCUUCCAGGGU	M.musculus	-13.1
CDC14A[b]	NM_003672	H.sapiens	3'	AUAUUUA.C.AUGUA.CAGUGU.UACAUUAUAUAU	M.mulatta,B.taurus, and R.norvegicus	-1.4
CDC42BPA[b]	NM_003607	H.sapiens	3'	UAGAAAA.C.ACUUG.CAGAGC.CAGGU.UUUGCUG	np[c]	
Cnbp1	NM_013493	M.musculus	3'	GAGGCUGUUC.C.CAGGC.CAGUGA.GCUUUACUUGCAGUGUA	H.sapiens and R.norvegicus	-8.3
D11Ertd498e	NM_145940	M.musculus	3'	GAGUUUGCGA.C.GGGAC.CAGUGU.GUCUAGACGACGAGAAU	H.sapiens	-6.3
DKFZP564B147	AL117556	H.sapiens	3'	GGACACAGCC.C.CUGGA.CAGUGA.UCCAGACAGCUGGCCGU	M.musculus	-14.4
Dsipi	NM_031345	R.norvegicus	3'	CCUAGUAACC.C.CAAGC.CAGUGA.GCUUGUCGUGCCACCGG	H.sapiens and M.musculus	-7.0

Gene name	Ref. sequence	Species	IRE at	IRE sequence	IRE in other species	E IRE[a]
FLJ34594[b]	XM_379386	H.sapiens	3'	GAGCUCC.C.UGACC.CAGAGA.GGUUA.AGGGUUU		-9.0
FLJ44675[b]	AK126633	H.sapiens	3'	GAAGAUU.C.UUUGG.CAGUGU.CCAAGAAUUAUC		-5.9
PSMA4	X91847	S.scrofa	3'	AUUUGGGGCA.C.CAGUU.CAGUGU.AAAAGCUGUCCUACUCU	H.sapiens	-6.3
SERTAD2	NM_014755	H.sapiens	3'	UAGUUUUUGC.C.UUUUU.CAGAGA.AAAAGAAAUUGCUUUGA	M.musculus	-3.8
SMARCC2	NM_003075	H.sapiens	3'	CCCUGUGC.C.ACCUC.CACAGU.GAGGAGCCAGCCAGACAUC	M.musculus	-9.9
TRAM1	NM_014294	H.sapiens	3'	CUGUUUGUGC.C.AUUUU.UAGUGU.AAAAGUUGCAGACCUAU	C.familiaris	-4.7
Vdac3	NM_011696	M.musculus	3'	AUAUCAGUCU.C.UGCUC.UAGUGA.GAGCUUUGGUUUUGCAU	H.sapiens	-6.3
ZC3H11A/ KIAA0663	NM_014827	H.sapiens	3'	UAGAGGAAUU.C.UUUUU.UAGUAU.GAAAAUUGUCCCUUUUC	M.musculus	-5.2

a *E, predicted IRE structure energy (14).*

b *IRE-like motifs selected from the UTRdb (23).*

c *np, not possible to determine*

Biocomputational Identification of Novel IRE-like Motifs

IRE-like motifs were identified from genomic nucleic acid databases by an algorithm combining primary nucleic acid sequence and RNA structural criteria (Bengert and Dandekar 2003; Mignone, Grillo et al. 2005). Depending on the choice of constraining criteria, such computational screens tend to generate a large number of false positives. To refine the search and reduce the number of false positive hits, additional constraints were introduced. First, we restricted the positive hits to those whose IRE folding energy was consistent with the energy of known IREs (at least ~3 kcal/mol or below). In addition, putative 5'-IRE motifs had to be located within the first 200 nucleotides of the mRNA leader, because the distance between functional 5'-IREs and the cap structure has been shown to be critical for efficient translational repression in mammals (Goossen and Hentze 1992). Such criteria identify all known 5'-IREs, including the experimentally confirmed IRE in SLC40A1 (Lymboussaki, Pignatti et al. 2003). For 3'-IRE motifs, only those conserved in at least two different species were selected. This refined screen yielded 15 IRE motifs (4 within the 5'-UTR and 11 within the 3'-UTR). A second approach made use of a reported list of 230 IRE-like sequences obtained from screening UTR databases (Mignone, Grillo et al. 2005). We selected 6 out of these 230 entries based on the ability of the lower IRE stem to form at least 6 out of 7 bp. Thus, 21 mRNAs with IRE-like motifs were selected in total (Table 4). Corresponding ESTs were spotted onto the human or mouse versions of the IronChip, a sensitive cDNA microarray platform covering ~500 genes involved in iron metabolism and connected metabolic pathways (e.g. copper and oxygen metabolism) (Muckenthaler, Richter et al. 2003)

IronChip Analysis of IRP/IRE mRNPs

To determine whether mRNAs with biocomputationally-predicted IRE-like motifs are contained within the immunopurified IRP/IRE mRNPs, we extracted their mRNAs for analysis on IronChips that include all known IRE-containing genes as well as those genes with IRE-like motifs identified by the biocomputational strategy (see above). Because the complexity of the mRNA population in the immunopurified (IP) fraction is not sufficient for global normalization protocols applied to dual colour microarray data analyses (Benes and Muckenthaler 2003), we used the supernatant (SN) fractions that offer high mRNA complexity and assayed these for the depletion of IRE-containing mRNAs. To exclude possible artefacts associated with uneven incorporation of cyanine dyes into cDNAs, we routinely performed dye switch experiments (Benes and Muckenthaler 2003).

Results

ICEP tool, described in this thesis, was used for IronChip analysis in many projects. One of the interesting examples is the identification of novel IREs by combination of biochemical and biocomputational methods (Sanchez, Galy et al. 2006).

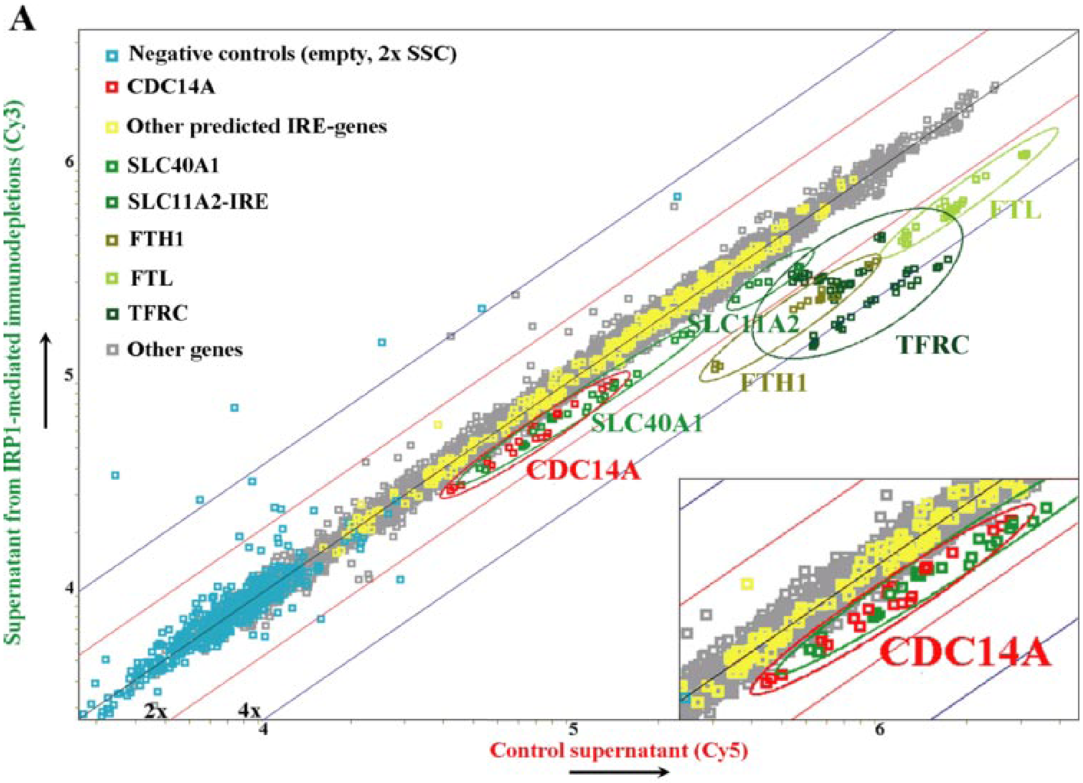
The work distribution was as follows: M. Sanchez, B. Galy, MU Muckenthaler contributed to the conceptualization of the method; M.Sanchez developed experimental procedure, performed experiments and wrote the manuscript; Y. Vainshtein, T. Dandekar, P. Bengert provided bioinformatical support of the project; M.U.Muckenthaler, M.W.Hentze and T.Dandekar provided substantial intellectual contribution to the manuscript.

Identification of novel IRE (Cdc14a)

From the 21 candidates that were selected biocomputationally, the majority was either undetectable or not significantly depleted from the SN fraction. The CDC14A mRNA was efficiently depleted from the supernatant of two human cell lines (293 and TPH1) and enriched in the immunopurified IRP1/IRE mRNPs from the human cell line 786-O (see Figure 5).

The depletion of IRE-containing mRNAs from the SN fractions was analysed using a dual colour cDNA microarray (IronChip) to generate a virtual scatterplot from Caco-2 and 293 human cell line data (Figure 5A). The reduction of known IRE-containing mRNAs (positive controls) from specific IRP1-mediated immunodepletions was clearly observed in the case of the TFRC mRNA and the FTH1 and FTL mRNAs in all tissues and cell lines tested (Figure 5, A and B). The tissue-specific SLC40A1 or IREG1 mRNA was detected in RNA samples of mouse duodenum and human Caco-2 cells and efficiently depleted from these input RNAs by an immunoselection procedure. Depletion of the SLC11A2-IRE (DMT1) mRNA was detectable in RNA from mouse duodenum, lung, and brain as well as human Caco-2 and MCF7 cell lines. The other tested tissues and cell lines did not detectably express SLC11A2 mRNA. The IronChip analysis therefore nicely mirrored the data generated by Northern blotting and qRT-PCR (Figure 6) and thus is a suitable tool to identify bona fide IRE-containing mRNAs in the supernatant fractions of IP reactions.

Figure 5. Microarray analysis detects known IRE-regulated mRNAs and identifies the novel IRE-containing gene CDC14A



B

Mouse tissues	Gene *TFRC*	*FTL*	*FTH1*	*SLC40A1*	*SLC11A2*	*CDC14A*
Liver	+++	+++	+++	n.a.		
Spleen	++	++	+++			
Brain	++	+	++	n.a.	(+)	
Duodenum	++	+	++	+	(+)	
Heart	+++	+++	++++	n.a.		
Lung	++	++	++	n.a.	(+)	
Kidney	+++	+++	+++	n.a.		

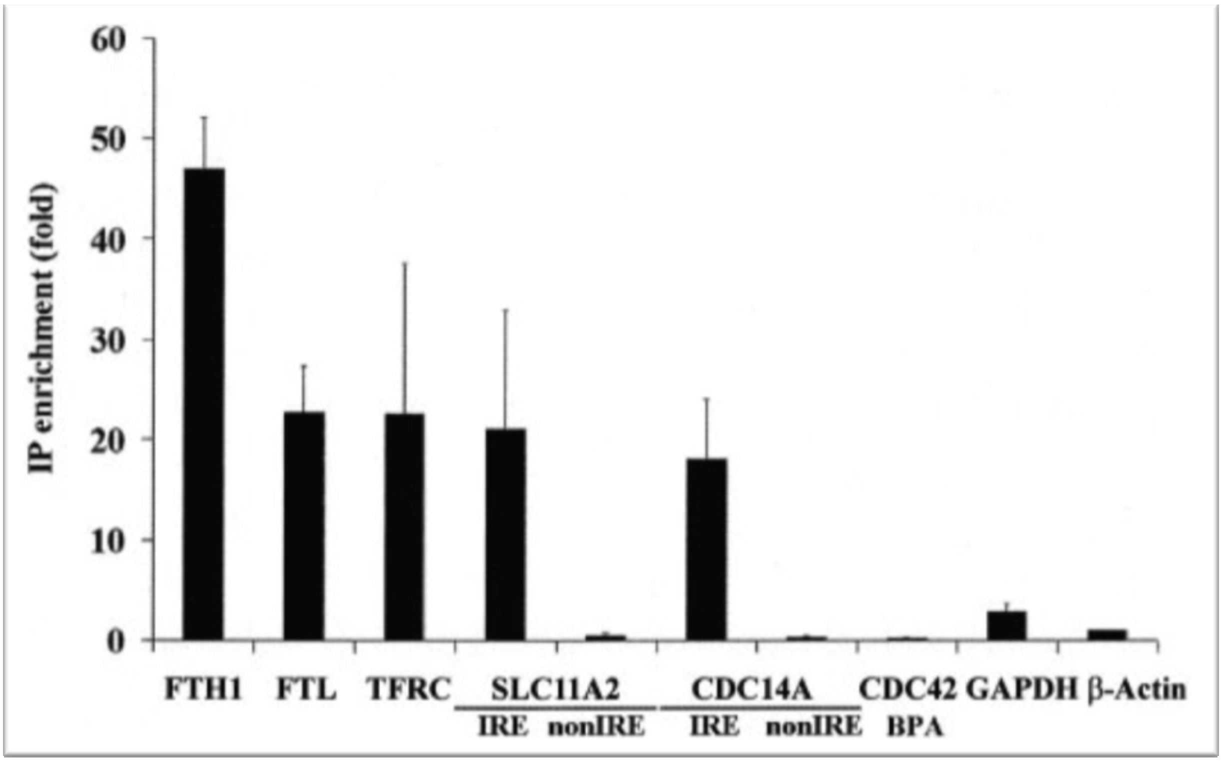
Human cell lines	Gene *TFRC*	*FTL*	*FTH1*	*SLC40A1*	*SLC11A2*	*CDC14A*
HepG2 (liver)	+	+	+	(+)		
Caco-2 (intestine, colon)	+++	++	++	+	(+)	
HeLa (uterus cervix)	+++	++	+			
MCF7 (breast)	++	+	+		(+)	
K562 (blood)	+	+	+			
TPH1 (bone marrow)	++	+	+			(+)
293 (kidney)	++	+	(+)			(+)

A. Virtual scatterplot generated from Caco-2 and 293 human cell line data. Known IRE-regulated genes are represented by green dots and predicted IRE-genes by red or yellow dots. Known IRE-regulated genes and CDC14A are circled. The 2-fold and 4-fold depletion limits are indicated by red and blue lines, respectively.

B. Summary of the data obtained from all mouse tissues and human cell lines analysed in this study. The -fold depletions are indicated as follows: +,>1.4;++,>2.0;+++,>3.0; and ++++,>4.0. (+) indicates that not all ESTs represented on the IronChip microarray platform performed as expected in dye switch experiments. n.a., the corresponding gene was not available on the microarray version used. (Sanchez, Galy et al. 2006)

SLC11A2 isoforms (IRE versus non-IRE) cannot be distinguished by Northern blotting therefore we also analysed the IP fractions from three independent experiments by qRT-PCR. In agreement with the Northern blot data the FTH1 and FTL mRNAs, as well as the TFRC mRNA, were strongly enriched in the IP fraction (46.9 ± 5-fold, 22.7 ± 5-fold, and 22.6 ± 15-fold, respectively) (Figure 6). In addition, the SLC11A2-IRE mRNA isoforms were selectively enriched in the IP fraction (21.0 ± 12-fold), whereas the non-IRE isoforms expectedly behaved as negative control (0.4 ± 0.4-fold) confirming the high selectivity and specificity of the procedure. Taken together, these data show that the immunoselection/microarray strategy is a feasible approach for screening bioinformatically predicted IRE genes and the detection of novel IRE-containing mRNAs.

