BIOINFORMATICS AND BIOMOLECULAR TOOLS FOR BIOMARKER DISCOVERY IN PERIPHERAL BLOOD LYMPHOCYTES FROM PATIENTS WITH SPORADIC AMYPTROPHIC LATERAL SCLEROSIS

by

Cristina Baciu

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2012

Approved by:

____
Dr. Jennifer W. Weller

____
Dr. Jean-Luc Mougeot

____
Dr. Susan Sell

____
Dr. Ann Loraine

____
Dr. Jun-Tao Guo

©2012
Cristina Baciu
ALL RIGHTS RESERVED

ABSTRACT

CRISTINA BACIU. Bioinformatics and biomolecular tools for biomarker discovery in peripheral blood lymphocytes from patients with Sporadic Amyotrophic Lateral Sclerosis. (Under the direction of DR. JENNIFER W. WELLER)

Sporadic Amyotrophic lateral sclerosis (sALS) is a complex, invariably fatal, disease with a poorly understood cause, despite many studies. Diagnostic biomarkers that precede active symptoms would be an immense help to clinicians, for patient management, following the progress of clinical studies, and uncovering early events in the development and progression of the disease.

Combining bioinformatics of microarrays and molecular biology assays we analyzed and extended the results from experiments performed on peripheral blood lymphocyte (PBL) fractions from an sALS and a normal-matched coronary artery disease (CAD) study. We developed a novel computational pipeline (LO-BaFL) to improve the power and discrimination of identifying differentially expressed (DE) genes on long-oligonucleotide arrays. From sALS samples we performed quantitative polymerase chain reaction (qPCR) validation assays that linked three novel genes, ACTG1, B2M, and ILKAP, to sALS. Selected regions of the DE transcripts were sequenced, which revealed a new, albeit non ALS-linked mutation. Genes revealed as DE by LO-BaFL were examined through pathway and network interaction analysis. Heightened profiles are seen in the immune response signature, apoptosis and responses to chemical stimulus; these correspond well to phenotypes associated with sALS and are good candidates for a simplified blood-based biomarker signature.

DEDICATION

I dedicate this work to my family and to all who are touched in any way by ALS.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jennifer W. Weller for her guidance, in-depth knowledge and for helping me to finalize this extraordinary journey. Throughout the years, her passion for science, optimism and ability to manage, at times, the non-manageable was both educational and inspiring.

I thank also to Carolinas Neuromuscular/ALS-MDA Center, Neuroscience and Spine Institute, Carolinas Medical Center, Charlotte, in particular to my co-advisor, Dr. Jean Luc Mougeot for providing the microarray data and samples, for his help and support with the research and proof reading the manuscript.

I thank my committee members for their patience to listen to my talks and for their pertinent and useful advices. Thank you to the Bioinformatics and Genomics department for guidance and financial assistance, to Weller group in particular for their encouragement and help. Thank you to the Bioinformatics Research Center at North Carolina Research Campus in Kannapolis for granting access to software and expertise. Thank you to International Office staff for smoothing the transition when moving abroad. I also give my thanks to my first year advisor, Dr. Andrij Baumketner for encouraging me to enroll into Bioinformatics PhD program and for financial aid.

At last, but not at least, I thank to my family, especially to my husband and my son for their patience and unconditional support, but also to my parents and my sister from overseas for having tremendous confidence in me.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

LIST OF ABBREVIATIONS x

CHAPTER 1: THE LO-BaFL PIPELINE FOR MICROARRAY EXPRESSION ANALYSIS 1

- 1.1 Introduction 1
- 1.2 Materials and Methods 8
- 1.3 Results and Discussions 14
- 1.4 Conclusions 29

CHAPTER 2: VALIDATION OF DIFFERENTIALLY EXPRESSED GENES ASSOCIATED WITH sALS BY qRT-PCR ASSAYS 31

- 2.1 Introduction 31
- 2.2 Materials and Methods 32
- 2.3 Results and Discussions 40
- 2.4 Conclusions 40

CHAPTER 3: SEARCHING FOR SEQUENCE VARIANTS ASSOCIATED WITH DIFFERENTIALLY EXPRESSED GENES DETERMINED BY LO-BaFL AND COMPARATIVE METHODS 43

- 3.1 Introduction 43
- 3.2 Materials and Methods 45
- 3.3 Results and Discussions 46
- 3.4 Conclusions 61

CHAPTER 4: PATHWAY ANALYSIS 63

- 4.1 Introduction 63
- 4.2 Materials and Methods 66
- 4.3 Results and Discussions 67
- 4.4 Conclusions 89

SUMMARY 92

REFERENCES 97

APENDIX A: ADDITIONAL NETWORK INTERACTIONS 111

VITA 115

LIST OF TABLES

TABLE 1.1: The % probes removed in total by applying the filtering steps of LO-BaFL pipeline and comparison with percent probes eliminated in the original BaFL pipeline. 18

TABLE 1.2: ALS experiment: Selected differentially expressed genes with $p < 0.05$, determined by LO-BaFL-Wilcoxon, TM4-W and SAM-W. 27

TABLE 2.1: The list of reference genes and DE genes as determined by qRT-PCR and their corresponding designed primers. 35

TABLE 2.2: Relative expression ratio for DE genes. 39

TABLE 3.1: Selected genes with known SNPs for sequencing and their dbSNP IDs. 45

TABLE 3.2: PCR primers for sequencing purposes. 47

TABLE 3.3: Allelic variants for *ACTG1*, from OMIM database. 62

TABLE 4.1: List of biological networks formed by selection of candidate genes. 70

TABLE 4.2: Transcription factors networks for current selection of genes. 80

LIST OF FIGURES

FIGURE 1.1: Flowchart of the LO-BaFL pipeline. 10

FIGURE 1.2: The LO-BaFL pipeline and the TM4 and SAM pipelines. 15

FIGURE 1.3: Graphical representation of $\Delta G_{cut\text{-}off}$ and ΔG vs. Probe signal. 21

FIGURE 1.4: Q-Q plot distribution of diseased and healthy controls. 24

FIGURE 1.5: Correlation between healthy controls in ALS and CAD studies. 29

FIGURE 2.1: Schematic diagram of the experimental design. 36

FIGURE 2.2: qRT-PCR standard curves for selected genes. 41

FIGURE 3.1: The flow chart for sequencing assay. 48

FIGURE 3.2: Electropherograms showing the presence of a novel mutation. 52

FIGURE 4.1: GenoGo map of networks with known interactions and Symbol legend. 72

FIGURE 4.2: MetaCore analysis of signaling pathways in PBLs from sALS samples. 73

FIGURE 4.3: GeneGo maps of transcription factors interaction networks. 82

FIGURE A1: Other relevant networks for our selection of genes. 111

LIST OF ABBREVIATIONS

ALS	Amyotrophic Lateral Sclerosis
CAD	Coronary Artery Disease
cDNA	complementary Deoxyribonucleic acid
DE	Differential Expression
DNA	Deoxyribonucleic acid
fALS	familial Amyotrophic Lateral Sclerosis
gDNA	genomic Deoxyribonucleic acid
LO-BaFL	Long Oligonucleotides Biologically Applied Filters
qRT-PCR	quantitative Real Time - Polymerization Chain Reaction
PBLs	Peripheral Blood Lymphocytes
PBMC	Peripheral Blood Mononulcear Cells
RNA	Ribonucleic acid
sALS	sporadic Amyotrophic Lateral Sclerosis
SNP	Single Nucleotide Polymorphism

CHAPTER 1: THE LO-BaFL PIPELINE FOR MICROARRAY EXPRESSION ANALYSIS

1.1 Introduction

Biomolecular component: Biology and biomarker discovery in ALS

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is one the most well known of the motor neuron diseases, being described for the first time in 1869 [1]. It is characterized by the progressive degeneration of upper (in the brain) and lower (in spinal cord) motor neurons that in turn alters the muscle normal functions, causing muscle weakness and atrophy that ultimately leads to death, within 1.5-5 year following diagnosis [1-3]. ALS usually affects people in their 50s, with an incidence of 1-3 cases in 100,000/year. About 5 to 10 % of cases are familial ALS (fALS), caused mainly by autosomal dominant genetic mutations, the remaining 90 to 95% being sporadic (sALS), with an etiology still to be determined [1, 4, 5].

Extensive research in the pathogenesis of fALS has been stimulated by the discovery of mutations in Cu/Zn superoxide dismutase 1 (SOD1) that are responsible for approximately 20% of fALS cases. This fact is substantiated by the demonstration that SOD1 mutations in mouse models reproduce a motor neuron disease phenotype [6, 7]. Recent studies have shown that mutations in two additional genes are associated with premature degeneration of the motor neurons in both forms of ALS: the TARDBP at the ALS10 locus on chromosome 1, and the FUS/TLS gene at the ALS6 locus on chromosome 16 [2, 8-13]. Mutations in these genes determine, however, independent neurodegeneration events in patients with amyotrophic lateral sclerosis, as FUS/TLS mutations are not present in fALS patients with *SOD1* or *TARDBP* mutations and *vice versa* [2].

In recent years, the development of high-throughput and targeted sequencing and sequence interrogation methods has increased, at a very large scale, the number of available human genome sequences for molecular analysis. Several whole-genome association studies (WGAS) have been conducted in order to discover single nucleotide polymorphisms, SNPs, conferring susceptibility to sALS [13-15]. While in some studies weak associations were found, there were no overlapping results and two of the studies could not find any significant association of SNPs with ALS [14, 15]. The other studies associated independent, distinct SNPs with sALS, corresponding to the *FLJ10986, ITPR2* and *DPP6* genes, respectively [16-18]. Only the association with DPP6 has been successfully replicated (in an Italian population) as shown in one study, but not in a second pool of samples (from Poland) [19, 20]. In addition, two studies that determined copy number variants (CNVs) associated with sALS in geographically different populations have been conducted. Again, correlated mutations were not replicated between studies [21, 22]. Therefore, either very large studies or different targets will be needed in order to clearly demonstrate the link between specific loci and sALS.

There are several proposed mechanisms associated with motor neuron degeneration in ALS: oxidative stress, toxicity induced by mutant SOD1 through different cellular processes, formation of intracellular aggregates (which are sometimes observed), mitochondrial abnormalities, deficiency in axonal transport, apoptosis, and others [1, 23, 24]. Many of these processes are found in sALS, but are also common to neurodegenerative diseases in general such as Alzheimer's, Parkinson's, Fronto-Temporal Dementia or prion diseases.

Many biological and genetic studies have been conducted on biofluids or biopsies from sALS patients, but these studies have not yet led to the identification of a common aberrant process. Riluzole^{TM}, a drug that inhibits glutamate uptake by CNS neurons, is the only approved FDA drug for ALS treatment. It extends the life of ALS patients by only 2-3 months, on average. When ALS is diagnosed earlier, treatment with Riluzole is generally more effective. Therefore, early diagnosis biomarkers such as differentially expressed genes that can be measured by assays performed on drawn blood are sought for quality of life issues alone, although a cure is obviously the eventual goal. It is hoped that early biomarkers may provide new insights into causal agents or pathways involved in degenerative mechanisms that can be potentially exploited for drug target discovery. Early detection might help us capture the mechanisms of initiation and early progression in which processes are reversible, or holding patients in pre-symptomatic stages as some of the multiple sclerosis treatments appear to do [25]. Even without understanding the causal event, markers of disease progression are needed to study responses to new drugs and facilitate clinical trials.