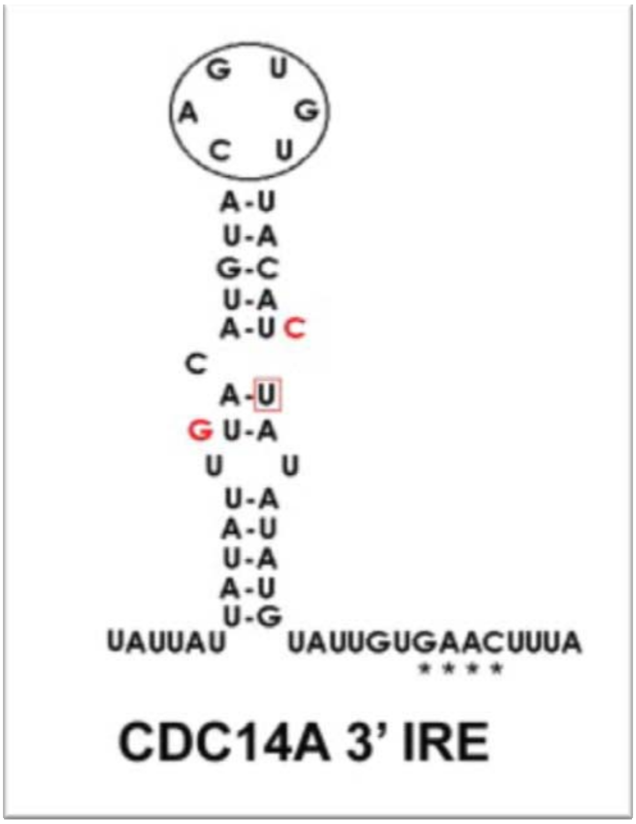
Figure 6. qRT-PCR analysis of IP samples



The histogram shows ratios (mean ± S.D.) of mRNA levels (normalized to β-actin mRNA) in IP fractions obtained in the presence versus absence of recombinant IRP1 (Sanchez, Galy et al. 2006)

Cdc14a fulfils all criteria of an IRE as tested by my Perl scrip or other IRE prediction programs, however, the energy E is too low (2.3 kcal) for correct recognition of a characteristic hairpin-like IRE structure by corresponding prediction software. The structural approach based on recognition of conserved sequence elements of IRE gives a following structure of a Cdc14a (Figure 7):

Figure 7. Structure of the CDC14A 3'-UTR IRE motif



Representation of the IRE structures present in the 3'-UTR of the human CDC14A mRNA. Nucleotides disrupting the IRE structure in the mouse sequence are indicated in red. A conserved four-nucleotide sequence is marked with asterisks (Sanchez, Galy et al. 2006)

Iron Regulation of CDC14A mRNA

This work identifies CDC14A as a novel IRE-containing and iron-regulated mRNA. CDC14A is one of the two human orthologs of the yeast CDC14 (cell division cycle 14) gene that has been shown to encode a phosphatase involved in the dephosphorylation of several critical cell cycle proteins (Wong, Chen et al. 1999). Loss of function mutations of the CDC14A gene have been described in various human cancer cell lines, suggesting that CDC14A could act as a tumor suppressor (Wong, Chen et al. 1999). Indeed, CDC14A has been shown to dephosphorylate cdk substrates such as p27kip1 and cyclin E that are critical for the G1 to S phase progression (Kaiser, Zimmerman et al. 2002). Alteration of CDC14A expression by RNA interference or transgenic overexpression has been found to cause abnormal mitotic spindle assembly and chromosome segregation (Kaiser, Zimmerman et al. 2002; Mailand, Lukas et al. 2002), arguing that CDC14A plays an important role in cell division. Present and previous studies (Wong, Chen et al. 1999; Kaiser, Zimmerman et al. 2002) revealed several mRNA isoforms with heterogeneity at both the 5’- and 3’-ends of the CDC14A mRNA. These mRNA isoforms are predicted to encode different protein products that differ by their N and C termini. This heterogeneity is highly reminiscent of SLC11A2, another 3’-IRE-containing gene that encodes several protein isoforms (Hubert and Hentze 2002) with distinct subcellular localizations (Tabuchi, Tanaka et al. 2002; Lam-Yuk-Tseung and Gros 2006). Analogously, the N- and C-terminal heterogeneity of CDC14A proteins could affect the targeting of the phosphatase within the cell. Immunofluorescence studies using antibodies that do not discriminate between the CDC14A isoforms revealed both cytoplasmic and centrosomal staining (Mailand, Lukas et al. 2002). Here, we identify a novel 5’-exon (exon 1A) predicted to contain an N-myristoylation site (GNFLSR) that may target the protein to membranes (Farazi, Waksman et al. 2001). Further work will explore this important aspect of CDC14A biology.

IronChip Evaluation Package (ICEP)

The major goal of this study is to develop a novel method of customized microarrays analysis, to confirm feasibility of the method on the IronChip microarrays platform and to apply analysis to study gene expression pattern under different conditions and to identify novel IREs.

We developed for this our tool called “IronChip Evaluation Package” (Vainshtein, Sanchez et al. 2010). The work distribution for this was as follows: Y.Vainshtein, M.U.Muckenthaler, M.Sanchez and A.Brazma contributed to the conceptualization of the method; Y. Vainshtein developed and implemented the methods, wrote the code and the manuscript. , M.U.Muckenthaler, M.W.Hentze, T.Dandekar and A.Brazma provided substantial intellectual contribution to the manuscript; M.U.Muckenthaler and M.Sanchez provided the application cases.

Implementation

ICEP exploits a collection of Perl programs and utilities with a Perl Tk GUI (graphical user interface). The Perl routines were all newly custom written for the purpose of rapid and solid microarray data analysis. ICEP features a decision-tree based algorithm to optimize spot selection and exploit here in particular multiple repetitions of ESTs. ICEP applies grouping rules in its decision tree algorithm to calculate signal intensity ratios for each individual group of ESTs representing the same transcript (this is step by step explained in the application of the analysis pipeline together with supporting online material. The pipeline is summarized in the Results section, Figure 10). ICEP does not use or require any existing software libraries and it can directly process simple tab delimited tables of array data of any type. It adds its optimized spot selection, filtering and normalization procedure to standard software such as Bioconductor (Gentleman, Carey et al. 2004) and can be used in combination with these or equally well alone.

ICEP is packaged to a Windows executable with a PDK (Perl developer kit). It can be operated in a command line mode or in a batch mode for the analysis of multiple arrays. A simple editor allows specifying all microarray data files for batch analysis. In a command line mode the ICEP analysis core itself can be executed under any operating system supporting Perl. Some Windows specific features, such as exporting of output data to an Excel table or Perl Tk GUI requires adaptation to a specific operating system. In general, the modular structure of ICEP allows porting it to any other operating systems supporting Perl.

We are using the following Perl and PDK version:

- ActiveState Perl v5.8.8 built for MSWin32-x86-multi-thread
- ActiveState Perl Development Kit v.7.3.0 for MSWin32-x86

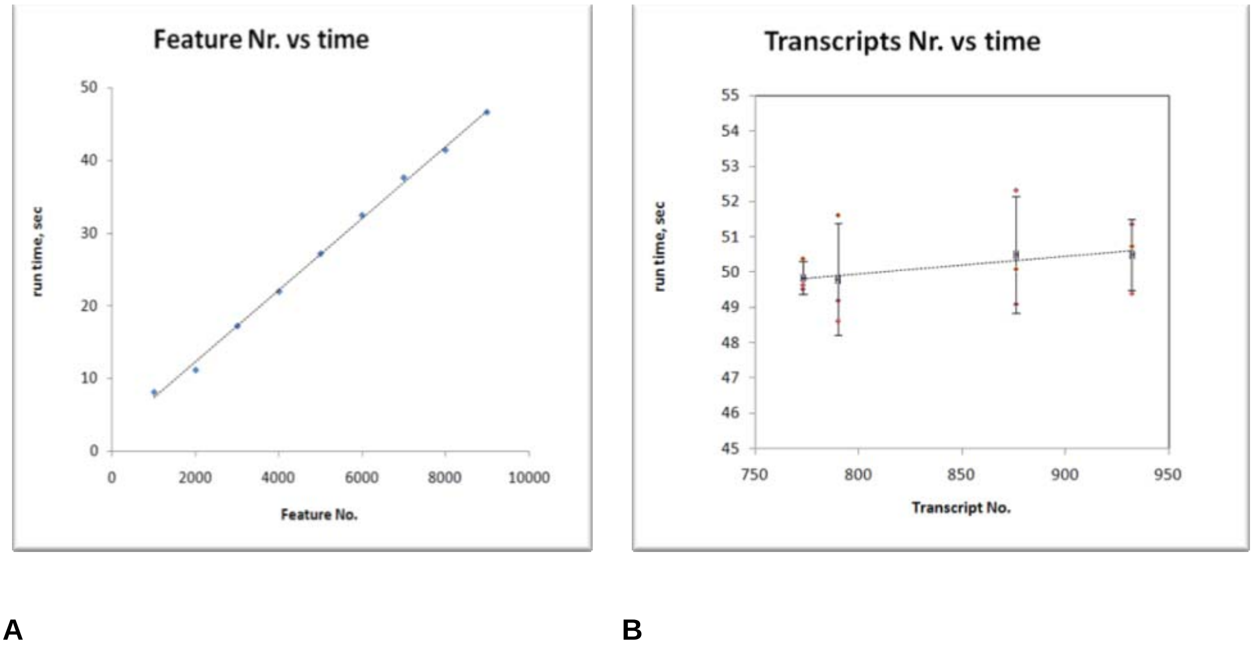
Data formats

ICEP recognizes any generic tab-delimited text tables from any type of gene microarray containing the normalized signal intensities and background data (e.g. from the ChipSkipper application; (Schwager 2002).ChipSkipper generates a tab-delimited text table containing raw and normalized signal intensity values, background signal intensity, physical coordinates of a spot relative to an upper left corner of a glass slide, cDNA sample position on original PCR spotting plate (row, column, plate Nr), some flags and statistical values related to the spot geometry and other internal values. The ICEP uses only few of those columns: spot and clone coordinates, comments and background-compensated and normalized signal intensity value from both channels. The build-in utility recognizes not only different formats of a ChipSkipper output file, but any generic tab-delimited text file gene expression array data can be processed by ICEP using the provided flexible configuration tool and alternative input file formats are added using the provided flexible configuration tool. Results are saved in tab-delimited format or Microsoft Office Excel formats.

Performance

We tested ICEP performance by measuring time consumption to analyse microarray data from different IronChip versions (version 2.0 contains 559 transcripts, while version 7.0 contains 932 transcripts) (Figure 8A) or by analysing a set of virtual arrays (1000 to 9000 features, with 1000 features step). On average, ICEP could evaluate 208 features per second. The time per run increases linearly with an increasing number of analysed features (Figure 8B).

Figure 8. ICEP run time chart

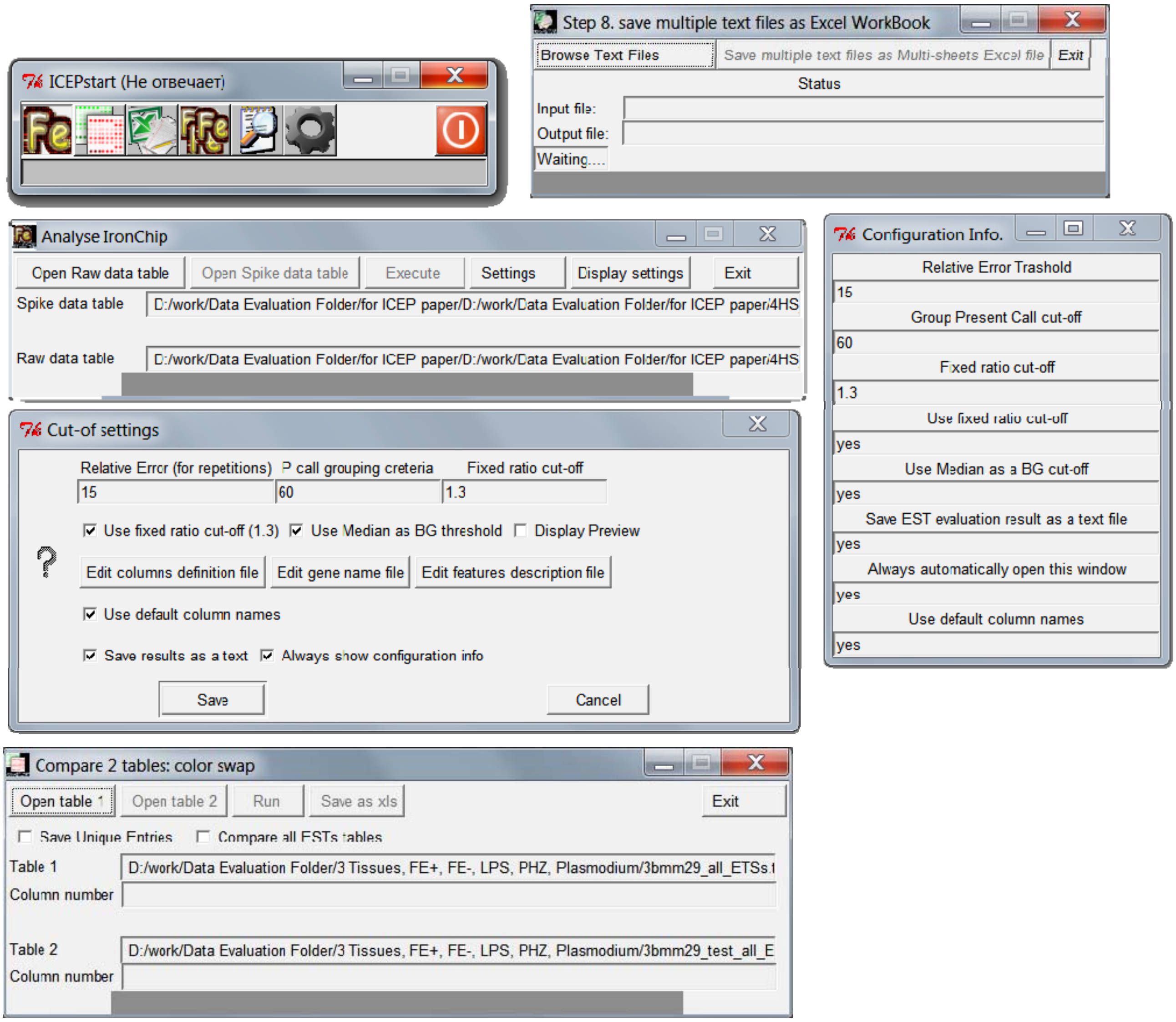


- *A. Run time analysis of different microarray versions containing an increasing number of features. The plot shows the resulting increase in ICEP run time for different IronChip versions.*
- *B. A set of virtual arrays of 1000 to 9000 features was analysed. We used a general tab delimited format. Robust statistical analysis included analysis of background noise, ratio cut-o_, evaluation of multiple repetitions, detailed feature extraction and grouping results. ICEP Run time increases linearly with the increase of the total number of analysed features. On average, ICEP evaluates 208 features per second. (Vainshtein, Sanchez et al. 2010)*

User interface

ICEP has been developed as a stand-alone application and does not require any special environment. It runs under any Windows operating system (it was tested on Windows 2000, XP and Vista). The interface has been designed to be highly user-friendly and interactive (Figure 9):

Figure 9. ICEP Graphical User Interface



Input file:

Output file:

Waiting....