Bioinformatics component

Oligonucleotide microarrays are extensively used for genomics studies (particularly transcriptomics and genotyping). The platforms were the first fully parallel instruments for assessing cell state, and remain powerful tools for pursuing biological mechanisms in the context of their full complexity, i.e., covariation in gene expression levels, detection of both alleles and haploblocks for SNPs and CNVs in genotyping, etc.[26-29]. However, despite their wide-spread use and frequent success, the correct handling of the measurements is still subject to debate, and conflicting interpretations are common [26]. Many factors contribute to the controversy. An individual's divergence from the 'reference standard' used in platform design is one factor [5, 6, 7], whose impact will become clearer as more genomes and variants are described [30]. Biophysical properties of the sensors are also important factors [27, 28]. Alternate transcript forms are a variable for eukaryotic genomes [29]; probes are unlikely to report on all variants. Noise has both biological and technical sources, including factors such as availability of a homogeneous sample and the completeness of amplification and fragmentation steps [31]. The effect of these factors on measurements is amenable to description and modeling: doing so improves the processing of the data [26].

In developing the data cleansing pipeline presented here, we considered those factors that can be identified with respect to a reference genome and databases of common variants, as well as biophysical factors for the most prevalent of the long-oligonucleotide arrays used to produce public datasets, the Agilent human 4x44k platform. The pipeline logically resembles the BaFL pipeline that was developed for short oligonucleotide probes [32], but differs in the particulars because the Agilent platform uses longer probes (60-mers) and has less intentional redundancy, with 1-2 probes per gene compared to Affymetrix arrays 11-16 probes (25-mers) per gene. Accommodating these differences requires modifying parameters in the algorithms and tests used to identify and map probes to the genome, since the length of a duplex affects it's stability under given hybridization conditions. For example, SNPs affect a measurement when they lie in the probe-target duplex, but the number required to eliminate the signal is correlated to the length of the duplex [33-35]. Similarly, internal probe or target structures compete with duplex, usually lowering the signal [36-38]; G runs (> 3) are a well-known special case [14-16]. Confirming the target requires remapping the probe to its genomic context, outcomes of which include identifying: (i) cross-hybridization to additional distant genomic locations; (ii) loss of binding site, where no stable complement exists; (iii) mis-location, requiring reassignment of the probe to a new gene (re-annotation); (iv) confirming correct, unique matches to the intended target. Where a sequence-based problem is identified a probe's measurement should be removed from all samples – this is most simply handled by altering the file describing the array layout e.g. with Aroma [39-41].

Not all error comes from sequence bias, sample handling and scanners also contribute. The instrument has important response characteristics to consider [42, 43], and the upper and lower limits of signal detection must be adjusted by experiment and platform. Variance from sample handling steps is examined after problematic probes have been removed: the pipeline incorporates several statistical tests to determine sample membership in the designated classes [32].

For rare, sporadic diseases such as ALS, it is difficult to obtain large sample sizes, therefore for statistical rigor and to make sure that a consistent effect is identified, meta-experiments are needed. The challenge is to identify samples that can legitimately be grouped and then to process the data in such a way that responses are similarly scaled. Only high quality sample annotation can ensure the first criterion, while removing probes known to have flaws and observing scanner response limitations helps with the second. Since all DE predictions require a robust normal control, and the sample size in our original study was very small [44], we obtained the CAD study [45] for its normal samples, whose age, gender and cell mixture characteristics were well-matched to our samples, as an independent control of the quality of our ALS normal cohort.

The effectiveness of a data processing pipeline is generally assessed by the accuracy of subsequent data mining efforts, which at the lowest level are tests for differential expression across states [46]. The most accepted confirmatory tests are sample-based, using an independent assay method (usually qRT-PCR), but may be meta-analysis based when samples are unavailable, using literature reports to reinforce the analysis findings. After processing data with both a standard pipeline, TM4 [47], and our LO-BaFL pipeline, we used SAM [24] to generate the DE predictions upon which effective processing is judged. Since a small amount of the ALS material amplified for the microarrays was available, some predictions were tested by qRT-PCR assays. To provide meta-analysis support, and because the ALS sample numbers were so small, the CAD normal sample analysis was added. We note that in the original CAD study qRT-PCR was used to test some of the predictions, and we have assumed that the reported results were accurate. Microarray and qRT-PCR results were declared concordant when the direction and the degree of change in expression compared to a control gene were accurately captured [48]. Finally, we performed a literature search for independent reports on a number of the genes, or pathway and interaction data predicted and confirmed to be important in these ALS samples [49].

We mention above that a 'standard' pipeline is used as well as the one we developed. As an open source for microarray data analysis, TM4 [47] consists of a series of applications under a graphical interface that facilitates analyses of microarray data. Among the TM4 suite of tools, MIDAS (Microarray Data Analysis System) includes several normalization steps (e.g., total intensity normalization, Lowess normalization, standard deviation regularization), and filtering to remove low intensity signals. There are several options for statistical analysis on filtered data to determine differentially expressed genes, i.e. parametric versus non-parametric tests. The TM4 pipeline uses statistical rather than biophysical criteria to remove poor measurements [47, 50] and it does not explicitly list the deprecated probes, so understanding directly what response changes have lead to different outcomes is not possible. To test whether the LO-BaFL processing pipeline has advantages over TM4 when array studies using small sample sizes are involved, we used each pipeline to process two independent data sets. We then used a significance test for DE genes, applying a simple Wilcoxon non-parametric test because the distributions did not meet criteria to use a parametric test [28]. For the ALS samples, we performed the qRT-PCR assays on the unused portion of the products sent for array hybridization [51] while for the subset of samples from the CAD experiment we relied on the published qRT-PCR results and literature references.

1.2 Materials and methods

For data storage, data organization, and recording the order and parameters used in the pipeline transformations, we have used DataFATE (Data - Feature Analysis Transformation Extraction), a software system based on a relational model that includes a toolset with data import and organization tools for relational database management systems (RDBMS), tools for factor (quantitation type, QT) definition, QT set construction, and storage of data from processing steps. The RDBMS is currently PostgreSQL 8.0.3. [52]. The project instance of DataFATE was installed into a 64 bit, 22-processor, 120 GB of RAM computer running Ubuntu 9.04 version for Kernel LINUX™ 2.6.28, as the operating system. Querying, extraction and manipulation of data stored in DataFATE has been made with scripts written with Python 2.6 [53], SQL (via PGAdminIII) [54] and R [55]. Additional software installed on this hardware and used for this project includes TM4 microarray software suite [56], and OligoArrayAux [29] for biophysical modeling. For the results using the packages TM4 [47] and Significance Analysis of Microarrays (SAM) [57], we set up the relational database in order to maintain stable output of intermediate and final results of both pipelines.

Data acquisition

Microarray image files and corresponding spot intensity values for the ALS study were provided by Carolinas Neuromuscular/ALS-MDA Center, Neuroscience and Spine

Institute, Carolinas Medical Center, Charlotte, NC. The microarray experiment used Agilent 4x44K human genome microarrays [58] in a pooled reference design [59]. Sample and microarray processing were performed at Cogenics [60], producing arrays contrasting each sample (healthy and diseased) to the healthy reference pool. The raw data sent back by Cogenics includes extracted spot intensities and the background-subtracted intensity ratios for each contrast.

CAD raw data was downloaded from GEO, Accession No.GSE10195.

The LO-BaFL pipeline

The steps in the pipeline, described below, are summarized in Figure 1.1.

A. In this section the probe-sequence based filters are described.

(i) Re-map the Agilent probes to assembly version 36.1 of the human genome (36.1) using the accelerated Tera-BLAST algorithm, as implemented by a TimeLogic-Decypher [61] server. The corresponding matches were deposited into an instance of the DataFATE database. Parameters were: nucleic match = 1; nucleic mismatch = -3; open penalty = -5; extend penalty = -2; threshold significance = 10. The input and output files can be found in Supplementary Material section, at: *http://webpages.uncc.edu/~cbaciu/LO-BaFL/supplementary_data.html* under Input Files/agilent_fasta or Cleansing Process/tera_blast_results.

(ii) Determine the cross-hybridization potential of probes to other sites in the genome, using the Kane criteria [62]. Briefly this is an empirical rule stating that any target sequence with similarity greater than 75% across the length of a probe can contribute a detectable amount of signal to the total intensity.

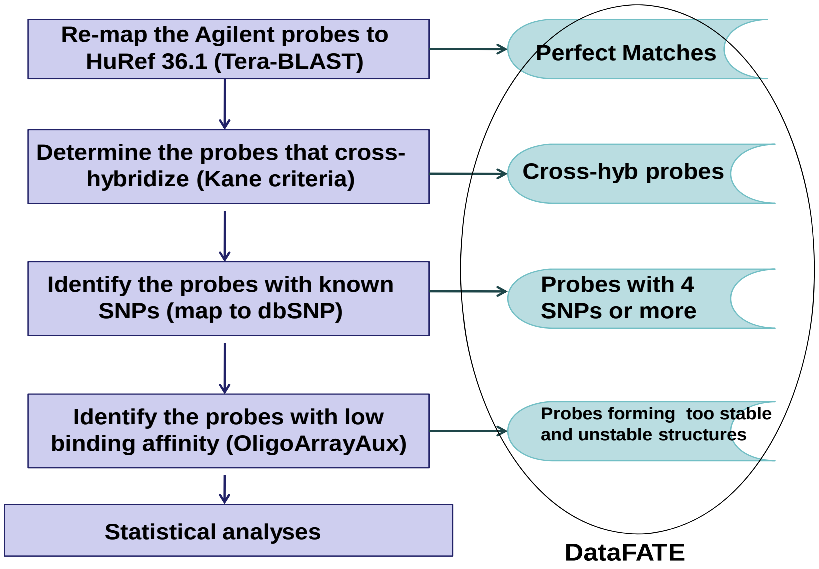

Re-map the Agilent probes to HuRef 36.1 (Tera-BLAST)
Determine the probes that cross-hybridize (Kane criteria)
Identify the probes with known SNPs (map to dbSNP)
Identify the probes with low binding affinity (OligoArrayAux)
Statistical analyses
Perfect Matches
Cross-hyb probes
Probes with 4 SNPs or more
Probes forming too stable and unstable structures
DataFATE

Figure 1.1 Flowchart of the LO-BaFL pipeline: the pipeline step is given on the left and the right indicates where intermediate datasets were stored in the project database. Note that this output has been made available as flat files.

This rule includes some constraints concerning the positions and lengths of mismatch regions. For a probe to cross-hybridize, we input the following conditions: percent identity ≥ 85%; presence of 50 matches out of 60 possible; minimum of 15 consecutive nucleotides in the Agilent probe sequence. We stored the output, consisting of all the Kane-criteria cross-hybridizing probes into DataFATE.