Spike data table D:/work/Data Evaluation Folder/for ICEP paper/D:/work/Data Evaluation Folder/for ICEP paper/4HS

Raw data table D:/work/Data Evaluation Folder/for ICEP paper/D:/work/Data Evaluation Folder/for ICEP paper/4HS

Relative Error (for repetitions)	P call grouping creteria	Fixed ratio cut-off
15	60	1.3

☑ Use fixed ratio cut-off (1.3) ☑ Use Median as BG threshold ☐ Display Preview

☑ Use default column names

☑ Save results as a text ☑ Always show configuration info

Relative Error Trashold
15
Group Present Call cut-off
60
Fixed ratio cut-off
1.3
Use fixed ratio cut-off
yes
Use Median as a BG cut-off
yes
Save EST evaluation result as a text file
yes
Always automatically open this window
yes
Use default column names
yes

☐ Save Unique Entries ☐ Compare all ESTs tables

Table 1 D:/work/Data Evaluation Folder/3 Tissues, FE+, FE-, LPS, PHZ, Plasmodium/3bmm29_all_ETSs.t

Column number

Table 2 D:/work/Data Evaluation Folder/3 Tissues, FE+, FE-, LPS, PHZ, Plasmodium/3bmm29_test_all_E

Column number

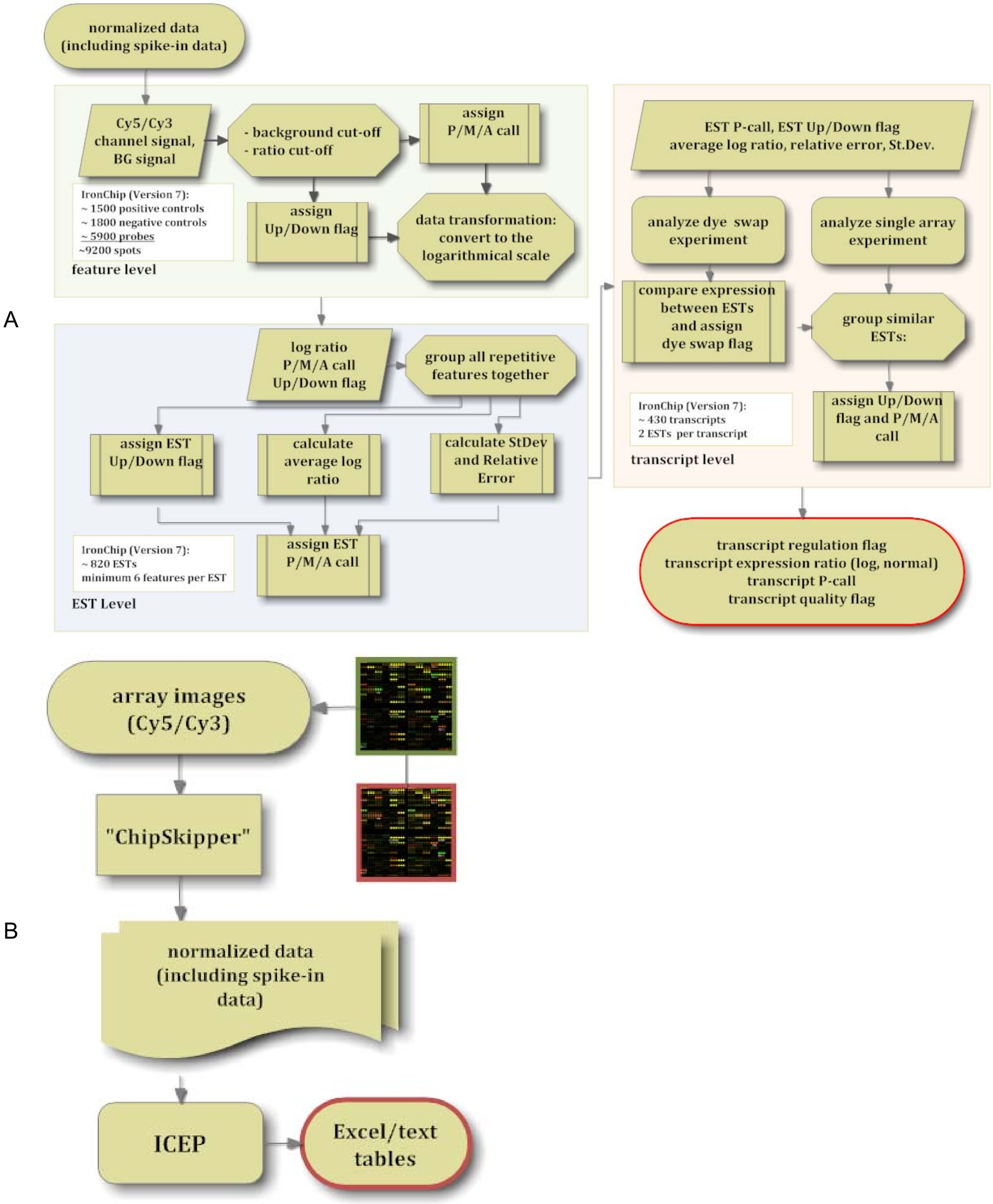
ICEP utilize a start panel, containing shortcuts to all utilities and ICEP analysis settings. From Analysis settings, in turn, it is possible to call Default column names configuration and gene name/gene description configuration files- 3 major files which allows adding new types of microarrays.

It helps the user to apply all analyses while hiding the complexity of the underlying statistical methods. The interface provides easy access to different layers of microarray analysis, including single array analysis, dye-swap analysis and the generation of a final report. A step by step user manual is available at: http://www.alice-dsl.net/evgeniy.vainshtein/ICEP/

Application of the analysis pipeline

The ICEP data analysis package was designed to be both highly flexible and user friendly. Data analysis involves an analysis pipeline that is composed of three elements: single array analysis, dye-swap experiment analysis and final report generation (Figure 10A).

Figure 10. IronChip analysis work flow


normalized data (including spike-in data)
Cy5/Cy3 channel signal, BG signal
- background cut-off
- ratio cut-off
assign P/M/A call
IronChip (Version 7):
~ 1500 positive controls
~ 1800 negative controls
~ 5900 probes
~9200 spots
assign Up/Down flag
data transformation: convert to the logarithmical scale
feature level
A
log ratio
P/M/A call
Up/Down flag
group all repetitive features together
assign EST Up/Down flag
calculate average log ratio
calculate StDev and Relative Error
IronChip (Version 7):
~ 820 ESTs
minimum 6 features per EST
assign EST P/M/A call
EST Level
EST P-call, EST Up/Down flag
average log ratio, relative error, St.Dev.
analyze dye swap experiment
analyze single array experiment
compare expression between ESTs and assign dye swap flag
group similar ESTs:
IronChip (Version 7):
~ 430 transcripts
2 ESTs per transcript
assign Up/Down flag and P/M/A call
transcript level
transcript regulation flag
transcript expression ratio (log, normal)
transcript P-call
transcript quality flag
array images (Cy5/Cy3)
"ChipSkipper"
B
normalized data (including spike-in data)
ICEP
Excel/text tables

A. *Flow chart of ICEP data processing and evaluation. Data evaluation with ICEP is organized into three functional modules: Single feature, EST and transcript evaluation.*

B. *In our application example, hybridized microarrays were scanned on a GenePix 4000B Microarray Scanner (Axon Instruments, Union City, CA, USA) and processed (feature background subtraction and normalization) by the ChipSkipper software [8] (Schwager 2002). ICEP uses these output files (generic tab-delimited text tables containing the normalized signal intensity and background data) for further analysis. (Vainshtein, Sanchez et al. 2010)*

Microarray analysis pipeline can either be executed directly or in a batch analysis mode.
In our application example, we utilized the mouse IronChip microarray that contains 520 genes as well as positive and negative controls. The controls are represented by 1400 spike-in control spots (positive controls) and 2400 background control spots (negative and specificity controls). 520 genes are represented by 880 ESTs. Genes are represented by 2 or more ESTs. Each EST was spotted on the array at least 6 times. 5400 spots in total are located on the array.

At the first level of analysis (single feature level), ICEP performs logarithmical transformation of the data and separates background and control spots from the rest of the data. ICEP then calculates a background cut-off value based on median signal intensities of all background and negative control spots, and an intensity ratio cut-off value, based on signals from the spike-in controls. Intensity ratios of all remaining genes are calculated as well. At the same time, ICEP performs a feature extraction procedure, whereby all repetitive features, representing the same EST are grouped together. After calculating the background and ratio cut-off values ICEP assigns the following flags: (1) the P-call flag (true positive call), which is based on a comparison of a signal intensity of each channel with the background cut-off value (Table 5A); (2) the regulation flag (significant difference in gene expression between two channels), which is based on comparison of a signal ratio to the ratio cut-off value (Table 5B).

Table 5. Flags definition

Feature P-call	Conditions
P	The signal intensity is higher than the background cut-off value in both channels
M	The signal intensity is higher than the background cut off value in one channel and lower in the other channel
A	The signal intensity is lower than the background cut off in both channels

A.

Regulation flag	Conditions
UP	Ratio between signal intensities is higher than corresponding ratio cut-off
DOWN	Ratio between signal intensities is lower than corresponding ratio cut-off
NONE	Ratio between signal intensities is in between upper and lower ratio cut-off

B.

A. P-call: Definition of a P-call flag (at the feature level) to distinguish bona fide signals from background noise
B. Regulation flag: Definition of a regulation flag (at the EST and the transcript level) to distinguish UP-/DOWN-/NONE-regulated ESTs and transcripts from each other

At the second level of analysis (EST level) ICEP assigns further flags to ESTs and estimates the data quality based on flags recorded on a single feature level. EST P-call flags are calculated by ICEP according to the rules given in Table 6. Definition of a P-call flag (at the EST level) is based on the P-calls of individual features. Corresponding threshold is set to 60%, due to the fact that a control microarray experiment (hybridization of a Hemin- and Desferrioxamine-treated HeLa cells) shows similar results to a published data only when a EST P-call threshold is about 60%. Significant increase of EST P-call threshold causes additional false negative while decreasing of this value cause additional false positive results.

Table 6. EST P-call definition.

Rule Nr.	EST P-call	Conditions	Description
1	P	P > 60% ≥ P	If more than 60% of features representing one EST have a p-call “P” then assign an EST P-call “P”
2	M	M > 60% ≥ M	If more than 60% of features representing one EST have a p-call “M” then assign an EST P-call “M”
3	A	A > 60% ≥ A	If more than 60% of features representing one EST have a p-call “A” then assign an EST P-call “A”
4	M	not 1,2 and 3: M + P > 60% ≥ M	If criteria 1,2 and 3 do not apply and more than 60% of features representing one EST have the p-calls “M” and “P” than assign an EST P-call “M”
5	A	not 1,2 ,3 and 4: M + A > 60% ≥ A	If criteria 1,2,3 and 4 do not apply and more than 60% of features representing one EST have the p-calls “M” and “A” than assign a EST P-call “A”

Definition of a P-call flag (at the EST level) is based on the P-calls of individual features and allows to distinguish UP-/DOWN-/NONE-regulated ESTs from each other.
60% was selected empirically as a threshold for P-call calculation according to many IronChip experiments to avoid additional false negative results (higher value) or false positive results (lower value). Confirmed by control microarray experiment with known data (hybridization of Hemin and Desferrioxamin treated HeLa cells)

The comparison of average or median signal intensity ratios to the previously calculated ratio cut-off value yields UP/DOWN/NONE-flags, similar to the flag calculations described above. At the EST level ICEP calculates the relative error: the ratio between the standard deviation and the average of signal intensity ratios of all features representing single ESTs. At the transcript level ICEP uses the relative error as a measure of reliability of technical and biological replicates.

Preceding transcript level analysis, ICEP analyses whether any bias has occurred as a consequence of the dye (Cy5 or Cy3 labelled nucleotides) incorporated into hybridization probes. To avoid such dye bias in two-colour microarray hybridizations the experimental and the control sample are routinely labelled with Cy5 and Cy3 labelled nucleotides, respectively, plus the other way around (dye swap). Such analysis avoids inconsistent signal intensity ratios that are artefacts due to the dye incorporated into the hybridization probe. Depending on whether an EST shows a similar average signal intensity ratio within the dye-swap data set, ICEP defines a dye-swap reliability flag (Table 7)

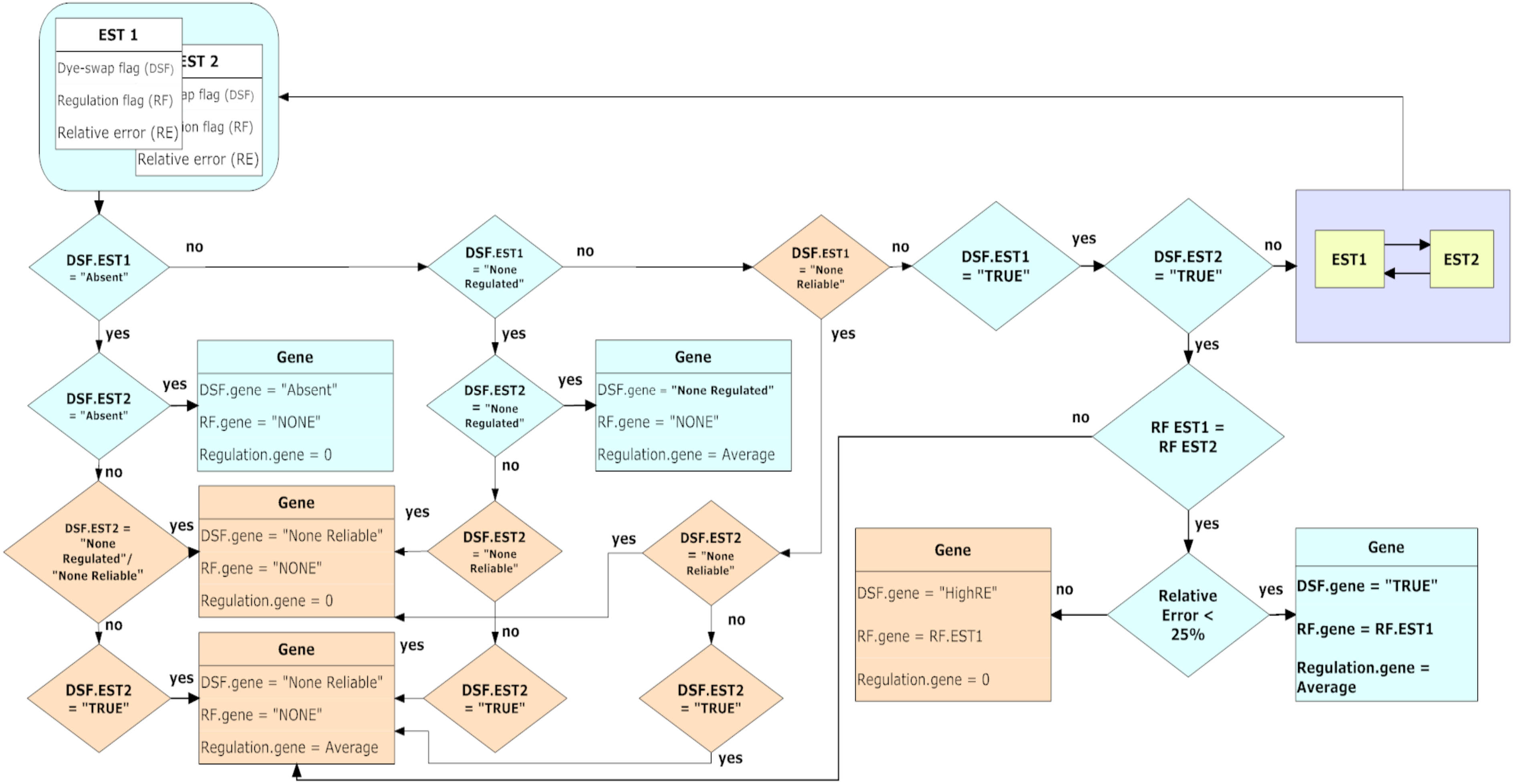
Table 7. Dye-swap reliability flag definition

Dye-swap flag (DS flag)	Conditions
Absent	EST shows the p-call “A” in the Cy5 and Cy3 experiment
Non reliable	The Cy5 and the Cy3 experiment show identical regulatory behavior (both UP, both DOWN)
Non regulated	EST does not show any regulation in both experiments
TRUE	EST shows “P” or “M” p-calls and is UP-regulated in the Cy5 and DOWN-regulated in the Cy3 experiment, or vice versa.
TRUE	EST shows “P” p-call and is UP-regulated or DOWN-regulated in the Cy5, while the Cy3 experiment shows a tendency towards the correct direction based on the ratio cut-off value, or vice versa
TRUE	EST shows “P” or “M” p-call, but both experiments show NONE-regulated with a tendency of regulation towards the correct direction based on a ratio cut-off

ICEP determines a reliability flag by evaluating the P-call and regulation flags of ESTs. The reliability flag is used by ICEP to distinguish reliable from unreliable expression changes.