(iii) Identify probes that no longer anchor to the reference genome. This information is acquired when a TeraProbe query returns ‘no hit’, and this is stored as an explicit type.

(iv) Identify SNPs and short indels known to occur in the probe-binding region. Probes were mapped to the human instance of dbSNP [63], taking all possible alternate alleles into consideration. The minimum number of SNPs expected to significantly degrade the signal is a parameter in the BaFL pipeline. Using the Kane criteria, the presence of six SNPs will reduce the signal to the point of background, but the presence of any SNP will cause the signal to reflect both sequence variation and transcript concentration and the question of degree is not simple since it depends on sequence context and competition. For the case study we set the 'deprecate' flag to 3 SNPs or more, assuming that this many competing alleles would make the intensity information useless for differential expression analysis. All of the information was retained, however, so another researcher could modify the query to adjust the number of SNPs to allow in retrieved probes.

(v) Employ the OligoArrayAux software [29] to determine the free energy of internal probe structures versus heterodimers. Parameters chosen were: temperature 55 to 62º C, concentrations of 1.0 M Na^+ and 0.0 M Mg^{++}, output was used to define probes inaccessible to target (except at very high concentrations) under experimental conditions. Probes that predominantly form very stable internal structures have a lower effective concentration, and so bind less target. Heterodimers with low stability under specific experimental conditions do not yield signal [64-66]. The predicted value of the most stable form is stored in the database as an attribute of the probe, allowing adjustment of the cut-off value.

(vi) It has been shown for Affymetrix arrays that four or more consecutive guanines (G-runs) lead to unusual probe structures that cause very high signal [14-16]. We identified the probes with this feature; we note that most of them are removed based on other filters as well (data not shown).

(vii) The presence of any member of the transposable elements family, short (SINE), long (LINE) or primate-specific (Alu) repeat elements, can have a great affect on gene expression [45-47]. Using the TranspoGene database[67] we examined the entire set of genes for these elements; none were identified.

Note that the order of operations is independent for the above filters; some probes fall into multiple categories so the total number of 'bad' probes identified per step will be greater than the total number removed.

B. Background (Noise Estimation)

The probes that form very stable heterodimers and have a single target (do not cross-hybridize) can provide insight into expected noise from the Agilent scanner (it must be estimated since the information is not given). Once this value has been determined, it can be used as a filter for eliminating the probes that have signal below the detection boundary. Candidate probes were identified using queries for uniqueness and free energy; measurement values were then retrieved, from each dataset independently, and the mean, median and lowess values were determined.

C. Sample Outlier Detection

We compared the signal intensities in the normal and diseased samples to the mean and variance of each class, and the distributions in order to identify samples outliers. Two contrasts were examined:

(i) Sample to class comparison of average number or probes and intensity per probe: Using only probes with acceptable measurement-yielding profiles, for each sample class and accepting only measurements above the lower detection boundary, we determined the mean signal per probe per array and across all arrays in the class. Samples whose probe-signal mean fell more than two standard deviations outside of the array mean were rejected. We then determined the number of acceptable probes that yielded good measurements per array, and across all arrays in the class, and similarly rejected any sample for which the number of informative probes fell more than two standard deviations from the class mean.

(ii) Sample distribution comparisons using filtered probe intensities: We compared the within- category and between-category distributions of probe intensities in the measurement-quality class. Non-normal distributions would suggest application of a log transformation, while sufficiently dissimilar distributions (test is described below) preclude the use of some statistical tests.

Statistical Analyses for Distribution

For the arrays that passed the LO-BaFL pipeline, the distributions of the intensities of the final set of acceptable probe values were tested for each class and experiment. Individual normal and diseased samples were labeled with Cy3 and the pooled reference was labeled with Cy5, which means the pooled reference group had twice as many members. The Shapiro-Wilk test [68-70], implemented in R, was used to check for normal distribution within and between sample classes. Since the results of both experiments show a non-normal distribution (data not shown), the Wilcoxon non-parametric test[71],[72] for unpaired groups was applied when performing comparisons for differential expression.

Significance of Differential Expression

Microarrays are the poster child for the multiple-hypothesis testing conundrum [73]. We addressed this issue using the Benjamini and Hochberg FDR procedure [74] implemented in R. The output consists of a list of DE genes and associated p-values. The R scripts for statistical analysis and the output file with DE genes can be found in Supplementary Material / Scripts/ stats_R.txt.

The control method, TM4 [47, 56] takes as input the spot intensity values for a set of arrays categorized by experimental design. TM4 allows for signal normalization (total intensity normalization, Lowess normalization), standardization (standard deviation regularization) and low-signal intensity filtering. Modified intensity values are the output, which are then used by the Significance Analysis of Microarray (SAM) package [57] or non-parametric tests to identify the differentially expressed genes. The output is a list of DE genes and associated p-values. The sequences of steps performed as statistical analyses are shown in Figure 1.2.

1.3 Results and Discussions

The pipeline for probe filtering

Data: the raw microarray data is publicly available at NCBI Gene Expression Omnibus, GSE28253 [75]. The specified intermediate pipeline output and the final results of our analyses are available in the Supplementary Material section unless noted.

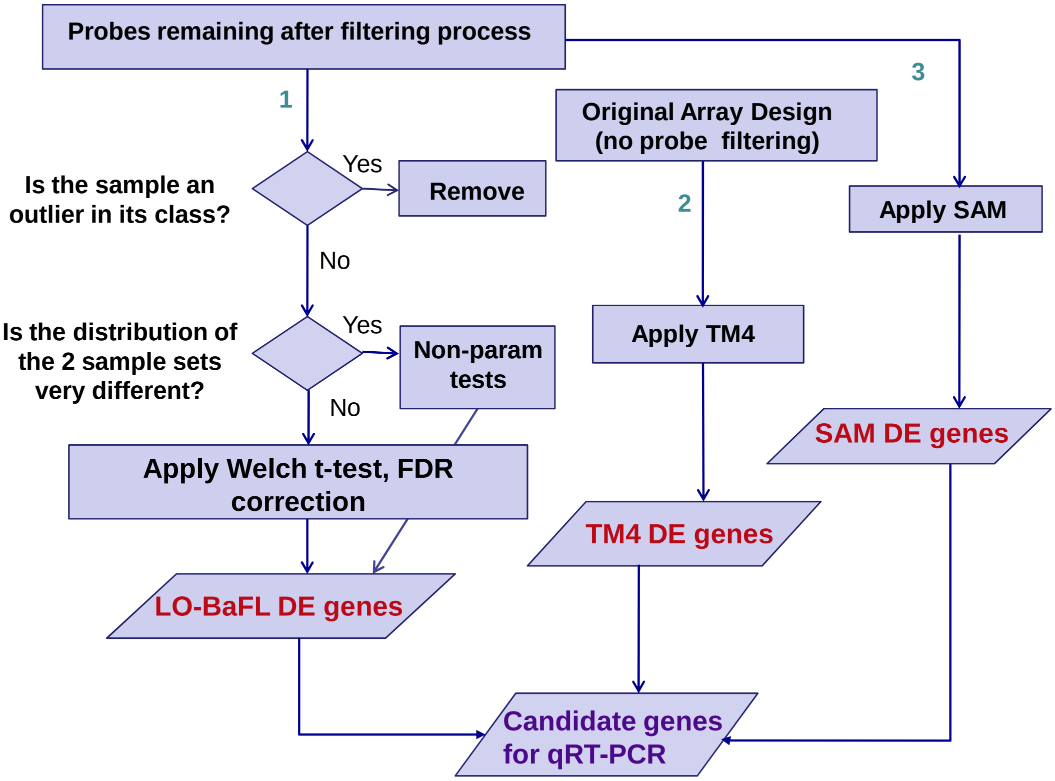

Probes remaining after filtering process
1
2
3
Original Array Design (no probe filtering)
Is the sample an outlier in its class?
Yes
Remove
No
Apply SAM
Is the distribution of the 2 sample sets very different?
Yes
Non-param tests
No
Apply TM4
SAM DE genes
Apply Welch t-test, FDR correction
TM4 DE genes
LO-BaFL DE genes
Candidate genes for qRT-PCR

Figure 1.2 The LO-BaFL pipeline (1) and the TM4 (2) and SAM (3) pipelines.

Pipeline:

(i) Re-mapping the Agilent probes to the Human Reference (HuRef) Genome 36.1 build. The 41,000 Agilent probes were scanned against the human reference genome using the Tera-BLAST algorithm implemented on the FPGA-accelerated platform from TimeLogic [61]. This search was implemented to find near-perfect as well as perfect matches, using parameter settings mentioned in the Materials and Methods section. We found a total of 370,139 hits to the probes, averaging ~9 matches per probe. The full list of matches is available on the project Web site / Cleansing process/tera_blast_results/tera_blast_raw.zip. In a gene expression array the impact of a perfect match to a secondary target depends on whether it is an expressed sequence from the complementary strand. The majority of additional locations did map to genes and did not appear to map across exon/intron junctions, although this does depend on the gene model used (data not shown).

(ii) Perfect and partial matches. Where a secondary target is not a perfect match there must be some boundary conditions for determining where sufficient signal contamination will occur to confound the interpretation of the data. We chose to follow the Kane criteria [62], adapted to Agilent 60-mers, such that 50/60 nucleotides have to match overall, with a minimum nucleation length of 15 nucleotides somewhere in the duplex. Applying these filters to the output above suggests that ~8.63 % of the probes would produce confounded measurements (signal coming from distinct loci), and that subset of probes was flagged. Oligonucleotides that report on multiple loci are usually eliminated from the measurement pool, at least in initial data cleansing efforts, since interpretation of the values is problematic [76, 77]. The file so modified is provided on the project Web site under Cleansing Process/cross-hybridization_filter/total_probes_no_crosshyb.csv.

(iii) Identifying the loss of probes. There are 407 probes that no longer map to the HuRef Genome 36.1 build (file = probes_info_no_pm_no_crosshyb_not_mapped.csv in Supplementary Data/Cleansing Process / loss_of_target_filter), these were flagged for removal from the active probes list.

(iv) Identifying the presence of SNPs. Although the cut-off is somewhat arbitrary, probes with four or more SNPs were removed (~2.53% of the remainder), since their presence would significantly distort the apparent concentration [65, 66]. The file showing the probes and the major/minor alleles for each SNP position is in Supplementary Material/Cleansing process/SNP filter/snp_info_probes_gt_4snp.csv. A separate file, 'agilent_probe_info_3SNPs.csv' gives the information on probes with less than four SNPs, for those wishing to adjust the stringency of SNP cut-offs.

(v) Delta G filter. The logic for deciding the cut-off for internal stability is described in the original BaFL report. For these 60-mers and hybridization conditions, $\Delta G = -5.2$ kcal mol^{-1} shows the comparable response [32], resulting in filtering out ~21.5 % of the probes, listed in the Supplementary Material file under "DeltaG_filter" tables.

The summary of the pipeline effects (as percent of probes filtered out per step) is shown in Table 1.1.

(vi) Poly-G filter. A factor added to the BaFL pipeline subsequent to the published report is the presence of poly-G (a run of G's >3) in the probe (Thompson, personal communication). The phenomenon of 'bright spots' from such runs has only been reported for short oligo arrays [14-16], but it is reasonable to check for them on longer oligonucleotide arrays as well. There are 4,742 such probes in the original data set (see file 'log_signal_4G_probes_total_no_filters.csv,' under polyG_filter of Cleansing Process in Supplementary Material). Of these ~10.2% had unusually high intensity ($\log_{10}(I) > 3.5$) while 50% had $\log_{10}(I) < 2.0$. Only 11 probes with this feature were present in the final list of acceptable probes (explained below), and these were flagged for removal.

Table 1.1. The % probes removed in total by applying the filtering steps of LO-BaFL pipeline and comparison with percent probes eliminated in the original BaFL pipeline.

Applied filter	**% Probes filtered out (LO-BaFL)**	**% Probes filtered out (BaFL) [32]**
Cross-hybridization	8.63 %	60.30%
Loss of target	0.99 %	2.19 %
SNP	2.53 %	1.78%
ΔG	21.46 %	5.17%

(vii) Repeated elements filter. A screen of the remaining probes for LINE, SINE and Alu subsequences was performed against the TranspoGene database [67]; no matches were identified.

Background estimation (instrument cut-off value)

The lower detection limit for our Agilent scanner was not available as a technical specification, so it was necessary to estimate it. Since this is not a standard method in most analysis pipelines the rationale for the steps is given here. Probes that do not cross-hybridize are used, in order to limit the possibility that a high concentration of target comes from an unexpected source. From these we selected those with a very stable internal structure so that little was available for duplex formation with target ($\Delta G < -5.2$ kcal mol^{-1}); this subset is given in (Supplementary Data/Cleansing process/Determining_instrument_cutoff/delta_g_mean_and_log_int_probes_no_crosshyb .csv). The goal is to identify a cutoff below which there is little variation in signal across many different probes, indicating that the response to changes target concentration has been compressed. In examining the intensities of this subset of probes in both normal and diseased samples in the ALS study, $\log_{10}(I)_{mean}$ and $\log_{10}(I)_{median}$ yielded values of 3.51 and 3.49 respectively, while for the CAD samples the values were 3.0 and 2.9 respectively, which indicates that this is either an experiment- or a scanner-specific value (we cannot separate labeling and scanner sensitivity factors). This is ten-fold higher than the value we find for most experiments using Affymetrix scanners (Thompson, personal report). If the extremely stable probes are eliminated ($\Delta G < -10$ kcal mol^{-1}), the background cut-off values approach the Lowess smoothing values, shown in Figure 1.3 for the ALS samples (lower panel), and also in Figure S1 for the CAD data (see 'Supplementary Data/CAD study'), and now approximate the values (200-300 fluorescent units) seen for the Affymetrix scanners. In the absence of calibration standards we cannot discriminate scanner detection limits from target fragmentation and labeling efficiency, but clearly the noise limit is experiment dependent and should be carefully determined for each experiment, and standardized for meta-experiments. This filter caused the largest single-step removal of probes for ALS experiment with all samples (~27,000 probes = ~ 95% of the total removed in this step, Supplementary

Material/Cleansing Process / Instrument_cutoff / all_samples_gt_instrument_cut-off_log_intens.csv). If a smaller value (2.5) for instrument cut-off is used, the percent of probes removed by this filter is reduced to ~ 78%, but the list of DE genes obtained is not any more similar to TM4 than before (see Results).

Sample outlier identification

We mirrored the BaFL approach for detecting *sample* outliers: briefly, one determines for each sample in a class the number of probes whose signal is above background and the the average signal per probe and compares the values to the sample-class means [32]. In experiments conducted on human patients with long-standing debilitating disease there is a strong likelihood that multiple conditions are present; a large difference in the number and identity of genes expressed may mean that part of the response is due to a second agent, so samples are screened for large overall response differences. A second difference from the earlier pipeline arose because of the experimental design of one these studies: the ALS experiment used a common reference design, so we added a step to determine how reproducible the signal of that reference is across all of the samples. Any sample (here represented by the array) for which the number of probes or intensity per probe falls more than two standard deviations from the mean for the category is considered an outlier and is not used. Defining an outlier as more than two standard deviations away from the class mean for either of these criteria, none of the samples in the ALS experiment failed, including those that were not tested by gene-specific PCR because of poor RIN numbers, indicating that the original samples were most likely of acceptable quality. No sample outliers were detected in the CAD experiment.

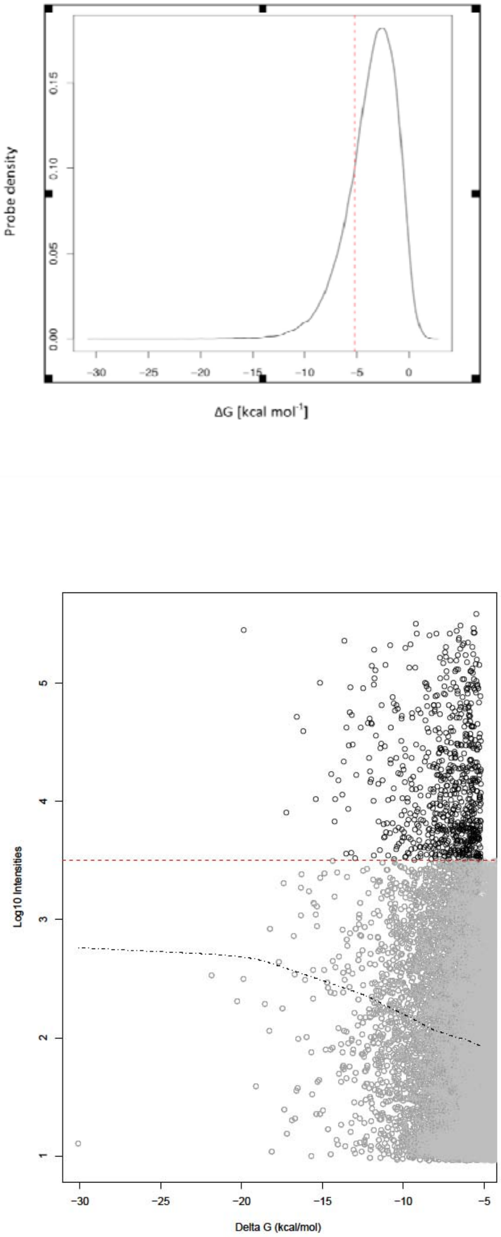


Figure 1.3. (upper) Graphical representation of $\Delta G_{\text{cut-off}}$ results: the probes having free energy, $\Delta G < -5.2$ kcal mol^{-1}, represented by the red line, were filtered out and the proportion to the left of the line is 21.5%; (lower) ΔG vs. Probe signal: the red line denotes background cut-off value; grey line is the Lowess smoothing line between ΔG and log10 intensities; grey dots represent the probes with very stable structures that have been eliminated in the process; black dots represent the probes with signal higher than the background cut-off value and $\Delta G < -5.2$ kcal mol^{-1}.

Additional Probe Restrictions

In the set of 12 samples whose cDNA passed the quality control (QC) step (see below), a set of 1,552 common probes was retained by the filtering process. Including the 12 samples with somewhat degraded cRNA (poor RIN scores) decreases the number to 1,327 probes to test for expression differences.

Sample Distribution Testing and Predicting Differential Expression

Once unreliable sensors (flawed probes) and measurements (scanner limitations) have been screened out, comparisons of the remaining intensity distributions allow one to select a valid statistical method to identify differentially expressed targets. In both studies, the F test indicated unequal variances between the two groups of samples. Probes for the same gene in the two sample classes failed the Shapiro-Wilks test for normality [68-70]. Our results indicated the presence of unequal and non-normal distributions (see Figure 1.4 and Figures S2, S3 in Supplementary Data/CAD Study); therefore, the use of a nonparametric equivalent to the t-test was chosen: in this case, the Wilcoxon two-sample test for unpaired groups [71, 72] was applied. The multiple-comparison problem is well-known for these experiments; we controlled for the false discovery rate (FDR) [74] using a setting of 0.20 and corrected the p-values accordingly with either the Bonferroni correction [78] or the Benjamini and Hockberg correction [74]. In each case no DE genes remained, suggesting that the criteria were too stringent. If we accept the argument that the multiple-testing criteria are too stringent [79], and use a $p < 0.05$ for significance, in the ALS experiment 87 probes were returned as DE for the complete dataset, with a subset of 60 of those reported as DE in the high-quality samples.

For the CAD study, 386 genes were found to be differentially expressed. The list with all DE genes for ALS is provided for each set, in Supplementary Material/Data post filtering/DE_genes_12(or all)_samples tables. DE genes for the CAD experiment are listed in:

Supplementary Data / CAD Study / DE_genes_CAD / DE_genes_CAD_data.csv.

TM4 and Significance Analysis of Microarrays (SAM) for ALS Data set

TM4, a widely-accepted platform for analyzing Agilent microarray data, incorporates SAM, which included a choice of non-parametric tests for the statistical analysis; thus we selected it as the standard pipeline against which to compare the LO-BaFL pipeline. Using the ALS experiment samples, analyses using several parameter settings were performed so that we could compare outcomes, and we used both the complete sample set and the highest RIN quality-validated subset of samples for each. The base-line analysis used TM4 default settings, which includes several normalization steps (e.g., total intensity normalization, Lowess normalization, standard deviation regularization), and filtering for the lowest 5% intensity signals, a common signal detection boundary of 100 for Cy5 and Cy3 intensities, and the Wilcoxon non-parametric test to determine DE genes (for convenience we label the results of using this method ‘TM4-W’). A second analysis path used LO-BaFL to remove problematic probes and the Wilcoxon non-parametric t-test to determine DE genes. It is possible that the data cleansing results are essentially the same even if the approach is not, so we used the LO-BaFl results as input to SAM, and within SAM opted for the Wilcoxon test (labelled as ‘SAM-W’).