On the transcript level ICEP applies grouping rules to calculate signal intensity ratios for a group of ESTs representing the same transcript. For this purpose ICEP is using all quality flags described before. On this level ICEP decides whether to average signal intensity ratios from different ESTs to a single value, to treat each EST as a separate transcript or to mark the complete set of ESTs as non-reliable. ICEP is able to analyse and group values from up to six similar ESTs representing a single gene (six is the maximum value in the current version of the IronChip microarray). To illustrate the grouping procedure a scheme is presented in Figure 11 which is based on 2 ESTs representing a single gene. Table 8 represents the possible flag combinations.

Figure 11. Schema of grouping two ESTs



On a block scheme boxes represent final or intermediate calculation steps; jewel boxes represent logical comparison. Rounded box contains incoming data and flags necessary for grouping procedure. Brown colour represents a none-reliable data. Blue colour represents reliable data. The box named "Gene" contains array of data and flags. It used as a result of a grouping procedure.

Table 8. ESTs grouping rules (Rule 1: 2 ESTs)

EST 1		EST 2		Relative Error	Transcript		
DS flag	**Regulation**	**DS flag**	**Regulation**		**Flag**	**value**	**Regulation**
absent	N.A.	absent	N.A.		absent	0	NONE
		non regulated	NONE		non reliable	0	NONE
		non reliable	N.A.		non reliable	0	NONE
		TRUE	UP/DOWN		non reliable	Average	NONE
non regulated	NONE	non regulated	NONE		non regulated	Average	NONE
		non reliable	N.A.		non reliable	0	NONE
		TRUE	UP/DOWN		non reliable	Average	NONE
non reliable	N.A.	non reliable	N.A.		non reliable	0	NONE
		TRUE	UP/DOWN		non reliable	Average	NONE
TRUE	UP	TRUE	UP	≥15	non reliable	Average	UP
			UP	<15	TRUE	Average	UP
		TRUE	DOWN		non reliable	Average	NONE
TRUE	DOWN	TRUE	DOWN	≥15	non reliable	Average	DOWN
			DOWN	<15	TRUE	Average	DOWN
			UP		non reliable	Average	NONE

The EST grouping rules represent rules for a decision-tree based algorithm to summarize expression data from several ESTs, representing one transcript. These are contained within a core table.

The variety of different flags combinations, we can separate on a 3 groups: regulated statistically significant and non-regulated; not detected on the array (the signal intensity is below the background cut-off level); and many intermediate stages: regulated, but with a high relative error (high deviation from a median or average), mostly regulated (several ESTs show one regulation while other ESTs representing single transcript showing other regulation or no regulation) and completely non-reliable when the dye-swap experiments fails for some ESTs or transcripts.

Such flexible gradation of a results helps to define a strategy for follow-up experiments.

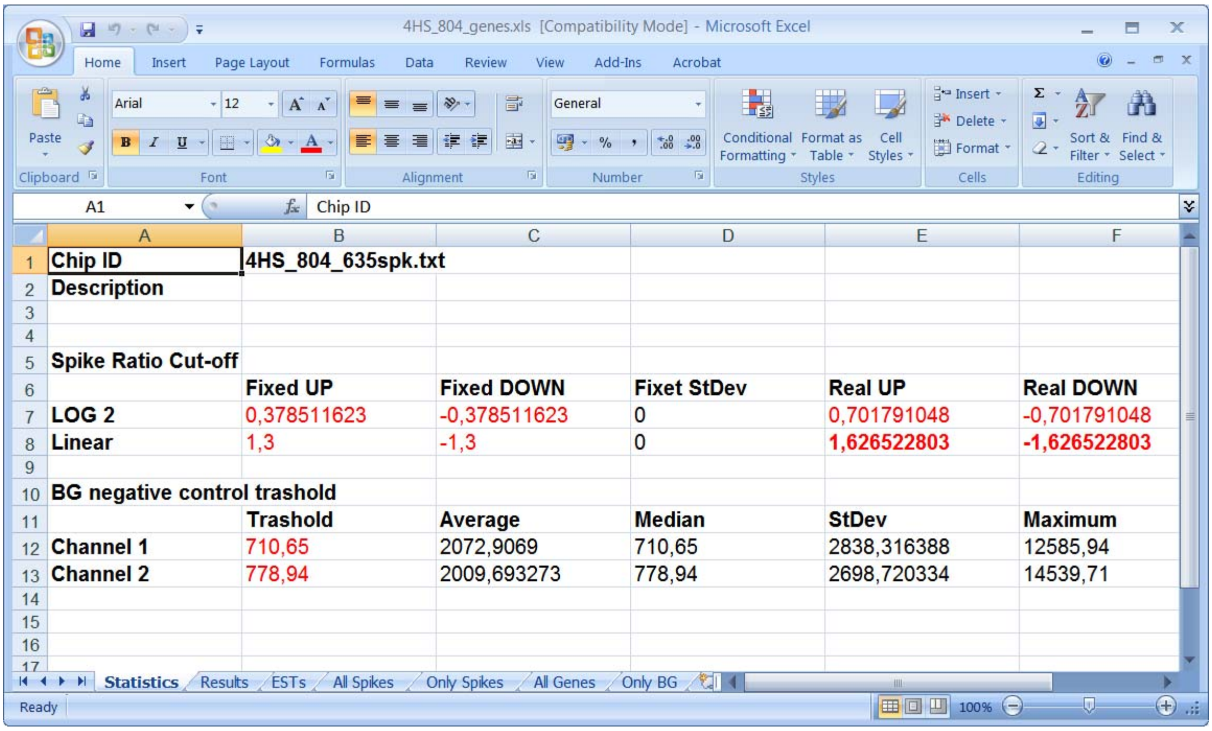
Using IronChip and ICEP for gene expression studies

Analysis by ICEP lead to generation of 3 resulting tables per every single microarray experiment (2 tab-delimited text files and 1 Excel table) and 2 resulting table per dye swap experiment (1 tab-delimited text file and 1 Excel table). For the moment there are more than 800 hybridisations of the IronChip done. This huge amount of data requires special approaches to organise easy access and made it possible to request all details. IronChip Laboratory Information Management System (IronLIMS) database was developed to store, search and manipulate this data. Currently it contains more than a 4000 samples more than 1000 samples amplified and labelled with either Cy3 or Cy5 and 800 unique hybridisations.

To illustrate in details how to use ICEP and to verify a method we selected a previously reported experiment that analyses iron-loaded (Hemin-treated) and iron-deficient (Desferrioxamine-treated) HeLa cells. Cellular iron overload or deficiency caused the expected changes in gene expression (Richter, Schwager et al. 2002; Muckenthaler, Richter et al. 2003). All illustration of ICEP output files is based on the described experiment (Table 2).

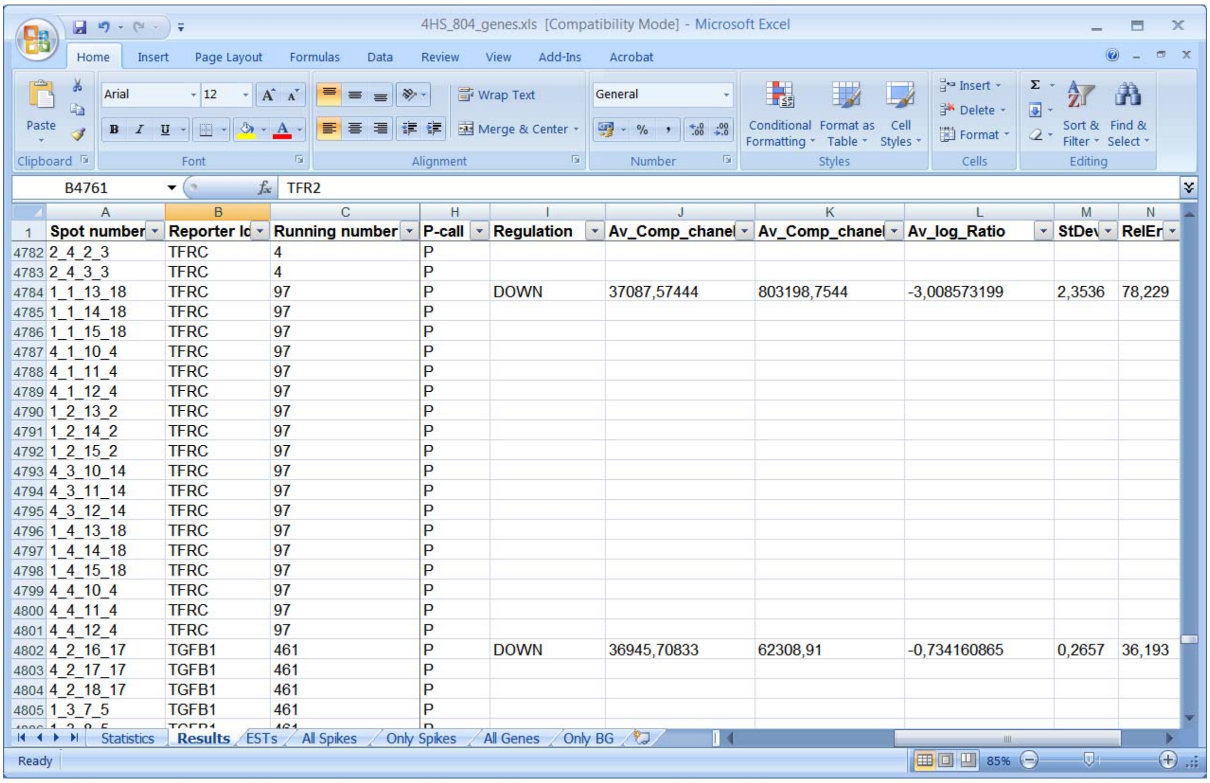
For hybridization we used mRNA material extracted from the Hemin-treated HeLa cell-line and compared it with mRNA extracted from the Desferrioxamine-treated HeLa cell-line. Both samples were labelled with Cy3 or Cy5 loading dye or vice versa and hybridized on the IronChip. Resulting arrays were scanned using Axon scanner and pre-processed with ChipSkipper application. With a ChipSkipper we read signal intensity information for each spot on the array, performed local background noise subtraction and normalization to compensate bias in signal intensity due to differences in Cy3/Cy5 dye activity.

Figure 12. ICEP output example.



A1 Chip ID

	A	B	C	D	E	F
1	Chip ID	4HS_804_635spk.txt				
2	Description					
3						
4						
5	Spike Ratio Cut-off					
6		Fixed UP	Fixed DOWN	Fixet StDev	Real UP	Real DOWN
7	LOG 2	0,378511623	-0,378511623	0	0,701791048	-0,701791048
8	Linear	1,3	-1,3	0	1,626522803	-1,626522803
9						
10	BG negative control trashold					
11		Trashold	Average	Median	StDev	Maximum
12	Channel 1	710,65	2072,9069	710,65	2838,316388	12585,94
13	Channel 2	778,94	2009,693273	778,94	2698,720334	14539,71
14						
15						
16						



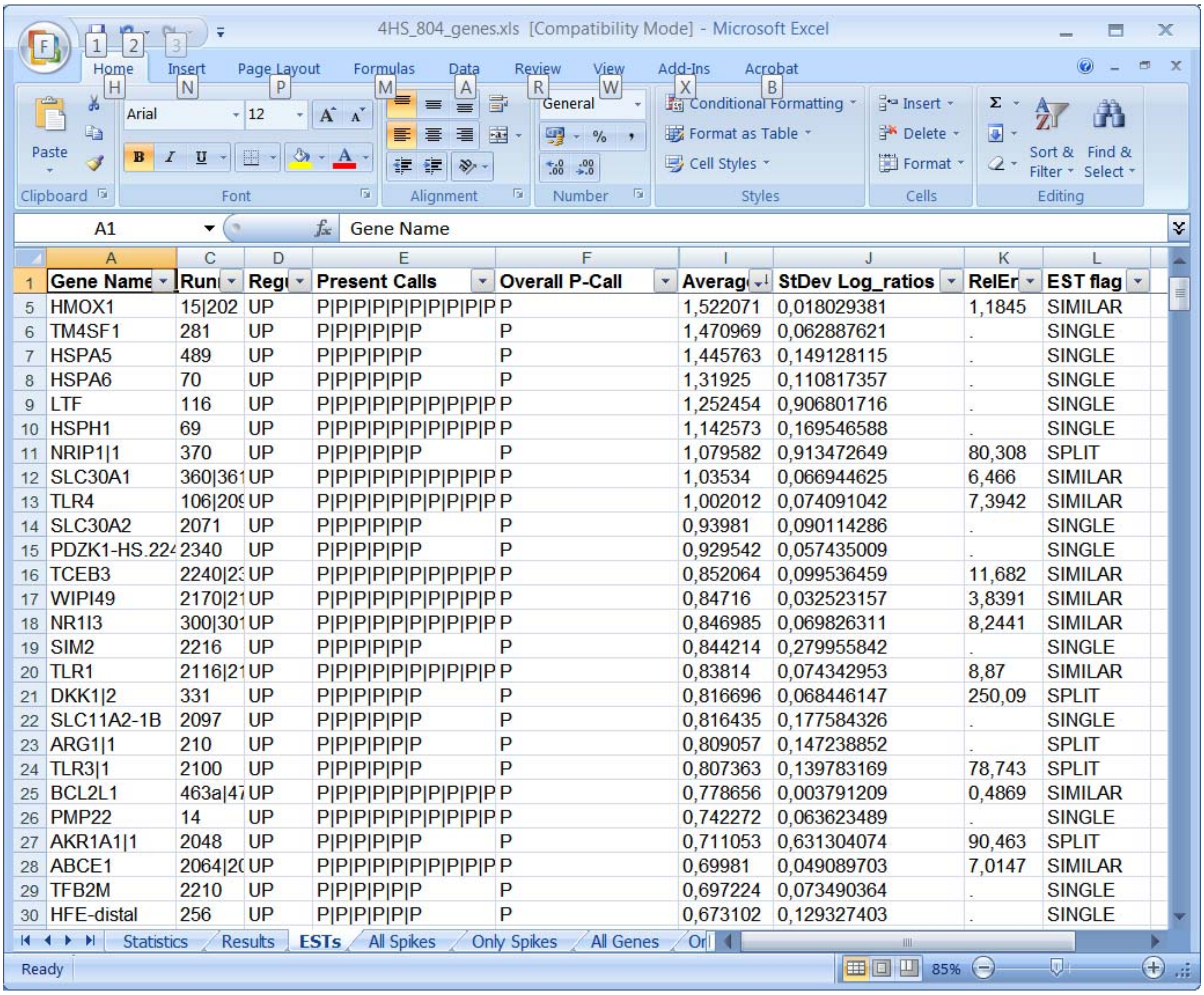
B4761 TFR2

	A	B	C	H	I	J	K	L	M	N
1	Spot number	Reporter Id	Running number	P-call	Regulation	Av_Comp_chanel	Av_Comp_chanel	Av_log_Ratio	StDev	RelEr
4782	2_4_2_3	TFRC	4	P						
4783	2_4_3_3	TFRC	4	P						
4784	1_1_13_18	TFRC	97	P	DOWN	37087,57444	803198,7544	-3,008573199	2,3536	78,229
4785	1_1_14_18	TFRC	97	P						
4786	1_1_15_18	TFRC	97	P						
4787	4_1_10_4	TFRC	97	P						
4788	4_1_11_4	TFRC	97	P						
4789	4_1_12_4	TFRC	97	P						
4790	1_2_13_2	TFRC	97	P						
4791	1_2_14_2	TFRC	97	P						
4792	1_2_15_2	TFRC	97	P						
4793	4_3_10_14	TFRC	97	P						
4794	4_3_11_14	TFRC	97	P						
4795	4_3_12_14	TFRC	97	P						
4796	1_4_13_18	TFRC	97	P						
4797	1_4_14_18	TFRC	97	P						
4798	1_4_15_18	TFRC	97	P						
4799	4_4_10_4	TFRC	97	P						
4800	4_4_11_4	TFRC	97	P						
4801	4_4_12_4	TFRC	97	P						
4802	4_2_16_17	TGFB1	461	P	DOWN	36945,70833	62308,91	-0,734160865	0,2657	36,193
4803	4_2_17_17	TGFB1	461	P						
4804	4_2_18_17	TGFB1	461	P						
4805	1_3_7_5	TGFB1	461	P						

A. First worksheet of an Excel workbook contains important statistical information corresponding to single array analysis

B. Resulting table for an EST analysis level. All features grouped by a "running number", on the left side all statistical data for each feature, at the right side- average values for a EST. EST-related values calculated based on features grouping rules.

Figure 13. ICEP output example



A1 Gene Name

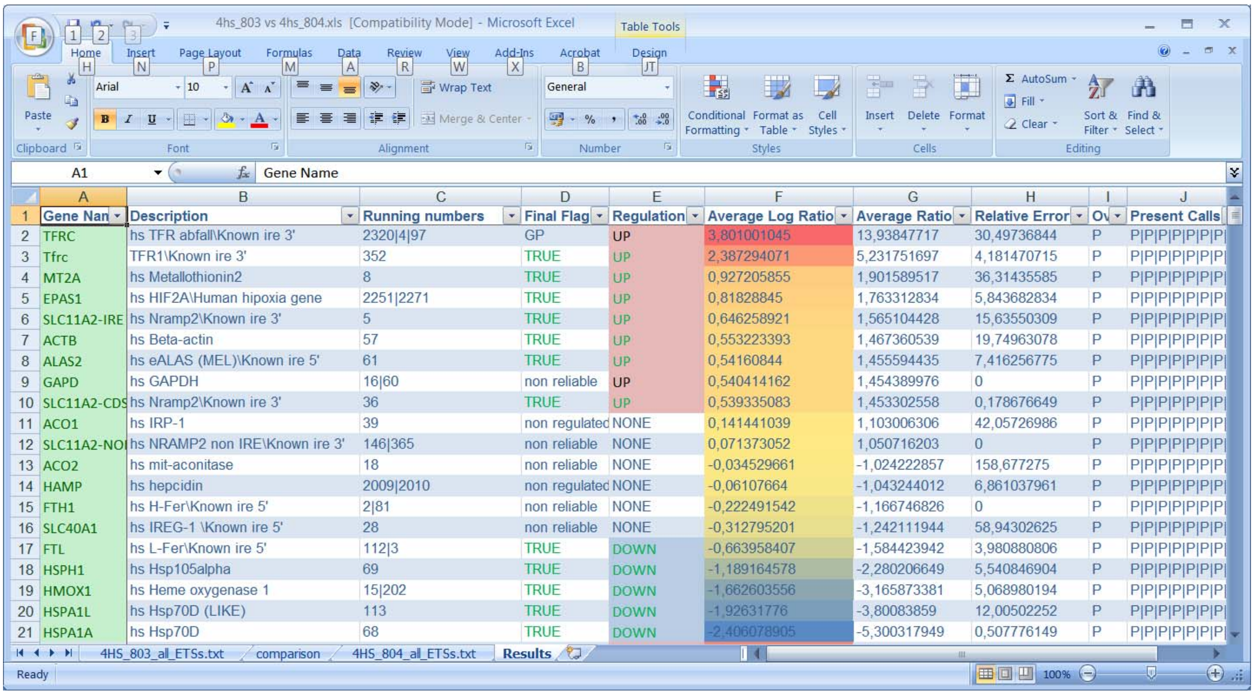
	A	C	D	E	F	I	J	K	L
1	Gene Name	Runi	Regı	Present Calls	Overall P-Call	Averag	StDev Log_ratios	RelEr	EST flag
5	HMOX1	15|202	UP	P|P|P|P|P|P|P|P|P|P	P	1,522071	0,018029381	1,1845	SIMILAR
6	TM4SF1	281	UP	P|P|P|P|P|P	P	1,470969	0,062887621	.	SINGLE
7	HSPA5	489	UP	P|P|P|P|P|P	P	1,445763	0,149128115	.	SINGLE
8	HSPA6	70	UP	P|P|P|P|P|P	P	1,31925	0,110817357	.	SINGLE
9	LTF	116	UP	P|P|P|P|P|P|P|P|P|P	P	1,252454	0,906801716	.	SINGLE
10	HSPH1	69	UP	P|P|P|P|P|P|P|P|P|P	P	1,142573	0,169546588	.	SINGLE
11	NRIP1|1	370	UP	P|P|P|P|P|P	P	1,079582	0,913472649	80,308	SPLIT
12	SLC30A1	360|361	UP	P|P|P|P|P|P|P|P|P|P	P	1,03534	0,066944625	6,466	SIMILAR
13	TLR4	106|209	UP	P|P|P|P|P|P|P|P|P|P	P	1,002012	0,074091042	7,3942	SIMILAR
14	SLC30A2	2071	UP	P|P|P|P|P|P	P	0,93981	0,090114286	.	SINGLE
15	PDZK1-HS.224	2340	UP	P|P|P|P|P|P	P	0,929542	0,057435009	.	SINGLE
16	TCEB3	2240|23	UP	P|P|P|P|P|P|P|P|P|P	P	0,852064	0,099536459	11,682	SIMILAR
17	WIPI49	2170|21	UP	P|P|P|P|P|P|P|P|P|P	P	0,84716	0,032523157	3,8391	SIMILAR
18	NR1I3	300|301	UP	P|P|P|P|P|P|P|P|P|P	P	0,846985	0,069826311	8,2441	SIMILAR
19	SIM2	2216	UP	P|P|P|P|P|P	P	0,844214	0,279955842	.	SINGLE
20	TLR1	2116|21	UP	P|P|P|P|P|P|P|P|P|P	P	0,83814	0,074342953	8,87	SIMILAR
21	DKK1|2	331	UP	P|P|P|P|P|P	P	0,816696	0,068446147	250,09	SPLIT
22	SLC11A2-1B	2097	UP	P|P|P|P|P|P	P	0,816435	0,177584326	.	SINGLE
23	ARG1|1	210	UP	P|P|P|P|P|P	P	0,809057	0,147238852	.	SPLIT
24	TLR3|1	2100	UP	P|P|P|P|P|P	P	0,807363	0,139783169	78,743	SPLIT
25	BCL2L1	463a|47	UP	P|P|P|P|P|P|P|P|P|P	P	0,778656	0,003791209	0,4869	SIMILAR
26	PMP22	14	UP	P|P|P|P|P|P|P|P|P|P	P	0,742272	0,063623489	.	SINGLE
27	AKR1A1|1	2048	UP	P|P|P|P|P|P	P	0,711053	0,631304074	90,463	SPLIT
28	ABCE1	2064|20	UP	P|P|P|P|P|P|P|P|P|P	P	0,69981	0,049089703	7,0147	SIMILAR
29	TFB2M	2210	UP	P|P|P|P|P|P	P	0,697224	0,073490364	.	SINGLE
30	HFE-distal	256	UP	P|P|P|P|P|P	P	0,673102	0,129327403	.	SINGLE

Final results for a single array experiment. Here expression values for all ESTs, representing single transcript, averaged applying grouping rules for ESTs

ChipSkipper generates tab-delimited text files containing information about physical localisation of a spot on the array, signal intensity information some quality flags and many other additional data.

A result of the analysis of each array illustrates Figure 12 and 13. Resulting Excel tables contain important statistical information about the analysis: background cut-off, ration cut-off, St.Dev, average, median and maximum signal from each channel (Figure 12A). Background cut-off level determines a sensitivity level of the hybridised array and ratio cut-off determines selectivity of the array analysis. Maximum, minimum and average signal levels helps to estimate overall hybridisation quality (signal/noise ratio, spread of the data). Figure 12B illustrates results on the single feature level and Figure 13 analysis results on the EST level. In the experiment described above, we analysing a colour-swap data. Therefore, on the final step ICEP summarizes results from direct and dye-swapped experiments. Based on ESTs regulation/P-call flags, statistical data and between dye-swapped arrays comparison ICEP applies grouping rules to generate a final table, containing only summarized expression data for every transcript. Figure 14 illustrates this for our example experiment:

Figure 14. Dye-swap experiment results



	Gene Nan	Description	Running numbers	Final Flag	Regulation	Average Log Ratio	Average Ratio	Relative Error	Ov	Present Calls
2	TFRC	hs TFR abfall\Known ire 3'	2320|4|97	GP	UP	3,801001045	13,93847717	30,49736844	P	P|P|P|P|P|P|P|
3	Tfrc	TFR1\Known ire 3'	352	TRUE	UP	2,387294071	5,231751697	4,181470715	P	P|P|P|P|P|P|P|
4	MT2A	hs Metallothionin2	8	TRUE	UP	0,927205855	1,901589517	36,31435585	P	P|P|P|P|P|P|P|
5	EPAS1	hs HIF2A\Human hipoxia gene	2251|2271	TRUE	UP	0,81828845	1,763312834	5,843682834	P	P|P|P|P|P|P|P|
6	SLC11A2-IRE	hs Nramp2\Known ire 3'	5	TRUE	UP	0,646258921	1,565104428	15,63550309	P	P|P|P|P|P|P|P|
7	ACTB	hs Beta-actin	57	TRUE	UP	0,553223393	1,467360539	19,74963078	P	P|P|P|P|P|P|P|
8	ALAS2	hs eALAS (MEL)\Known ire 5'	61	TRUE	UP	0,54160844	1,455594435	7,416256775	P	P|P|P|P|P|P|P|
9	GAPD	hs GAPDH	16|60	non reliable	UP	0,540414162	1,454389976	0	P	P|P|P|P|P|P|P|
10	SLC11A2-CDS	hs Nramp2\Known ire 3'	36	TRUE	UP	0,539335083	1,453302558	0,178676649	P	P|P|P|P|P|P|P|
11	ACO1	hs IRP-1	39	non regulated	NONE	0,141441039	1,103006306	42,05726986	P	P|P|P|P|P|P|P|
12	SLC11A2-NOI	hs NRAMP2 non IRE\Known ire 3'	146|365	non reliable	NONE	0,071373052	1,050716203	0	P	P|P|P|P|P|P|P|
13	ACO2	hs mit-aconitase	18	non reliable	NONE	-0,034529661	-1,024222857	158,677275	P	P|P|P|P|P|P|P|
14	HAMP	hs hepcidin	2009|2010	non regulated	NONE	-0,06107664	-1,043244012	6,861037961	P	P|P|P|P|P|P|P|
15	FTH1	hs H-Fer\Known ire 5'	2|81	non reliable	NONE	-0,222491542	-1,166746826	0	P	P|P|P|P|P|P|P|
16	SLC40A1	hs IREG-1 \Known ire 5'	28	non reliable	NONE	-0,312795201	-1,242111944	58,94302625	P	P|P|P|P|P|P|P|
17	FTL	hs L-Fer\Known ire 5'	112|3	TRUE	DOWN	-0,663958407	-1,584423942	3,980880806	P	P|P|P|P|P|P|P|
18	HSPH1	hs Hsp105alpha	69	TRUE	DOWN	-1,189164578	-2,280206649	5,540846904	P	P|P|P|P|P|P|P|
19	HMOX1	hs Heme oxygenase 1	15|202	TRUE	DOWN	-1,662603556	-3,165873381	5,068980194	P	P|P|P|P|P|P|P|
20	HSPA1L	hs Hsp70D (LIKE)	113	TRUE	DOWN	-1,92631776	-3,80083859	12,00502252	P	P|P|P|P|P|P|P|
21	HSPA1A	hs Hsp70D	68	TRUE	DOWN	-2,406078905	-5,300317949	0,507776149	P	P|P|P|P|P|P|P|

Formatting of data cells was done manually to visualize UP/DOWN regulated transcripts