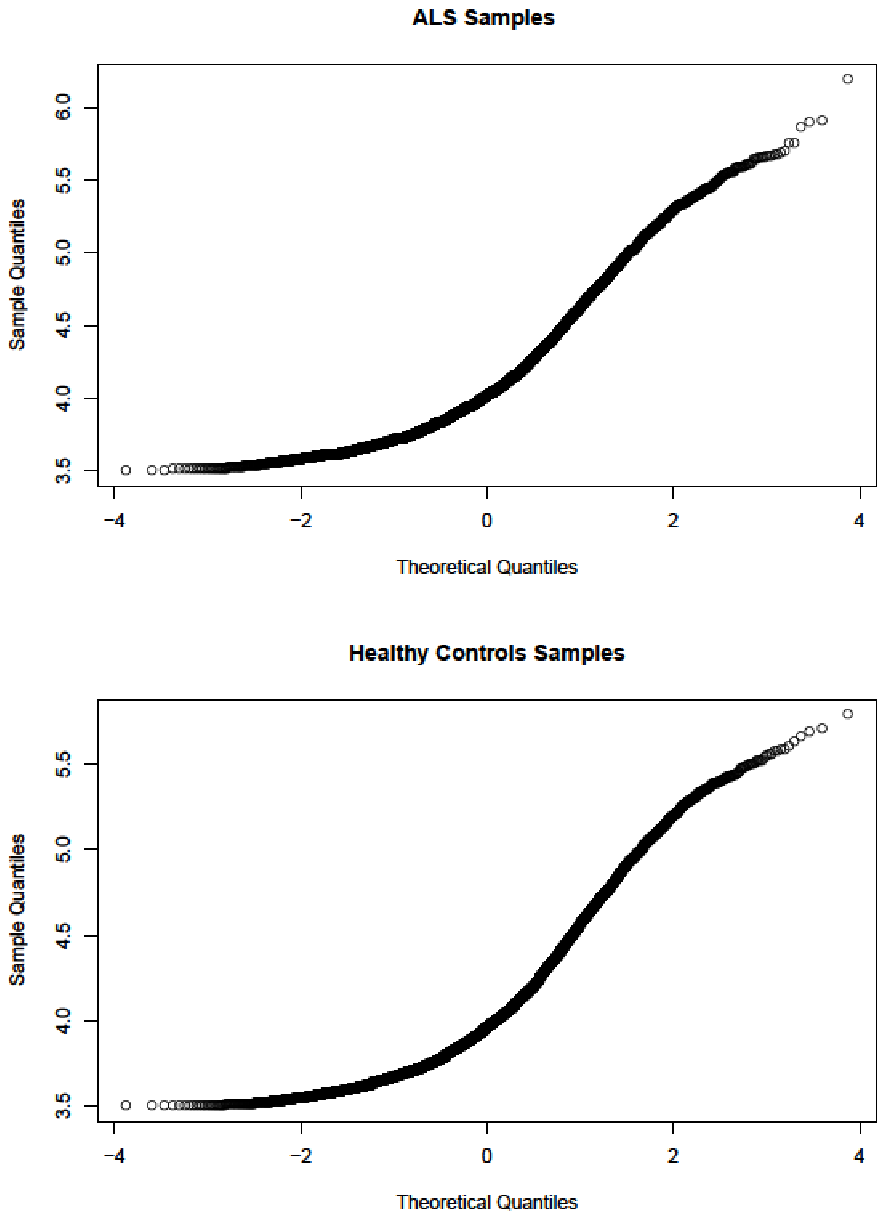


Figure 1.4. Q-Q plot distribution of diseased (upper) and healthy controls (lower).

The main caveat in this comparison is that there must be a sufficient number of observations in the classes for the method to be valid. SAM takes measurements and response variable category as input, and uses permutations to determine the strength of association; we set the permutation number to 100. Table 1.2 compares the six lists of 5 most significantly DE genes obtained from our analyses. The R implementation and SAM implementation of the Wilcoxon are very similar, with SAM perhaps being slightly more stringent since it eliminates one gene allowed by the other algorithm. The number of samples made a large difference, with only 1 of 5 genes being in common when 12 or 22 samples were processed with LO-BaFL (that being *JUNB*) or with TM4 (the gene being *DYNLT1*). An obvious reason for the disparity is if the probes have been deprecated in the list of acceptable probes. Checking the list of such probes showed that four of the TM4 DE genes fell below the minimum signal boundary set for LO-BaFL, explaining their absence. The fifth TM4-predicted gene, *DYNLT1* did not appear on the LO-BaFL list because it did not meet the p-value criterion.

Confirmatory analyses on DE genes in CAD samples

The list of DE genes determined by LO-BaFL was compared with results reported in [45]. Two of the genes appear on both lists (*CSPG2, ALOX5*), four are close variants of the DE genes, while the remainder of the genes reported in the paper were eliminated from out list based on specific criteria including the strong structure (delta G), low signal (scanner limitation) or p-values that failed our significance criterion (See ‘Supplementary Material / CAD study / comparison_with_DE_genes_CAD / comparison_LO-BaFL_CAD_DE_genes.csv’). The file listing DE genes for this experiment as determined by our pipeline is found as ‘LO-BaFL_DE_genes_CAD_data.csv’, located in the same directory mentioned above.

Analysis of Healthy Control samples from a different experiment and comparison with Healthy Controls from ALS data.

A major shortcoming of many analysis pipelines is that they are over-tuned to a particular experiment, so that parameters that yield excellent results in one study give poor results in another.Our intent is that the LO-BaFL filters be mostly experiment-blind (except for the determination of scanner detection limit and sample outlier status); if this is true LO-BaFL should predict the behavior of genes in similar samples but different experiments relatively well.

We looked for experiments in which human peripheral-blood samples and the Agilent arrays were used. One such studied coronary artery disease (CAD) [45] from controls (n=14) and diseased (n = 27). The data is accessible at GEO, Accession No.GSE10195. We compared the behavior of the two control groups: those without CAD in one study (six randomly selected samples out of 14) and without ALS in the other (six controls). An anomaly in the CAD study was a number of spots with 'negative' intensities (often saturated spots that the software does not know how to handle), which were removed. Prior to probe filtering, all of the samples have measurements for 24,336 genes. After LO-BaFL filtering being applied we graphed and compared the probe intensities in Normal samples from each study. We found a good correlation across genes between the two groups, indicated by the near-linearity of the Lowess smoothing line, shown in red (Figure 1.5). The genes that are most highly expressed in each set of samples (e.g. *RPS2, RPLP1, RPS28, HLA-C*) and expressed at low but detectable levels (e.g. *CD28, CDV3, CD79A, CCD12*) are characteristic of white blood cells.

Table 1.2 ALS experiment: Selected differentially expressed genes with $p < 0.05$, determined by LO-BaFL-Wilcoxon (LO/W12, LO/W22), TM4-W (TM4/W12, TM4/W22) and SAM-W (SAM/W12, SAM/W22).

List of DE genes	**Gene/Accession**	**Description**	**p-value /q-value**
LO/W12	FTH1/ NM_002032	Ferritin, heavy polypeptide 1	1.59E-3
	JUNB/ NM_002229	Jun B proto-oncogene	3.67 E-3
	B2M/ NM_004048	Beta-2-microglobulin	1.54 E-3
	ACTG1/ NM_001614	Poly(A) binding protein, cytoplasmic 1	3.7 E-3
	SLC25A3/NM_005888	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3	4.46 E-3
LO/W22	EXOC3L2/NM_138568	Exocyst complex component 3-like 2	5.73 E-3
	FAU/ NM_001997	Finkel-Biskis-Reilly murine sarcoma virus	1.96 E-3
	GLTSCR1/ AF182077	Glioma tumor suppressor candidate region gene 1	2.56 E-3
	JUNB/ NM_002229	Jun B proto-oncogene	1.24 E-3
	IRS2/ NM_003749	Insulin receptor substrate 2	1.66 E-3
TM4/W12	CSE1L/ NM_001316	CSE1 chromosome segregation 1-like (yeast)	0.0
	NUP88/ NM_002532	Nucleoporin 88kDa	0.0
	PARP1/NM_001618	poly (ADP-ribose) polymerase 1	0.0
	DYNC1I2/NM_001378	Dynein, cytoplasmic 1, intermediate chain 2	0.0
	DYNLT1/ NM_006519	Dynein, light chain, Tctex-type 1	0.0

Table 1.2 (continued)

TM4/W22	TARDBP/NM_007375	TAR DNA binding protein	0.0
	DYNLT1/ NM_006519	Dynein, light chain, Tctex-type 1	0.0
	SKIV2L2/ NM_015360	Superkiller viralicidic activity 2-like 2 (S. cerevisiae)	0.0
	C12orf35/NM_018169	Chromosome 12 open reading frame 35	0.0
	IRS2, NM_003749	Insulin receptor substrate 2	0.0
SAM/W12	FTH1/ NM_002032	Ferritin, heavy polypeptide 1	0.0
	JUNB/ NM_002229	Jun B proto-oncogene	0.0
	B2M/ NM_004048	Beta-2-microglobulin	0.0
	ACTG1/ NM_001614	Poly(A) binding protein, cytoplasmic 1	0.0
	SLC25A3/NM_005888	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3	0.0
SAM/W22	IRS2/ NM_003749	Insulin receptor substrate 2	0.0
	GLTSCR1/ AF182077	Glioma tumor suppressor candidate region gene 1	0.0
	FAU/ NM_001997	Finkel-Biskis-Reilly murine sarcoma virus	0.0
	EXOC3L2/NM_138568	Exocyst complex component 3-like 2	0.0
	JUNB/ NM_002229	Jun B proto-oncogene	0.0

This result encourages us that we can extend our studies to additional microarray data of ALS samples, since LO-BaFL predictions of DE genes are often confirmed by qRT-PCR results.

1.4 Conclusions

We performed two related case studies, using data obtained from independent experiments, one on ALS and one on CAD. Transcript levels for both experiments were measured with the Agilent 4x44k platform. The data was processed using two pipelines: LO-BaFL and TM4.

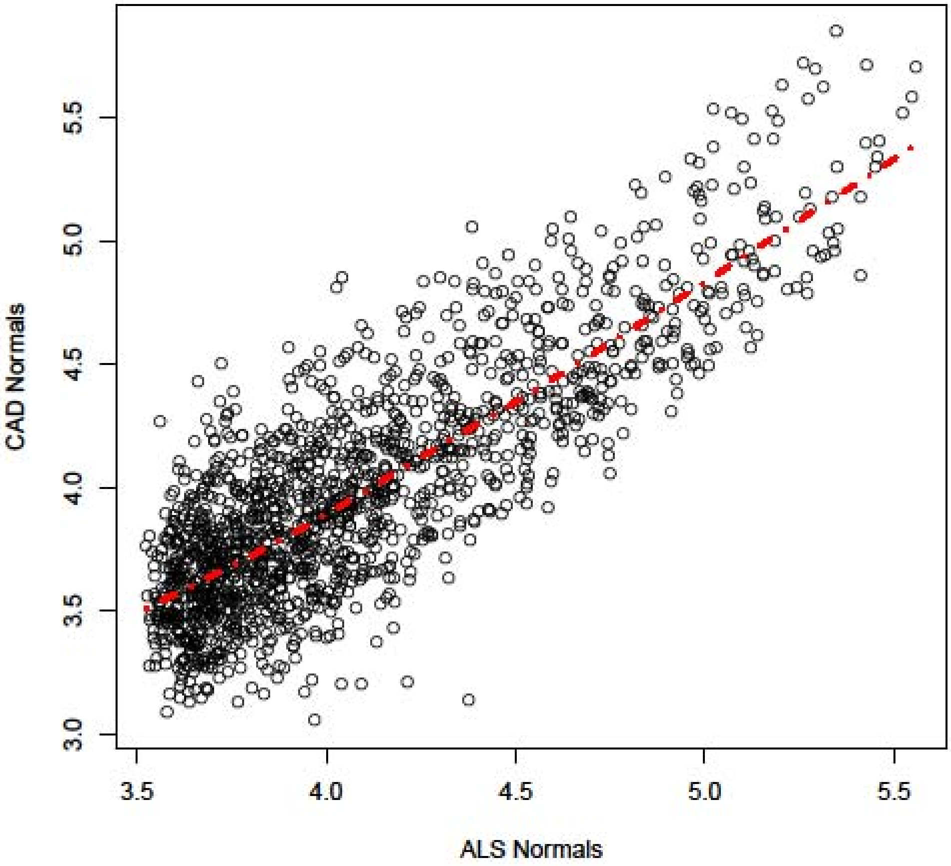


Figure 1.5. Correlation between healthy controls in ALS and CAD studies denoted by the Lowess smoothing line in red.

Comparing normal samples from the independent CAD experiment [47] to the ALS-normal samples indicated that the latter microarray hybridizations gave very similar results (albeit with somewhat different labeling efficiency), giving us more confidence in the differences observed with the diseased samples for the quite small ALS study. Several of the most significant DE genes in ALS were related to immune responses, while in the CAD study the DE genes were involved in atherosclerosis, cell motility, as signaling receptors or transcription factors [45]. Our pipeline was applied to the CAD data as a whole (healthy n = 14, diseased n = 27), paying particular attention to those DE genes that the researchers of the original study tested with qRT-PCR assays. We compared the DE genes determined by LO-BaFL with their list and the result shows that except for several genes that have been eliminated by ΔG filter or by the background cut-off filter, the rest are found in our list with significant expressed genes. Two of them are confirmed to be DE, four are close variants and several others were dropped because the p-value fell just below our cut-off (See Supplementary Material/CAD study). Furthermore, a comparative analysis of controls in ALS study vs. controls in CAD study shows a very good correlation between the two groups.

We note that these studies had considerably fewer disease samples than were available for the original BaFL study, which used a large, publicly available lung cancer dataset. ALS is sporadic, rare, and has a mysterious etiology [1, 7, 13]; the inherent small number of samples means that methods for increasing the power of studies are even more important.

CHAPTER 2: VALIDATION OF DIFFERENTIALLY EXPRESSED GENES ASSOCIATED WITH sALS BY qRT-PCR ASSAYS

2.1 Introduction

The research described in this chapter represents an extensive investigation of the differentially expressed genes, as determined by applying LO-BaFL pipeline, defined in details in Chapter 1, to available microarray data and by comparative bioinformatics methods, e.g. TM4, in peripheral blood lymphocytes (PBLs) from patients with sALS, and normal patients, samples provided by Carolinas Neuromuscular/ALS-MDA Center, Charlotte, NC.

Employing different methods to obtain DE genes on same the data set has the advantage of determining a more complete list. More specifically, the LO-BaFL pipeline is designed to eliminate and flag any probes that are cross-hybridizing, whereas TM4 and SAM do not have this filter. For instance, one Agilent probe representing TARDBP, one of the most important genes in sALS studies, was found to cross-hybridize with *IlKAP*. The latter was not identified by TM4 as being differentially expressed and therefore, it would have been excluded for further assays if only the TM$-SAM computational method was used. Because LO-BAFL identifies genes by category, cross-hybridizing genes can be specifically identified in the absence of microarray DE predictions. Thus we added to the list of genes to test by qPCR TARDBP and ILKAP, neither of which was predicted as significantly DE by LO-BAFL and only the first of which was predicted by TM4.

For any diagnostic test a confirmation of the results must be demonstrated using an independent method. This is particularly true of microarray results, which tend to produce long lists of DE genes that appear only in single studies. While the follow-up assay id not prescribed, it is most commonly a quantitative PCR assay, either absolute or relative, since the set of collective methods is sensitive and reproducible. Because reagents for the absolute quantitation method do not exist for the amplicons in our study we chose the relative quantitative approach. This requires that the different efficiencies of amplification be considered in the data analysis, explained in more detail in the Discussion section.

2.2 Materials and Methods

Gene selection

We selected *FTH1*, *JUNB*, *B2M*, *ACTG1*, *SLC25A3* as top common DE genes for LO-W and SAM-W (see Chapter 1), and in addition, *SKIV2L2, C12orf35, DYNLT1 and TARDBP* (determined by TM4-W) and its corresponding cross-hybridizing genes (e.g. *ILKAP, DIAPH3*), to be tested in the lab.

Also, we adopted a set of four reference genes to assess the sample and assay conditions, according to best practice recommendations [49]. A reference is context specific: these were selected based on an apparently consistent level of expression in the microarray data across sample classes and individuals, in the middle range of concentrations. These references included: *UBE2Z* (ubiquitin-conjugating enzyme E2Z), *PGK1* (phosphoglycerate kinase 1), COX4I1 (cytochrome c oxidase subunit IV isoform 1), and *SRRM1* (serine/arginine repetitive matrix 1).

The qRT-PCR assay reagents and quantitation templates were developed and the titrations of samples against standards performed according to standard methods [51]. Primers and reference template sequences are provided in Table 2.1. The instrument was the Bio-RAD MyiQ Single-Color Real-Time PCR Detection System [80]. We used the software that ships with the instrument to perform initial calculations; we chose the maximum correlation coefficient approach to determine the C_ts, from which the starting concentrations of the unknowns were estimated. Further analyses were performed with the Pfaffl method [81].

In the absence of a calibration standard the actual expression levels of genes in the individual samples are not readily available. Thus the wet-lab work had two goals: determine the level of expression that a microarray value yields in a qRT-PCR assay; determine whether either pipeline was accurate in its predictions of the predicted difference in expression levels between normal and diseased samples. Table 2.1 shows the genes selected and their category. Those marked as ‘reference’ are expected to be expressed in PBLs at moderate and consistent levels in all samples.

A diagram representing the four steps involved the experimental part of the present study is sketched in Figure 2.1.

Quantitative and qualitative assessment of RNA

The isolated RNA from peripheral blood lymphocytes samples of healthy controls and ALS patients, stored at -80° C, provided by Carolinas Neuromuscular/ALS-MDA

Center, Chalotte, NC was qualitatively checked using the Agilent 2100 Bioanalyzer [58], and quantified with the Nanodrop ND-1000 from ThermoScientific [82]. We carried forward only those samples that satisfied the condition that RIN >5.5.

cDNA synthesis and QC

In addition to the samples used in the microarray experiments, we extracted RNA from other samples of blood cells, suspended in Trizol and kept at -80º C, as positive controls, using the AllPrep DNA/RNA Mini Kit from Qiagen [83], following the manufacturer's instructions. This RNA was qualified and quantified as above. We then synthesized double-stranded cDNA from the ALS samples that passed the RNA quality/quantity test (6 normal controls and 6 diseased) and from the control RNA, using the Full Spectrum^{TM} Complete Transcriptome RNA Amplification Kit from System Biosciences [84], according to the supplier's manual. After quantification of the yield, and standardization of the concentrations, the cDNA products were qualified by determining whether the reference gene PCR primers yielded the expected size product on 12% acrylamide (native, in 1X TBE buffer) gels [85]. Even where the starting concentration of RNA was low, e.g; 20 ng, we obtained good yields of cDNA. The PCR reaction conditions were as follows, per 50ul final volume: 5.0 µl /reaction of 10X Buffer (Invitrogen^{TM} [86]), 3.5 µl /reaction $MgCl_2$ (Invitrogen^{TM} [86], 50 mM stock solution, for Mg^{++} 3.5 mM final concentration), 2.5 µl /reaction dNTP mixture (Invitrogen^{TM} [86], 10 mM stock, 2.5 mM final), 0.5 µl /reaction DNA Taq Polymerase (BioLabs$^{®}$ Inc.[87] concentration of 100 mM stock; final concentration of 5 mM), 2.0 µl /reaction cDNA as template (100 ng).

Table 2.1 The list of reference genes and DE genes as determined by qRT-PCR and their corresponding designed primers.

Gene information	**Gene role**	**Forward primer (5` to 3`)** **Reverse primer (3` to 5`)**
UBE2Z, NM_023079.3	Reference gene	GCAGAGCATGTCTGGCATAG TTCTCCTTCTGCCAAAACAAA
PGK1, NM_000291.3	Reference gene	TGCATCTCCACTTGGCATTA TGGGATCTTGAAGAATGTATGC
SRRM1, NM_005839.3	Reference gene	GGAAATCCTTGGGTTTGAAGA GGCCACAGTTCTCCCATAAA
COX4I1, NM_001861.2	Reference gene	GGCACTGAAGGAGAAGGAGA GGGCCGTACACATAGTGCTT
B2M, NM_004048	DE gene determined by LO-BaFL / SAM	GATGAGTATGCCTGCCGTGTG CAATCCAAATGCGGCATCT
ACTG1, NM_001614	DE gene determined by LO-BaFL / SAM	AGAGGCTGGCAAGAACCAGTTGTT CAATGACGTGTTGCTGGGGCCT
DYNLT1, NM_006519.1	DE gene determined by TM4	CCAGCCTATGGCCTTTCTCCTTTTGT CAACGCAGGCTGCAGGTGAC
SKIV2L2, NM_015360.4	DE gene determined by TM4	TGCAGAAGGAATCACCAAAA ATGGGAGAACCAAATCCACA
C12orf35, NM_018169.3	DE gene determined by TM4	CGGGGAAACAAGGTATTTGA TTCACATCACAGTGGGCATT
TARDBP, NM_007375.3	DE gene determined by TM4	TTTGCTGCAGTTCTGTGTCC TCCATCTCAAAAGGGTCAAAA
ILKAP, NM_030768.2	Cross-hybridizing gene with TARDBP	CACAGGAGTACACAAAACACAC TGCGGATAGGGCACTGAG

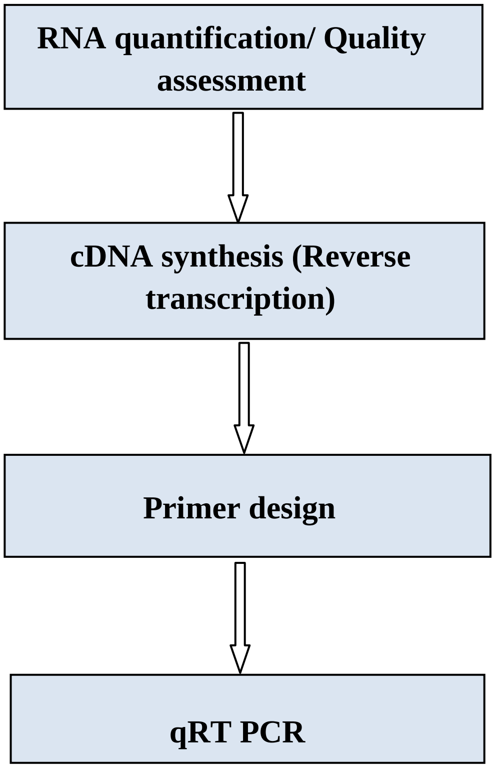

RNA quantification/ Quality assessment
cDNA synthesis (Reverse transcription)
Primer design
qRT PCR

Figure 2.1 Schematic diagram of the experimental design

The GeneAmp® PCR System 9700 from Applied Biosystems [88] was set up to the following profile: the initial DNA denaturation at 95° C for 5 minutes; 30 cycles of denaturation at 94° C for 30 seconds, primer annealing at 57° C for 30 seconds and extension at 72° C for 30 seconds; a final elongation at 72° C for 4 minutes and a 4° C hold.