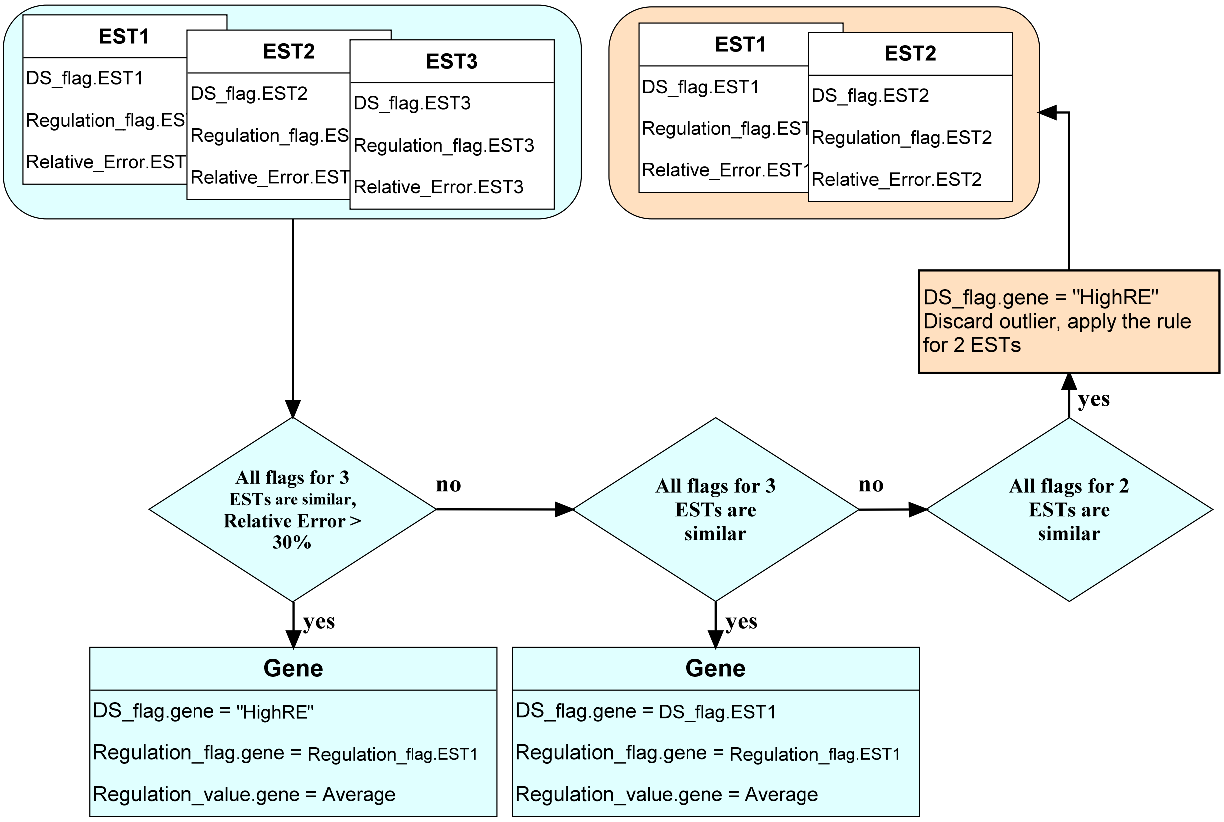
Resulting table contains 4 worksheets: resulting files of a single array analysis (automatically generated by ICEP with "_EST.txt" extension), EST by EST comparison of two dye-swapped arrays and a final result for each transcript on the array. Every transcript displayed in the "Results" worksheet is characterized by a combination of flags. "Regulation" and "Transcript P-call" flags are similar to corresponding EST and single feature flags. Regulation could take a "UP/DOWN/NONE" value and "Transcript P-call" flag could be "P/M/A". The "Final flag" is summarizing all flags mentioned above with a relative error and helps to distinguish several groups of transcripts, depending on its value: "TRUE" – all ESTs representing a transcript are regulated statistically significant in both dye swap experiments and have a proper regulation[1]; GP – "grouping problems" – there are 3 or more ESTs representing a transcript and most ESTs are regulated statistically significant in a proper way; "HighRE" – "high relative error"- is similar

---

[1] During dye-swap experiment the experimental sample is labeled first with a Cy3 dye while a control sample is labeled with a Cy5 dye. For a second hybridization experimental and control samples are labeled with Cy5 and Cy3 dyes correspondingly. Therefore if a cDNA of a certain EST is overrepresented in the experimental sample in comparison with the control sample it will be up-regulated in first hybridization experiment and down-regulated in the second dye-swap hybridization experiment. Such regulation we call here "proper regulation"

to the “TRUE” situation, but the deviations of a regulation from the average value is too big (relative error is bigger than 30%); “Non-regulated” – the regulation of all ESTs representing certain transcript is below the regulation cut-off value for both dye swap hybridisations; “Non reliable” – corresponds to the situation, when only one EST from the group representing a transcript shows proper regulation, while other ESTs shows no proper regulation. The ICEP dye swap comparison utility analyses every case using set of rules. If transcript represented by only 1 EST, ICEP used values and flags from single EST analysis. If there are 2 or more ESTs representing a transcript, ICEP applies a decision tree-based algorithm to define all flags (for example, see Table 8 and Figure 11 above for a description of rules applied in the case of 2 ESTs, representing single transcript). Figure 15 represents similar algorithm for a 3 ESTs grouping:

Figure 15. Grouping of 3 ESTs


EST1
DS_flag.EST1
Regulation_flag.ES
Relative_Error.EST
EST2
DS_flag.EST2
Regulation_flag.ES
Relative_Error.EST
EST3
DS_flag.EST3
Regulation_flag.EST3
Relative_Error.EST3
EST1
DS_flag.EST1
Regulation_flag.EST
Relative_Error.EST1
EST2
DS_flag.EST2
Regulation_flag.EST2
Relative_Error.EST2
DS_flag.gene = "HighRE"
Discard outlier, apply the rule for 2 ESTs
yes
All flags for 3 ESTs are similar, Relative Error > 30%
no
All flags for 3 ESTs are similar
no
All flags for 2 ESTs are similar
yes
yes
Gene
DS_flag.gene = "HighRE"
Regulation_flag.gene = Regulation_flag.EST1
Regulation_value.gene = Average
Gene
DS_flag.gene = DS_flag.EST1
Regulation_flag.gene = Regulation_flag.EST1
Regulation_value.gene = Average

Blue rounded box contains initial data and flags for the analysis using 3 ESTs grouping rule procedure. All logical operations are represented by a blue jewel boxes. The resulting flags are in the blue boxes with a title “Gene”. If analysis identifies that only 2 ESTs from 3 have similar flags, than 3ESTs procedure discards information about outlier and call 2ESTs procedure for the rest. Corresponding steps are represented by brown colour on a block scheme.

In the case of 3 and more ESTs situation can be simplified: if all ESTs behaves similarly (same regulation flags, same P-calls and low relative error) than ICEP average regulation values and assign final flags according to the Table 9 (see below). If one EST from the group of 3 ESTs or 2-3 ESTs from the group of 4, 5 or 6 ESTs behaving non-consistently with the major group of similar ESTs, than ICEP discard non-consistent values and process similar ESTs with one of the previous rules (for a group of 3 ESTs with 1 non-consistent- discard 1 non-reliable or non-consistent EST and use 2ESTs rule for the rest).

Table 9. 3ESTs rule: processing 3 ESTs representing one transcript

EST 1		EST 2		EST 3		Relative Error	Apply rule	Transcript		
DS flag	**Regulation**	**DS flag**	**Regulation**	**DS flag**	**Regulation**			**Flag**	**value**	**Regulation**
absent	N.A.	absent		absent				absent	0	NONE
		non-regulated		non-regulated				non-regulated	Average	NONE
		non reliable		non reliable				non reliable	0	NONE
TRUE	UP	TRUE	DOWN	TRUE	UP	≥ 15		Non reliable	Average	UP
					UP	< 15		TRUE	Average	UP
				TRUE	DOWN		2ESTs	-	-	-
TRUE	DOWN	TRUE	DOWN	TRUE	DOWN	≥ 15		Non reliable	average	DOWN
					DOWN	< 15		TRUE	average	DOWN
				TRUE	UP		2ESTs			
TRUE		TRUE		absent						
				non-regulated			2ESTs			
				non reliable						
TRUE		absent		absent						
		non reliable		non reliable			2ESTS			
		non-regulated		non-regulated						

The EST grouping rules represent rules for a decision-tree based algorithm to summarize expression data from 3 ESTs, representing one transcript. These are contained within a table.

Rules for 4 and more ESTs are difficult to represent in the form of tables or schemes, but we always use the same principle: group similar ESTs and discard outliers by comparing flags and verify values with the dye swap data. Discarding of outliers allows applying rules of lower order.

Cellular iron overload or deficiency caused the expected changes in gene expression (Richter, Schwager et al. 2002; Muckenthaler, Richter et al. 2003). Table 10 represents results of a dye-swap experiment. Application of ICEP reveals the previously reported and experimentally validated changes in mRNA expression of genes such as Tfrc (Transferrin receptor 1), Slc11a2 (NRAMP2; DMT1: Natural resistance-associated macrophage protein 2; Divalent metal transporter 1) and Ftl (Ferritin light chain; Ferritin L subunit).

Table 10. Expression increase in iron-loaded and iron-deficient HeLa cells

Gene name	Regulation (D/H)	Final flag	Ratio(av)	P-call (av)
TFRC	UP	TRUE	5,23 ± 0,22	P
MT2A	UP	TRUE	1,9 ± 0,69	P
EPAS1	UP	TRUE	1,76 ± 0,1	P
SLC11A2	UP	TRUE	1,57 ± 0,24	P
ACTB	UP	TRUE	1,47 ± 0,29	P
ALAS2	UP	TRUE	1,46 ± 0,11	P
FTL	DOWN	TRUE	-1,58 ± 0,06	P
HSPH1	DOWN	TRUE	-2,28 ± 0,13	P
HMOX1	DOWN	TRUE	-3,17 ± 0,16	P
HSPA1L	DOWN	TRUE	-3,8 ± 0,46	P
HSPA1A	DOWN	TRUE	-5,3 ± 0,03	P

The table represents the average ratios of differentially expressed genes in Hemin-and Desferrioxamine-treated HeLa cells. The relative errors are shown. The table contains only selected columns and genes.

The very last step of data evaluation before actual analysis by a biologist is a comparison of gene expression profiles in different experimental conditions. Partially this could be done by using one of the ICEP's built-in utilities: convert multiple text files to a multi-sheets Excel workbook. This utility helps to put together all expression profiles related to a biological experiment to a single file. Partially helps external utility for alignment of gene expression profiles. Of cause the very last step of actual analysis should be done manually. The figure below represents such a resulting table (Figure 16).

Sorting of transcripts according to the regulation and colour codes where done in Excel manually to simplify the representation of a final data.

Iron metabolism and gene expression analysis

The regulation and maintenance of iron homeostasis is critical to human health. As a constituent of haemoglobin, iron is essential for oxygen transport and significant iron deficiency leads to anemia. Eukaryotic cells require iron for survival and proliferation. Iron is part of hemoproteins, iron-sulfur (Fe-S) proteins, and other proteins with functional groups that require iron as a cofactor.

Individuals with hereditary hemochromatosis suffer from systemic iron overload due to duodenal hyperabsorption. Most cases arise from a founder mutation in HFE (845G-->A; ref. 2) that results in the amino-acid substitution C282Y and prevents the association of HFE with beta2-microglobulin. Mice homozygous with respect to a null allele of Hfe (Hfe-/-) or homozygous with respect to the orthologous 882G-->A mutation (Hfe(845A/845A)) develop iron overload that recapitulates hereditary hemochromatosis in humans, confirming that hereditary hemochromatosis arises from loss of HFE function.

At the cellular level, iron uptake, utilization, storage, and export are regulated at different molecular levels (transcriptional, mRNA stability, translational, and posttranslational). Iron regulatory proteins (IRPs) 1 and 2 post-transcriptionally control mammalian iron homeostasis by binding to iron-responsive elements (IREs), conserved RNA stem-loop structures located in the 5'- or 3'- untranslated regions of genes involved in iron metabolism (e.g. FTH1, FTL, and TFRC). To identify novel IRE-containing mRNAs, we integrated biochemical, biocomputational, and microarray-based experimental approaches.

Gene expression studies greatly contribute to our understanding of complex relationships in gene regulatory networks. However, the complexity of array design, production and manipulations are limiting factors, affecting data quality. The use of customized DNA microarrays improves overall data quality in many situations, however, only if for these specifically designed microarrays analysis tools are available.

ICEP tool allows analysing very complex data set and generating enough data to start building iron regulatory network. Table 11 contains detailed description of hybridisation experiments performed to understand the role of primary or secondary iron overload in different mouse tissues, such as liver, spleen or duodenum, due to iron treatment or knock-out mutation. The preliminary results of these experiments are not shown here due to their complexity. All experimental data for these user cases were provided by M.Sanchez and M.U. Muckenthaler.

Table 11 Experimental setup

Sample 1				Sample 2			
Genetic background	**Mutation**	**Treatment**	**Organ**	**Genetic Background**	**Mutation**	**Treatment**	**Organ**
SV129	None	Untreated	Liver	SV129	None	FE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	HFE -/-	NONE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	HFE -/-	FE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	B2M -/-	NONE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	B2M -/-	FE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	C282Y	NONE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	C282Y	FE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	B2M -/- HFE -/-	NONE	Liver
			Spleen				Spleen
			Duodenum				Duodenum
SV129	None	Untreated	Liver	SV129	B2M -/- HFE -/-	FE	Liver
			Spleen				Spleen
			Duodenum				Duodenum

All samples derived from same genetic background. Left column contains control samples, derived from naive animals. Right column contains samples of the same genetic background but undergoing different treatments or/and mutations.

Discussion

Identification of a novel IRE-containing genes

The search for IREs is a long term endeavour. It started with the first experimental discoveries of IREs (Hentze, Caughman et al. 1987). Subsequently Dandekar and Hentze started motif-searches to look for IREs (Dandekar, Stripecke et al. 1991; Gray, Pantopoulos et al. 1996). In the meantime many people got interested, for instance (Cox, Bawden et al. 1991): later work by Theil (1996) and a number of IRE searching programs, such as UTRscan etc. So when I came to the field there was a lot of research going on and different strategies available. The example shown in the first chapter, Cdc14a, shows however, why the search for IREs remains a continued challenge:

The IRE consensus got extended and modified. Different methods capture different things. IREs turn out to be always more wide-spread than previously thought. A nice example for the latter point is our claim (Dandekar, Beyer et al. 1998) that IREs may also occur in bacteria such as *E.coli*, now this is standard knowledge (bacterial aconitase moonlighting Joerg Stuehlke 2009). In the same context, Cdc14a has to be seen: The standard IRE search failed (Bengert algorithm) however with the iron-chip and ICEP it could be found.

Cdc14a – possible role in iron homeostasis

CDC14A is a novel IRE-containing and iron-regulated mRNA. CDC14A is one of the two human orthologs of the yeast CDC14 (cell division cycle 14) gene that has been shown to encode a phosphatase involved in the dephosphorylation of several critical cell cycle proteins (Wong, Chen et al.). Loss of function mutations of the CDC14A gene have been described in various human cancer cell lines, suggesting that CDC14A could act as a tumor suppressor (Wong, Chen et al.). Indeed, CDC14A has been shown to dephosphorylate cdk substrates such as p27kip1 and cyclin E that are critical for the G1 to S phase progression (Kaiser, Zimmerman et al. 2002). Alteration of CDC14A expression by RNA interference or transgenic overexpression has been found to cause abnormal mitotic spindle assembly and chromosome segregation (Kaiser, Zimmerman et al. 2002; Mailand, Lukas et al. 2002), arguing that CDC14A plays an important role in cell division.

Present and previous studies (Wong, Chen et al. 1999; Kaiser, Zimmerman et al. 2002) revealed several mRNA isoforms with heterogeneity at both the 5'- and 3'-ends of the CDC14A mRNA. These mRNA isoforms are predicted to encode different protein products that differ in their N and C termini. This heterogeneity is highly reminiscent of SLC11A2, another 3'-IRE-containing gene that encodes several protein isoforms (Hubert and Hentze 2002) with distinct sub-cellular localizations (Tabuchi, Tanaka et al. 2002; Lam-Yuk-Tseung and Gros 2006). Analogously, the N- and C-terminal heterogeneity of CDC14A proteins could affect the targeting of the phosphatase within the cell. Immunofluorescence studies using antibodies that do not discriminate between the CDC14A isoforms revealed both cytoplasmic and centrosomal staining (Mailand, Lukas et al. 2002). Here, we identify a novel 5'-exon (exon 1A) predicted to contain an N-myristoylation site (GNFLSR) that may target the protein to membranes (Farazi, Waksman et al. 2001). Further work will explore this important aspect of CDC14A biology.

Cell type-specific iron regulation also characterizes the SLC11A2-IRE mRNA isoforms (Gunshin, Allerson et al. 2001). These findings suggest that iron control of the CDC14A- and SLC11A2-IRE mRNA isoforms requires yet unidentified cell-specific regulatory factors, in contrast to the ubiquitous iron regulation of the *TFRC* mRNA.

RT-PCR analysis demonstrated that the increase in CDC14A mRNA levels is clearly selective for the presence of the IRE in the 3'-UTR, but it does not require transcription initiation at either exon 1A or 1B. This result makes it rather unlikely that iron regulation of CDC14A mRNA expression is due to transcriptional activation, although a transcriptional mechanism cannot be ruled out formally at present. The conservation of the proposed TFRC mRNA endonucleolytic cleavage site in the CDC14A mRNA rather suggests the possibility that a similar mechanism of mRNA stabilization could be involved. Interestingly, the sequence conservation does not extend into the flanking sequences, which could possibly contribute to the differential regulation observed. However, CDC14A mRNA turnover was not detectably slowed in iron-deprived cells treated with DRB. Although the mechanism underlying TFRC and CDC14 mRNA up-regulation could be distinct, it is also possible that our experimental conditions masked mRNA stabilizing effects of IRP binding. It is possible that the half-life of CDC14A mRNA significantly exceeds the time course of our DRB experiments, but this time course could not be prolonged because of the adverse toxic effects of DRB treatment. Notably, the regulation of the expression of the SLC11A2 isoforms appears to be complex and to occur at the transcriptional, posttranscriptional, and post-translational levels (Zoller, Theurl et al. 2002; Lis, Paradkar et al. 2005; Wang, Garrick et al. 2005; Garrick, Kuo et al. 2006). It is possible that the one or more mechanisms responsible for the iron regulation of mRNAs with single 3'-UTR IREs (i.e.

CDC14A and SLC11A2) differ from the mechanism that controls the TFRC mRNA with 5 IRE motifs in its 3’-UTR.

The selective iron modulation of the CDC14A-IRE mRNA isoforms points toward a possible, previously unrecognized link between iron metabolism, the IRE/IRP regulatory system, and cell cycle progression.

An IRE motif is characterized by a 6-nucleotide apical loop with the consensus sequence 5’-CAGWGH-3’ on an upper stem of 5 bp, a small asymmetric bulge with an unpaired cytosine and an additional lower stem of variable length. We performed a biocomputational screen for IRE-like motifs, including restricted search criteria to minimize the number of false positives. Nonetheless, most of the 21 candidate mRNAs could not be detected in immunoselected IRP/IRE mRNPs when assessed by IronChip or qRT-PCR in RNA samples from several sources. Possible reasons for falsely negative results include the lack of expression of these mRNAs in the cell types or tissues tested, the masking of an IRE-containing isoform by more abundant non- IRE splice variants, or the exclusive binding to IRP2 instead of IRP1. Therefore, we characterized the binding of some of these predicted IREs to IRP1 (11 candidates tested) or IRP2 (6 candidates tested) using competitive EMSA. With the notable exception of the human CDC14A IRE, none of the tested IRE candidates was able to compete efficiently for IRP1 or IRP2 binding to the *FTH1* IRE probe. These data clearly suggest that many of the IRE-like motifs identified by the biocomputational screen possess low IRP binding affinity. Previous biocomputational IRE motif searches included primary RNA sequence information and RNA folding criteria consistently guided by the folding energy of known IREs (at least -3 kcal/mol or below) (15). These searches missed the CDC14A IRE (-1.4 kcal/mol) that was identified here.

IREG1 and HIF2a were identified as well using a combination of biocomputational and biochemical approach (McKie, Marciani et al. 2000; Sanchez, Galy et al. 2007).

IRE-containing genes were identified previously not only in mouse or human, but as well in other species. For example, *Drosophila melanogaster* succinate dehydrogenase iron protein (SDH-IP) is an example of mRNA species being translationally regulated by an IRE. SDH is a three subunit Krebs cycle complex reducing succinate into fumarate and concomitantly generating FADH2. The functional SDH complex consists of a flavoprotein, a cytochrome and an iron protein (IP) subunit. The approximately 30 kDal IP subunit (SDH-IP) contains three different iron-sulfur clusters which are needed for electron transfer to quinones in the respiratory chain. Unlike other Krebs cycle enzymes, SDH is an integral part of the respiratory chain Complex II. (Melefors 1996)

A more recent study shows, for example that a Myotonic dystrophy kinase-related Cdc42-binding kinase alpha (MRCKalpha, formally known as CDC42BPA) a functional iron responsive element (IRE) in the 3'-untranslated region (UTR) of its mRNA, suggesting that it may be involved in iron metabolism and MRCKalpha protein expression is also regulated by iron levels; MRCKalpha colocalizes with transferrin (Tf)-loaded transferrin receptors (TfR), and attenuation of MRCKalpha expression by a short hairpin RNA silencing construct leads to a significant decrease in Tf-mediated iron uptake. Results of a study indicate that MRCKalpha takes part in Tf-iron uptake, probably via regulation of Tf-TfR endocytosis/endosome trafficking that is dependent on the cellular cytoskeleton. Regulation of the MRCKalpha activity by intracellular iron levels could thus represent another molecular feedback mechanism cells could use to finely tune iron uptake to actual needs (Cmejla, Ptackova et al. 2010).

The posttranscriptional control of iron uptake, storage, and utilization by iron-responsive elements (IREs) and iron regulatory proteins (IRPs) provides a molecular framework for the regulation of iron homeostasis in many animals. NICOLA K. GRAY et al. in 1996 demonstrated that *Drosophila melanogaster* IRP binds to an IRE in the 5' untranslated region of the mRNA encoding the iron-sulfur protein (Ip) subunit of succinate dehydrogenase (SDH). This interaction is developmentally regulated during Drosophila embryogenesis. In addition, he identified a regulatory link between energy and iron metabolism in vertebrates and invertebrates, and suggest biological functions for the IRE/IRP regulatory system in addition to the maintenance of iron homeostasis (Gray, Pantopoulos et al. 1996).

Developing of new IRE motifs identification methods.

In the moment we are developing new combined methods for biochemical and biocomputational identification of novel IRE-containing genes, responding to an iron.

To identify novel IRE-containing mRNAs, we integrated biochemical, biocomputational, and microarray-based experimental approaches: IRP/IRE messenger ribonucleoproteins were immunoselected, and their mRNA composition was analysed using an Affymetrix microarray.

High-throughput approach allows identifying more than 100 IRE-containing transcripts. Among them, the integrated experimental strategy reliably identifies known IRE-containing genes.

To complement microarray results with the structural information we used the IRE structure recognition tool written on Perl (sIREs- search for IREs) which is able to recognise 18 generalized IRE-like motifs. Each motif is a stem-loop structure with a characteristic C-bulge and 6th nucleotides-long unique IRE pattern. Each 19-mer was converted using Perl regular expressions to a cloud of possible motifs. This allows simulating in total ~$1.87*10^5$ unique IRE-like patterns (there are ~$2.75*10^{11}$ random patterns).

Using sIREs we confirmed that more than 60% of experimentally identified transcripts contain an IRE-like motif.

IronChip Evaluation Package (ICEP) – novel microarray analysis tool

We introduce a new flexible microarray analysis tool named ICEP, optimized for robust statistical analysis of specialized custom cDNA or oligonucleotide microarrays. There are many microarrays analysis methods available, one of the most flexible tool is a Bioconductor, based on statistical programming language R. The Bioconductor is a specialized statistical programming language, therefore it allows analysis of any type and difficulty but requires a deep knowledge of a programming language and methods.

ICEP in turn is very easy to use tool but it allows performing analysis of a broad range of two-colour microarrays. It could be used for standard analysis but as well to make use of specific design of a microarray. Specific design assumes high number if a positive and negative controls as well as a number of specifically designed replicates, representing single transcript on the array.

Our analysis yields values for ratio cut- off, background noise, multiple repetitions and de- tailed feature extraction as well as grouping rules. ICEP is easily extended to support further input and output data formats and different data transformation steps can be added. ICEP allows rapid post-processing of microarray data on a user-friendly platform. Software, example data and a tutorial are open source, free and available for downloading.

Challenges to understand iron network

The regulation and maintenance of systemic iron homeostasis is critical to human health. Iron overload and deficiency diseases belong to the most common nutrition-related pathologies across the globe. It is now well appreciated that the hormonal hepcidin/ferroportin system plays an important regulatory role for systemic iron metabolism.

Iron is essential for oxygen transport, cellular respiration, and DNA synthesis. Iron deficiency can cause cellular growth arrest and death. Conversely, iron excess and “free” reactive iron is toxic: Ferrous iron reacts with hydrogen peroxides or lipid peroxides to generate hydroxyl or lipid radicals, respectively.

These radicals damage lipid membranes, proteins, and nucleic acids. Since both iron deficiency and iron overload are detrimental to the cell, anomalies in iron metabolism are frequent causes of clinical disorders. For example, iron overload in hereditary hemochromatosis (HH) and the thalassemias leads to potentially fatal liver or heart failure. Iron deficiency represents the most common cause of anemia worldwide and can cause developmental retardation in children. Thus, iron homeostasis must be tightly controlled on both the systemic and the cellular levels to provide just the right amounts of iron at all times

Cellular iron homeostasis is achieved by the coordinated and balanced expression of proteins involved in iron uptake, export, storage, and utilization. Although genetic control is exerted at multiple steps, the posttranscriptional control mediated by the IRE/IRP system has emerged as central and essential (Hentze, Muckenthaler et al. 2004). Cellular iron metabolism is coordinately controlled by the binding of IRP1 or IRP2 to cis-regulatory mRNA motifs termed IREs; IRE/IRP interactions regulate the expression of the mRNAs encoding proteins for iron acquisition (transferrin receptor 1; TFRC-TFR1), divalent metal transporter 1 (SLC11A2/DMT1/DCT1/NRAMP2), storage (ferritin H, FTH1 and ferritin L, FTL), utilization (erythroid 5’-aminolevulinic acid synthase; mitochondrial aconitase, ACO2; Drosophila succinate dehydrogenase, SDH; hypoxia-inducible factor 2, HIF2/ Epas1), and export (SLC40A1/FPN1/IREG1/MTP1) (WE, Al et al. 2006; Sanchez, Galy et al. 2007). Past and recent evidence suggests that additional mRNAs are regulated by the IRPs (Kohler, Menotti et al. 1999; Cmejla, Petrak et al. 2006; Sanchez, Galy et al. 2006).

Microarrays-based techniques to improve a disease diagnosis

Since end of 90th when microarray technique became a standard lab practice like PCR, many attempts were performed to utilise microarrays for disease diagnosis. There are studies dealing with profiling of a different tumor specimens (Kononen, Bubendorf et al. 1998), attempts to diagnose autoimmune diseases (Joos, Schrenk et al. 2000) and many other (Yoo, Choi et al. 2009).

HFE, transferrin receptor 2 (TfR2), and hemojuvelin (HJ), 3 genes mutated in a group of frequent iron overload disorders called hereditary hemochromatosis (HH), control appropriate hepcidin expression. Analysis of patient biopsies samples hybridised on IronChip with a help of ICEP could reveal other characteristic genes, responsible for a HH disease. In a long-term it could lead to development of a novel HH diagnostic tool and improve a disease treatment strategy, using personalized drugs.

Conclusions

Applying microarray-based techniques to study patterns of gene expression is important to understand how cells adapt and change in time.

As a key stimulus we investigate here the response to iron. We provide different biocomputing methods to study this: In the first results chapter we detect new iron responsive elements using PERL scripts and the ICEP chip array analysis program. The second chapter details the performance, implementation and use of the ICEP program while the third chapter shows a number of gene expression studies where we used ICEP for pattern search and demonstrate its reliability (in controls) and its usefulness (discovery of new expression patterns).

Physiologically, iron regulation is far more intricate and detailed than previously estimated, the discovery of posttranscriptional regulation (as shown here applying ICEP) opens up new options for diagnosis, and maybe later, also treatment.

References

Beaumont, C., P. Leneuve, et al. (1995). "Mutation in the iron responsive element of the L ferritin mRNA in a family with dominant hyperferritinaemia and cataract." Nat Genet **11**(4): 444-446.

Benes, V. and M. Muckenthaler (2003). "Standardization of protocols in cDNA microarray analysis." Trends Biochem Sci **28**(5): 244-249.

Bengert, P. and T. Dandekar (2003). "A software tool-box for analysis of regulatory RNA elements." Nucleic Acids Res **31**(13): 3441-3445.

Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nat Genet **21**(1 Suppl): 33-37.

Chetverin, A. B. and F. R. Kramer (1994). "Oligonucleotide arrays: new concepts and possibilities." Biotechnology (N Y) **12**(11): 1093-1099.

Chu, S., J. DeRisi, et al. (1998). "The transcriptional program of sporulation in budding yeast." Science **282**(5389): 699-705.

Cmejla, R., J. Petrak, et al. (2006). "A novel iron responsive element in the 3'UTR of human MRCKalpha." Biochem. Biophys. Res. Commun. **341**: 158.

Cmejla, R., P. Ptackova, et al. (2010). "Human MRCKalpha is regulated by cellular iron levels and interferes with transferrin iron uptake." Biochem Biophys Res Commun.

Cox, T. C., M. J. Bawden, et al. (1991). "Human erythroid 5-aminolevulinate synthase: promoter analysis and identification of an iron-responsive element in the mRNA." EMBO J **10**(7): 1891-1902.

Dandekar, T., K. Beyer, et al. (1998). "Systematic genomic screening and analysis of mRNA in untranslated regions and mRNA precursors: combining experimental and computational approaches." Bioinformatics **14**(3): 271-278.

Dandekar, T., R. Stripecke, et al. (1991). "Identification of a novel iron-responsive element in murine and human erythroid delta-aminolevulinic acid synthase mRNA." EMBO J **10**(7): 1903-1909.

DeRisi, J., L. Penland, et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer." Nat Genet **14**(4): 457-460.

Dhiman, N., R. Bonilla, et al. (2002). "Gene expression microarrays: a 21s century tool for directed vaccine design." Vaccine **20**: 22-30.

Dudoit, S., R. C. Gentleman, et al. (2003). "Open source software for the analysis of microarray data." Biotechniques **Suppl**: 45-51.

Eickhoff, H., J. Schuchhardt, et al. (2000). "Tissue gene expression analysis using arrayed normalized cDNA libraries." Genome Res **10**(8): 1230-1240.

Eisenstein, R. S. and K. L. Ross (2003). "Novel roles for iron regulatory proteins in the adaptive response to iron deficiency." J Nutr **133**(5 Suppl 1): 1510S-1516S.

Farazi, T. A., G. Waksman, et al. (2001). "The biology and enzymology of protein N-myristoylation." J Biol Chem **276**(43): 39501-39504.

Galy, B., D. Ferring, et al. (2005). "Generation of conditional alleles of the murine Iron Regulatory Protein (IRP)-1 and -2 genes." Genesis **43**(4): 181-188.

Garrick, M. D., H. C. Kuo, et al. (2006). "Comparison of mammalian cell lines expressing distinct isoforms of divalent metal transporter 1 in a tetracycline-regulated fashion." Biochem J **398**(3): 539-546.

Gentleman, R., V. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Goossen, B. and M. W. Hentze (1992). "Position is the critical determinant for function of iron-responsive elements as translational regulators." Mol Cell Biol **12**(5): 1959-1966.

Gray, N., K. Pantopoulos, et al. (1996). "Translational regulation of mammalian and Drosophila citric acid cycle enzymes via iron-responsive elements." Proc Natl Acad Sci U S A **93**(10): 4925-4930.

Gress, T. M., J. D. Hoheisel, et al. (1992). "Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues." Mamm Genome **3**(11): 609-619.