Primer design and synthesis

The primers were designed using Primer3 software [89] in combination with NCBI Primer-BLAST [90] to check for the specificity. Whenever possible (exceptions are discussed below), the primers were designed to bridge the positions occupied by the corresponding Agilent probes, in order to account for sensitivity to transcriptional isoforms. Primers were purchased in dry form from Eurofins mwg|operon [91] and resuspended in DNA Suspension Buffer; concentrations were verified with the Nanodrop spectrophotometer, length was gel-verified using 12% Acrylamide in 1X TBE buffer [85]. PCR performance was checked with the cDNA made from the control RNA. PCR conditions were optimized by changing the Mg^{++} concentration in a range of 2.5 – 4.0 µM, the annealing temperature in interval 55 to 60° C or the dNTP mixture concentration from 2.5 µM to 3.5 µM. Where necessary new primers were designed and run again through the QC protocol mentioned above. The list of primers and their designed sequences are provided in Table 2.1.

qRT-PCR assay

Before proceeding with qRT-PCR assays with patient samples, we performed a quality-control process, using the reference samples, for each gene product in order to optimize the PCR reaction conditions. By adjusting the primer annealing temperature, the concentration of Mg^{++} or of the dNTP mixture concentration within the parameters described above, we amplified products with similar efficiency using a common set of PCR reaction and cycling conditions. These are: annealing temperature = 57° C; Mg^{++} concentration = 3.5 µM; dNTP mixture concentration = 2.5 µM; primer mixture concentration = 5.0 µM.

The quantification consistency was verified using parallel reactions, taking PCR reagents from a master mix to amplify the gene product reference at known concentrations against a mass-titration of a sample's cDNA product [51]. We used the following reagents: 10.0 µl/well of iQ™ SYBR® Green Supermix from Bio-Rad [80] that includes 2X reaction buffer with dNTPs, iTaq DNA Polymerase, 6mM $MgCl_2$, SYBR Green I, fluorescein and stabilizers according to the BIO-RAD specifications [80]; forward and reversed primer mixture (2.0 µl/well at 5 mM); 5.0 µl/well of template (either the standard gene, the unknowns or Accugene water-for the negative controls), and 3.0 µl/well Accugene water in final reaction volumes of 20.0 µl/well. Titration series were set up as follows: six 10- fold serial dilutions of the gene product reference and of the samples, in triplicates, with negative controls in all series to identify any cross-contamination problems. The reactions were set up in 96-well clear Multiplate® PCR Plates, covered with iCycler iQ™ Optical Tape from BIO-RAD [80]. The instrument employed for these reactions was MyiQ Single-Color Real-Time PCR Detection System from BIO-RAD [80]. We used a 2-step protocol with the following profile: *Cycle 1*: (1X) step 1: 95.0ºC for 03:00; *Cycle 2*: (40X) step 1: 94.0ºC for 00:15; step 2: 57.0ºC for 00:30 (data collection and real-time analysis enabled); step 3: 72.0ºC for 00:15; *Cycle 3*: (1X) step 1: 95.0ºC for 01:00; *Cycle 4*: (1X) step 1: 55.0ºC for 01:00.

The data were analyzed using the relative quantification method applied for reactions with different efficiencies, as described by Pfaffl [81]. According to this method, the relative expression ratio is calculated with formula:

$$Ratio = \frac{(E_{target})^{\Delta Ct_{target}(control-sample)}}{(E_{ref})^{\Delta Ct_{ref}(control-sample)}}$$

where: E_{target} = real time PCR efficiency of a target gene transcript;

E_{ref} = real time PCR efficiency of a reference gene transcript;

ΔCt_{target} = Ct deviation of control - sample of the target gene;

ΔCt_{ref} = Ct deviation of control - sample of the reference gene.

The corresponding real-time PCR efficiency of one cycle in the exponential phase was calculated, according to equation:

$$E = 10^{[-1/slope]}$$

were the slope was determined automatically by the machine software [81].

A standard curve is derived from serial dilutions, in our case six-point ten-fold dilutions and running standards in triplicates. Initial concentrations of standards and specific samples (unknowns), in logarithmic scale (base 10) are plotted against crossing points (Ct values). The regression coefficient R is calculated and adjusted for fitting the standards and unknowns: the greater and closer to 1 value of R, the better fit and thus, the better efficiency. We used the median value of efficiencies for E_{ref} . The summarized results are shown in Table 2.2 and selected standard curves can be found in Figure 2.2.

Table 2.2 Relative expression ratio for DE genes.

Gene Symbol	Gene Accession	Expression Ratio
ACTG1	NM_001614	48.5
SKIV2L2	NM_015360.4	37.3
C12orf35	NM_018169.3	22.4
B2M	NM_004048	18.2
DYNLT1	NM_006519.1	17.4
ILKAP	NM_030768.2	8.8
TARDBP	NM_007375.3	5.6

2.3 Results and Discussions

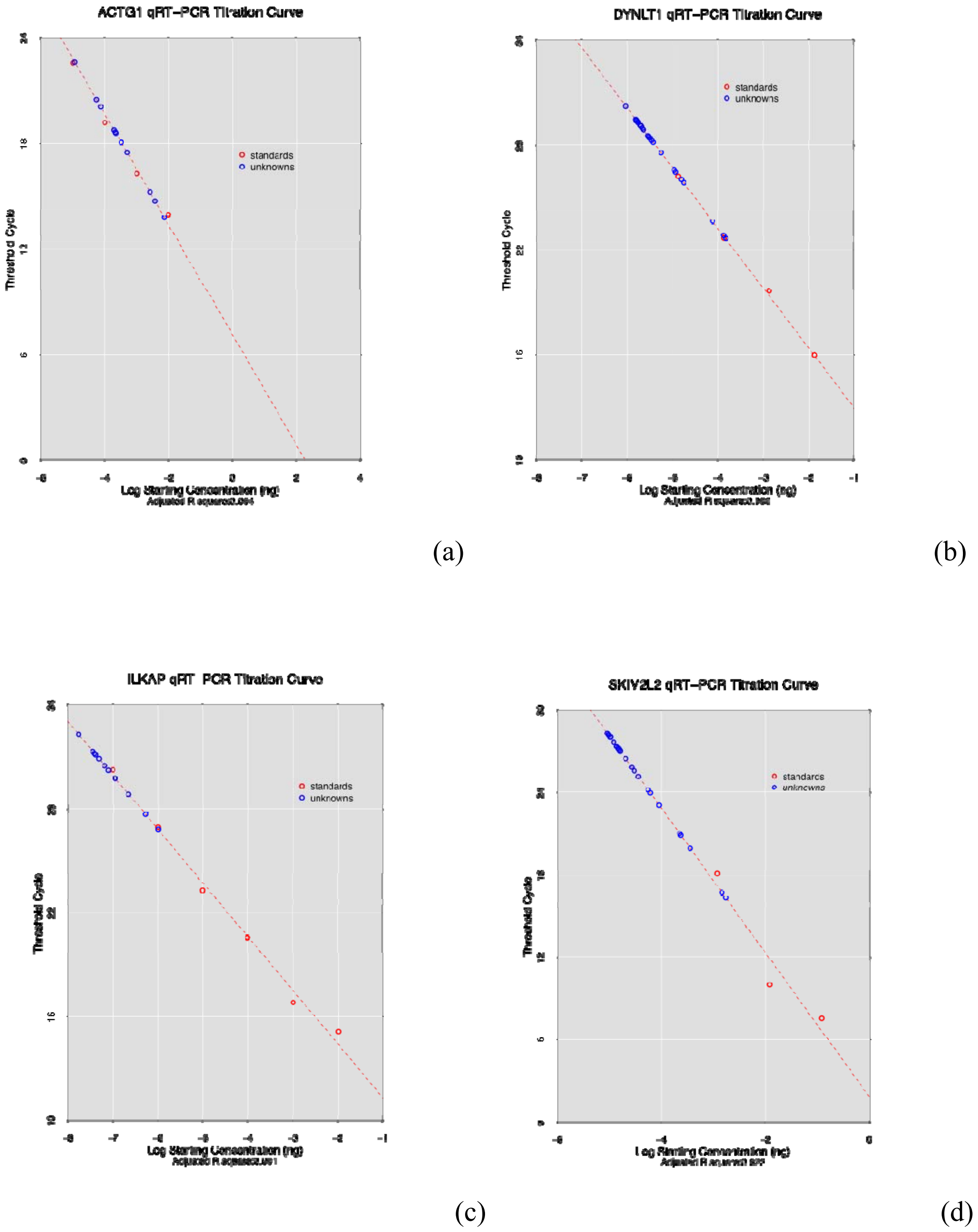
We observed significant differential expression ratios for *ACTG1, SKIV2L2, C12 orf35, B2M* and *DYNLT1* and differential but smaller differences in the values for *ILKAP* and *TARDBP*. Our experimental results are in good agreement with very recent findings by Mougeot *et al*. [75] showing, by computational methods, that *SKIV2L2, C12orf35, DYNLT1* were differentially expressed in PBLs samples from patients with sALS vs. Normals.

It is also confirmed here, as in previous studies, that *TARDBP* is among the genes with differential expression for ALS [8, 12]. The present work confirms the differential expression of three previously unreported genes (*ACTG1, B2M, ILKAP*) as determined by the LO-BaFL pipeline, with differential expression in the PBL samples from ALS patients vs. Controls.

Some of the TM4 genes were DE – that is LO-BaFL may exclude some genes that are actually DE, due to multiple and strict filtering steps, but the example of TARDBP and its cross-hybridizing gene, *ILKAP*, highlights why only TM4 is not efficient. In fact using both and then confirming the predictions with qRT-PCR may be the only way to be complete (pursuing mechanisms), while LO-BaFL is probably best for robust diagnostic predictions (less complete, but more likely to be right).

2.4 Conclusions

Testing of 12 genes with qRT-PCR, confirmed the microarray observations and most of our computational predictions when applying LO-BaFL and comparative methods for microarray analysis: *ACTG1, SKIV2L2, C12orf35, B2M, DYNLT1, TARDBP, ILKAP*



(a) (b)

(c) (d)

Figure 2.2 qRT-PCR standard curves for selected genes.(a) *ACTG1*; (b) *DYNLT1*; (c) *ILKAP*; (d) *SKIV2L2*.

were found to have higher expression ratio in patients with ALS vs. Healthy Controls. This confirms the results of previous and more recent studies [8, 12, 75, 92], with additional new candidate biomarkers in the genes *ACTG1, B2M, ILKAP*.