Gunshin, H., C. R. Allerson, et al. (2001). "Iron-dependent regulation of the divalent metal ion transporter." FEBS Lett **509**(2): 309-316.

Hentze, M. W., S. W. Caughman, et al. (1987). "Identification of the iron-responsive element for the translational regulation of human ferritin mRNA." Science **238**(4833): 1570-1573.

Hentze, M. W., M. U. Muckenthaler, et al. (2004). "Balancing acts: molecular control of mammalian iron metabolism." Cell **117**(3): 285-297.

Herzel, H., D. Beule, et al. (2001). "Extracting information from cDNA arrays." Chaos **11**(1): 98-107.

Hubert, N. and M. W. Hentze (2002). "Previously uncharacterized isoforms of divalent metal transporter (DMT)-1: implications for regulation and cellular function." Proc Natl Acad Sci U S A **99**(19): 12345-12350.

Joos, T. O., M. Schrenk, et al. (2000). "A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics." Electrophoresis **21**(13): 2641-2650.

Kaiser, B. K., Z. A. Zimmerman, et al. (2002). "Disruption of centrosome structure, chromosome segregation, and cytokinesis by misexpression of human Cdc14A phosphatase." Mol Biol Cell **13**(7): 2289-2300.

Kato, J., K. Fujikawa, et al. (2001). "A mutation, in the iron-responsive element of H ferritin mRNA, causing autosomal dominant iron overload." Am J Hum Genet **69**(1): 191-197.

Kohler, S. A., E. Menotti, et al. (1999). "Molecular cloning of mouse glycolate oxidase. High evolutionary conservation and presence of an iron-responsive element-like sequence in the mRNA." J. Biol. Chem. **274**: 2401.

Kononen, J., L. Bubendorf, et al. (1998). "Tissue microarrays for high-throughput molecular profiling of tumor specimens." Nat Med **4**(7): 844-847.

Kurian, K. M., C. J. Watson, et al. (1999). "DNA chip technology." J Pathol **187**(3): 267–271.

Lam-Yuk-Tseung, S. and P. Gros (2006). "Distinct targeting and recycling properties of two isoforms of the iron transporter DMT1 (NRAMP2, Slc11A2)." Biochemistry **45**(7): 2294-2301.

Lashkari, D. A., J. L. DeRisi, et al. (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proc Natl Acad Sci U S A **94**(24): 13057-13062.

Lipshutz, R. J., S. P. Fodor, et al. (1999). "High density synthetic oligonucleotide arrays." Nat Genet **21**(1 Suppl): 20-24.

Lis, A., P. N. Paradkar, et al. (2005). "Hypoxia induces changes in expression of isoforms of the divalent metal transporter (DMT1) in rat pheochromocytoma (PC12) cells." Biochem Pharmacol **69**(11): 1647-1655.

Lymboussaki, A., E. Pignatti, et al. (2003). "The role of the iron responsive element in the control of ferroportin1/IREG1/MTP1 gene expression." J Hepatol **39**(5): 710-715.

Mailand, N., C. Lukas, et al. (2002). "Deregulated human Cdc14A phosphatase disrupts centrosome separation and chromosome segregation." Nat Cell Biol **4**(4): 317-322.

McKie, A., P. Marciani, et al. (2000). "A novel duodenal iron-regulated transporter, IREG1, implicated in the basolateral transfer of iron to the circulation." Mol Cell **5**(2): 299-309.

Melefors, O. (1996). "Translational regulation in vivo of the Drosophila melanogaster mRNA encoding succinate dehydrogenase iron protein via iron responsive elements." Biochem Biophys Res Commun **221**(2): 437-441.

Meyron-Holtz, E. G., M. C. Ghosh, et al. (2004). "Genetic ablations of iron regulatory proteins 1 and 2 reveal why iron regulatory protein 2 dominates iron homeostasis." EMBO J **23**(2): 386-395.

Mignone, F., G. Grillo, et al. (2005). "UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs." Nucleic Acids Res **33**(Database issue): D141-146.

Muckenthaler, M., B. Galy, et al. (2008). "Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network." Annu Rev Nutr **28**: 197-213.

Muckenthaler, M., N. K. Gray, et al. (1998). "IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F." Mol Cell **2**(3): 383-388.

Muckenthaler, M., A. Richter, et al. (2003). "Relationships and distinctions in iron-regulatory networks responding to interrelated signals." Blood **101**(9): 3690-3698.

Muckenthaler, M., C. Roy, et al. (2003). "Regulatory defects in liver and intestine implicate abnormal hepcidin and Cybrd1 expression in mouse hemochromatosis." Nat Genet **34**(1): 102-107.

Poustka, A., T. Pohl, et al. (1986). "Molecular approaches to mammalian genetics." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 131-139.

Richter, A., C. Schwager, et al. (2002). "Comparison of fluorescent tag DNA labeling methods used for expression analysis by DNA microarrays." Biotechniques **33**(3): 620-628, 630.

Ross, D. T., U. Scherf, et al. (2000). "Systematic variation in gene expression patterns in human cancer cell lines." Nat Genet **24**(3): 227-235.

Rouault, T. A. (2001). "Iron on the brain." Nat Genet **28**(4): 299-300.

Sanchez, M., B. Galy, et al. (2006). "Iron regulation and the cell cycle: identification of an iron-responsive element in the 3'-untranslated region of human cell division cycle 14A mRNA by a refined microarray-based screening strategy." J Biol Chem **281**(32): 22865-22874.

Sanchez, M., B. Galy, et al. (2007). "Identification of target mRNAs of regulatory RNA-binding proteins using mRNP immunopurification and microarrays." Nat Protoc **2**(8): 2033-2042.

Sanchez, M., B. Galy, et al. (2007). "Iron-regulatory proteins limit hypoxia-inducible factor-2alpha expression in iron deficiency." Nat Struct Mol Biol **14**(5): 420-426.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-470.

Schena, M., D. Shalon, et al. (1996). "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes." Proc Natl Acad Sci U S A **93**(20): 10614-10619.

Schneider, B. D. and E. A. Leibold (2000). "Regulation of mammalian iron homeostasis." Curr Opin Clin Nutr Metab Care **3**(4): 267-273.

Schuchhardt, J., D. Beule, et al. (2000). "Normalization strategies for cDNA microarrays." Nucleic Acids Res **28**(10): E47.

Schwager, C. (2002). "ChipSkipper."

Smith, S. R., M. C. Ghosh, et al. (2006). "Complete loss of iron regulatory proteins 1 and 2 prevents viability of murine zygotes beyond the blastocyst stage of embryonic development." Blood Cells Mol Dis **36**(2): 283-287.

Southern, E. M. (1974). "An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose." Anal Biochem **62**(1): 317-318.

Southern, E. M. (1975). "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J Mol Biol **98**(3): 503-517.

Southern, E. M. (1995). "DNA fingerprinting by hybridisation to oligonucleotide arrays." Electrophoresis **16**(9): 1539-1542.

Southern, E. M. and A. R. Mitchell (1971). "Chromatography of nucleic acid digests on thin layers of cellulose impregnated with polyethyleneimine." Biochem J **123**(4): 613-617.

Staudt, L. and P. O. Brown (2000). "Genomic views of the immune system." Annual Review of Immunology **18**: 829–859.

Tabuchi, M., N. Tanaka, et al. (2002). "Alternative splicing regulates the subcellular localization of divalent metal transporter 1 isoforms." Mol Biol Cell **13**(12): 4371-4387.

Theil, E. C., R. A. McKenzie, et al. (1994). "Structure and function of IREs, the noncoding mRNA sequences regulating synthesis of ferritin, transferrin receptor and (erythroid) 5-aminolevulinate synthase." Adv Exp Med Biol **356**: 111-118.

Vainshtein, Y., M. Sanchez, et al. (2010). "The IronChip evaluation package: a package of perl modules for robust analysis of custom microarrays." BMC Bioinformatics **11**: 112.

Walden, W. E., A. I. Selezneva, et al. (2006). "Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA." Science **314**: 1903.

Wang, X., M. D. Garrick, et al. (2005). "TNF, IFN-gamma, and endotoxin increase expression of DMT1 in bronchial epithelial cells." Am J Physiol Lung Cell Mol Physiol **289**(1): L24-33.

Wong, A. K., Y. Chen, et al. (1999). "Genomic structure, chromosomal location, and mutation analysis of the human CDC14A gene." Genomics **59**(2): 248-251.

Yoo, S. M., J. H. Choi, et al. (2009). "Applications of DNA microarray in disease diagnostics." J Microbiol Biotechnol **19**(7): 635-646.

Zoller, H., I. Theurl, et al. (2002). "Mechanisms of iron mediated regulation of the duodenal iron transporters divalent metal transporter 1 and ferroportin 1." Blood Cells Mol Dis **29**(3): 488-497.

Abbreviations used

Cy3/Cy5 - reactive water-soluble fluorescent dyes of the cyanine dye family. Cy3 dyes are yellow-orange (550 nm excitation, 570 nm emission), while Cy5 is fluorescent in the red region (650/670 nm).

EST - Expressed Sequence Tags -ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene

GUI - Graphical User Interface
ICEP - IronChip Evaluation Package
IRE - iron-responsive elements
IRP - Iron regulatory protein
PDK - Perl developer kit
RNPs - ribonucleoparticles
UTR - untranslated region

Gene names

TFRC - Transferrin receptor 1
MT2A - Metallothionein 2
EPAS1 - Endothelial PAS domain-containing protein 1; Hypoxia-inducible factor 2 alpha (HIF-2 alpha)
SLC11A2 - Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2
NRAMP2 - Natural resistance-associated macrophage protein 2 (NRAMP2)
DMT1 - Divalent metal transporter 1
ACTB - Actin, beta, cytoplasmic
ALAS2 or eALAS - Aminolevulinic acid synthase 2, erythroid
FTL - Ferritin light chain; Ferritin L subunit
HSPH1 - Heat shock 105kDa/110kDa protein 1
HMOX1 - Heme oxygenase (decycling) 1
HSPA1L - Heat shock 70kDa protein 1-like
HSPA1A - Heat shock 70kDa protein 1A
FTH1 - ferritin heavy polypeptide 1
FTL - ferritin light polypeptide
CDC14A - cell division cycle 14 homolog A
DCT1 - Divalent Cation Transporter 1

Supplementary materials

List of Publications

- **Vainshtein Y*,** Sanchez M, Brazma A, Hentze MW, Dandekar T*, Muckenthaler MU*. The IronChip evaluation package: a package of perl modules for robust analysis of custom microarrays. BMC Bioinformatics. 2010 Mar 1;11:112.
- Sanchez M, Galy B, Dandekar T, Bengert P, **Vainshtein Y**, Stolte J, Muckenthaler MU, Hentze MW. Iron regulation and the cell cycle: identification of an iron-responsive element in the 3'-untranslated region of human cell division cycle 14A mRNA by a refined microarray-based screening strategy. J Biol Chem. 2006 Aug 11;281(32):22865-7
- Gunshin, H; Tohyama, C; **Vainshtein, Y**; Hentze, MW; Muckenthaler, M; Andrews, NC. Metal distribution and gene expression profile in DMT1 knockout mice. American Journal of Hematology June 2007 82(6)

Conferences presentations

- **Vainshtein Y. "**Introduction into Biostatistics".
  Special lectures on statistics for biologists. EMBL, Heidelberg, Germany; 2006 – presented
- **Vainshtein Y,** Sanchez M, Brazma A, Muckenthaler MU, Hentze MW and Reich J. "Systems Biology of Mammalian Iron Metabolism: Biocomputational Characterization of Iron Regulatory Networks"
  First Congress of the International BioIron Society Prague, Chech Republic; 2005 – oral presentation and poster
- **Vainshtein Y**, Horgan G.W., Avery P., Pfaffl M.W. "New REST-384 © : the way of simple and precise qRT-PCR data evaluation"
  qPCR Satellite Symposium & qPCR Workshop Leipzig, Germany; 2005 – oral presentation
- **Vainshtein Y,** et al. "Systems Biology of Mammalian Iron Metabolism: Biocomputational Characterization of Iron Regulatory Networks"
  EMBO Practical Course on Analysis and Informatics of Microarray Data, EBI, Hinxton, UK, 2005 – oral presentation
- **Vainshtein Y**, Muckenthaler MU, Brazma A and Hentze MW "First steps towards a multi-dimensional iron regulatory network"
  1st FEBS Advanced Lecture Course on Systems Biology Gosau, Austria; 2005 – poster
- Vainshtein Y, et al. "Expression analysis of Host-Cell response against Mycobacterium Tuberculosis"
  EUConference. Coord.Immunity; Greifswald , Germany; 2002 –poster

Curriculum vitae

Personal Data

Yevhen Vainshtein

Date of Birth: 29.03.1977
Place of Birth: Kiew, Ukraine
Family status: Maried since 10.2002
Nationality: Ukraine

Educational Background

12/2008–07/2010	*Ph.D. student,* **Department of Bioinformatics**, Biozentrum, Am Hubland, University of Würzburg “Applying microarray-based techniques to study gene expression patterns: a bio-computational approach”
12/2008–02/2009	*Visiting Scientist*, **Director’s Research Unit**, European Molecular Biology Laboratory, Heidelberg, Germany.
02/2008–12/2008	*Research Fellow,* **Molecular Medicine Partnership Unit,** Department of Pediatric Oncology, Haematology, Immunology and Pulmonology, Children’s Hospital, University of Heidelberg, Germany
12/2003–01/2008	*Research Fellow,* **Director’s Research Unit**, European Molecular Biology Laboratory, Heidelberg, Germany
9/2000–9/2003	*Research Fellow,* **Department of Immunology,** Max-Plank Institute of Infection Biology
1998, 2000	**BSc/MSc. Diploma** on subject: *“Physical and chemical properties of bacteriophage T4 tail sheath protein gp18“* with a grade “very good”. Average grade is “good”
1997–2000	**Laboratory of Molecular Bioengineering,** Institute of Bioorganic Chemistry, Moscow, Russia
1994--2000	Study on Applied a faculty of molecular biophysics, Moscow Institute of Physics and Technology (MIPT), Moscow, Russia
1990-1994	Secondary school Nr. 142 Physical-Mathematical Lyceum, Kiev, Ukraine

1987-1990	Physical-Mathematical secondary school Nr. 100, Kiev, Ukraine
1984-1987	Primary school Nr. 244, Kiev. Ukraine

Acknowledgements

On this place I would like to thank all people who was helping and supporting me during my thesis project. This work could not be done without generous help of many people. All my colleagues from Gene Expression Unit at EMBL and Molecular Medicine Partnership Unit at University of Heidelberg contributed to the success of this project. I am very grateful to all of you.

In particular I would like to acknowledge following people:

I want to thank first of all Prof. Matthias Hentze (associate director, EMBL) for introducing me to this exciting topic. Without him and his support it wouldn’t be possible for me to work on this project. I wish to express my gratitude to Prof. Martina Muckenthaler for her continuous and helpful supervision during all experiments in Heidelberg. We had many very fruitful discussions which helps me to implement our ideas to a novel application.

The same thanks go to Prof. Thomas Dandekar for thesis supervision here in Würzburg and regarding writing and program testing.

Finally, Prof. Jens Reich (MDC Berlin) was a constant and attentive listener in a Socratic way which helped to pave out challenging results.

From my colleagues on the first place I would like to thank Dr. Mayka Sanchez. She is the person who was always ready to help and open for discussions. During the time of this project, I was glad to have many helpful and motivating discussions with colleagues who became friends.