Chapter 3: SEARCHING FOR SEQUENCE VARIANTS ASSOCIATED WITH DIFFERENTIALLY EXPRESSED GENES DETERMINED BY LO-BaFL AND COMPARATIVE METHODS

3.1 Introduction

Although in the absence of a known cause no definitive statement can be made, it is thought at this time that ALS is a complex disease. This means that the condition is not determined by genetic mutations in a single gene, but is the result of accumulating errors (since it is usually late in onset, like many cancers) that change interactions between multiple genes. Sequence variants within transcripts, as distinct from their regulatory regions, are broadly of two types: single nucleotide polymorphisms that may or may not create a minor change in the coding sequence of a protein and may or may not change the processing of the transcript, and alternative splice forms that result in different exon presence and thus a significantly different protein form, possibly leading to different modification and localization. A likely example of the latter is seen with the *TARDBP* protein, for which aggregates are seen in all post-mortem sALS patients. Studies of genetic markers, such as SNPs, through linkage mapping or genome wide association studies (GWAS), can reveal genes associated with a particular disease if the association is sufficiently strong [16, 18, 93-95]. It is believed that single nucleotide polymorphisms are responsible for most of the genetic variation in humans, on average one site per 300 bases [96]. This variation conditions all biochemical responses and influences the variable responses seem for many human diseases, e.g.: altered responses to pathogens, chemicals, drugs and vaccines. Tuning treatments to specific variants may lead to interventions for genetic diseases and in gene therapy [97]. However, to uncover relatively weak effects over many gene combinations against a wide range of backgrounds requires sample sizes in the tens of thousands; with polygenic conditions the correct stratification of patients is often problematic. Worldwide the number of patients with ALS barely meets the required number and the majority will not be eligible for GWAS studies, for a variety of reasons. Thus studies of genetic variants present in ALS patients usually focus on genes with interesting molecular phenotypes, from expression of the transcript or expression and localization of the protein product. We investigated sequence variants in the samples that passed the quality criteria that occurred in our list of differentially expressed genes identified in Chapter 1 and Chapter 2.

We formulate the hypothesis that there are SNPs present in the differentially expressed genes that correlate to the disease state. Wet lab analysis of DNA from 10 patient samples (n=5, diseased; n=5, normals) were tested for the presence of sequence polymorphisms in specific regions of DE genes. Several of these genes have known SNPs, although they have not been linked to sALS, including *ACTG1, B2M* and *SLC25A3*; *ACTG1* is linked to muscle development, and one study showed reduced expression levels of the gene in an animal model correlating with human muscle weakness and myopathies [98]; *B2M* is found in amyloid particles characteristic of Alzheimer’s disease: its structure can adopt a fibrillar configuration seen in amyloid structures in certain pathological states [99]. We selected *FTH1* because defects in ferritin proteins are associated with several neurodegenerative diseases [100, 101], as well as two ribosomal proteins, *RPS10A* and *RPL21*, and *TARDBP*, which has been extensively studied at the protein level in both familial and sporadic ALS [108-111]. The genes and their dbSNP IDs are listed in Table 3.1. Of these genes, *SLC25A3* is a mitochondrial phosphate carrier (P_iC), was of particular interest, because Mayr *et al.* [102] have shown that a deficiency in P_iC in muscle is caused by a homozygous mutation in the alternatively spliced exon 3A of the gene. By replacing the guanine in position 215 with adenine (215G-to-A in the mRNA), a glycine becomes a glutamate (Gly72-to-Glu in the protein). The consequences are severe, including hypertrophic cardiomyopathy, muscular hypotomia, and lactic acidosis [102]. Studies on a separate glutamate solute carrier gene, *SLC1A2* did reveal an association with sALS [103-105]. Since LO-BAFL detected differential expression in the gene and the phenotype is relevant to the disease state we sequenced part of *SLC25A3*.

Table 3.1. Selected genes with known SNPs for sequencing and their dbSNP IDs.

Gene Symbol	**Accession**	**Exon / dbSNP ID**
B2M	NM_004048	Exon 1 / rs104894481
ACTG1	NM_001614	Exon 3 / rs28999111
		Exon 5 / rs28999112
		Exon 6 / rs104894547
SLC25A3	NM_005888	Exon 3A / rs104894375

3.2 Materials and Methods

Due to limited materials, only ten PBLs samples from patients with sALS (n=5) and Healthy Controls (n=5), were available from the original lot. Samples are described in detail in Chapter 1. Preparation for Sanger sequencing [106] required cDNA synthesis and PCR amplification: these protocols were described in Chapter 2. Design of amplification primers followed methods similar to those used for qRT-PCR (see Figure 3.1). From the list with differentially expressed genes (see Chapter 1) we targeted ten exons as follows: five for expressed genes with identified SNPs, (Table 3.1), and in addition, the exons comprising the Agilent probes for corresponding DE genes. The complete information for the sequenced genes (exon), primer sequences, annealing temperatures and product size are given in Table 3.2.

Obtained sequences were BLAST-ed against target sequences, with NCBI bl2seq, to check for similarity. The electropherograms were visualized and analyzed with FinchTV [107], a free tool for DNA trace view with enhanced capabilities for BLAST, reverse complement, and heterozygote detection.

3.3 Results and Discussions

We analyzed the outcome of 200 sequencing reactions to identify any sequence variants in our selected exons. There were no mutations that consistently segregated with the disease samples. A novel mutation in exon3 of the *ACTG1* gene, was found. This exon was screened for the SNP rs28999111, which has the sequence:

CGACATGGAGAAGATCTGGCACCACA[**C**/**T**]CTTCTACAACGAGCTGCGCGTGGCC

Shown below, the target amplified was designed to include the known mutation (shown in blue, the primer sites are also highlighted).

TGACCCTGAAGTACCCCATTGAGCATGGCATCGTCACCAACTGGGACGACATGGAGAAGATCTGGCACCACA**C**CTTCTACAACGAGCTGCGCGTGGCCCCGGAGGAGCACCCAGTGCTGCTGACCGAGGCCCCCCTGAACC**CCAAGGCCAACAGAGAGAAG**

Table 3.2 PCR primers for sequencing purposes. *Note*: The gene symbol in column 1 is followed by the exon number, parenthesis specifying the presence of a SNP or the fact that the primer was designed around the Agilent probe; F, R denote forward and reversed.

Exon	**Sequence**	**Annealing temperature (°C)**	**Product size (bp)**
B2M-e2(probe)-F	GTGTCTGGGTTTCATCCATCCGAC	57.5	176
B2M-e2(probe)-R	ACATGGTTCACACGGCAGGCAT	59.3	
FTH1-e4(probe)-F	CCCCATAGCCGTGGGGTGACT	60	170
FTH1-e4(probe)-R	CCCAAGACCTCAAAGACAACACCTG	58	
ACTG1-e3(SNP)-F	TGACCCTGAAGTACCCCATT	59	161
ACTG1-e3(SNP)-R	CTTCTCTCTGTTGGCCTTGG	60	
ACTG1-e5(SNP)-F	GTATGGAATCTTGCGGCATC	60	152
ACTG1-e5(SNP)-R	GGTGATCTCCTTCTGCATCC	60	
ACTG1-e6(SNP)-F	TGAGGCTAGCATGAGGTG TG	56.8	169
ACTG1-e6(SNP)-R	CCTTCCAGCAGATGTGGATT	55	
RPL21-e2,3(probe)-F	AGTTGTTCCTTTGGCCACATA	59.5	162
RPL21-e2,3(probe)-R	TTTACAACAATGCCAACAGCA	60	
RPS10-e1(probe)-F	CTCACAAGAGGGGAAGCTGA	60.5	151
RPS10-e1(probe)-R	TTTACTGAGGTGGCTGACCA	60	
SLC25A3-e3A(SNP)-F	CATTCCAGTGGCCTTAGTCA	54.5	203
SLC25A3-e3A(SNP)-R	TGCAAAACAAACCTGCATTC	52.5	
SLC25A3-e8(probe)-F	AGCTGTGGCACAACACATACAGC	59	152
SLC25A3-e8(probe)-R	AGCCAAGGAAAGTCGGAGCCCA	58	

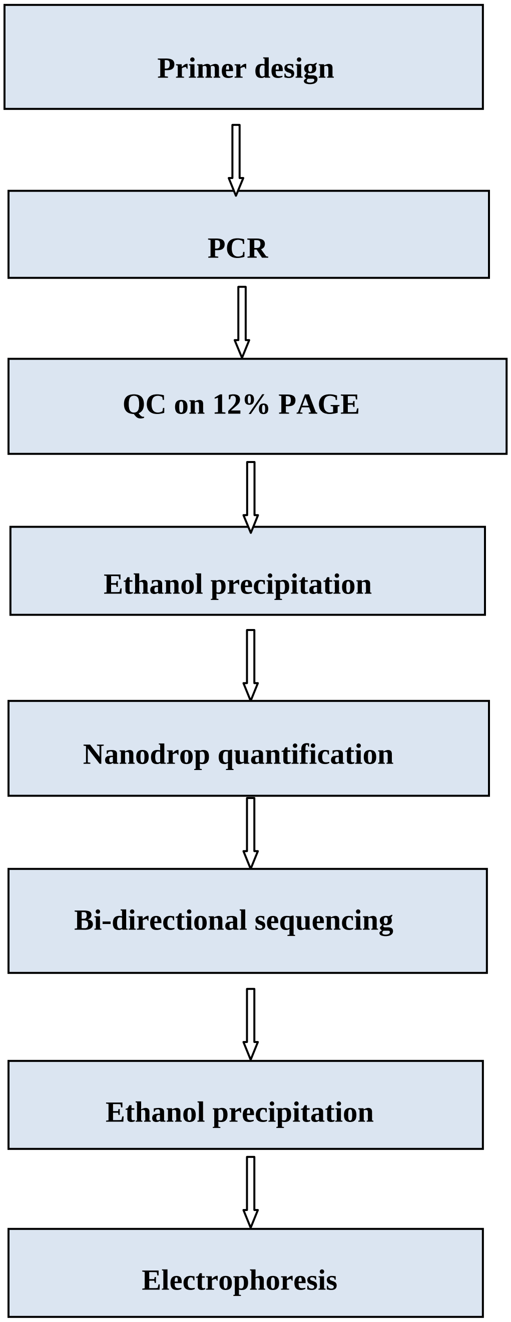


Figure 3.1. The flow chart for sequencing assay

The electropherograms were analyzed with FinchTV (PerkinElmer/ Geospiza), by which we were able to resolve the nucleotides originally shown as 'ambiguous' (highlighted in red below); several examples are given. The small grey letters above the alignment represent either the original majority call from the ABI software or minor known alleles. The nucleotides in green are the replacement calls we made that agree with the known major alleles, and the ‘C’ nucleotide in blue is the location of the dbSNP-characterized variant (present in all of our samples).

The nucleotides shown in magenta highlight the newly identified mutation. Text labels include F and R for the forward and reverse reactions and HC# or ALS# for the healthy control and ALS sample number.

F-HC7; Score = 170 bits (188), Expect = 7e-48, Identities = 105/111 (95%), Gaps = 1/111 (1%), Strand=Plus/Plus.

```
                    a                       c        c     c  g
Query  20   ATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGGCTCCCGAGGAGCAC  78
            |||||||| ||||||||||||||||||||||| |||||||| ||||| || |||||||||
Sbjct  51   ATGGAGAAGATCTGGCACCACACCTTCTACAACGAGCTGCGCGTGGCCCCGGAGGAGCAC  110
```

*R-HC7:*Score = 187 bits (206), Expect = 1e-52, Identities = 114/121 (94%), Gaps = 0/121 (0%), Strand=Plus/Minus.

```
                    t           c  g     g        g
Query  13   AGCAGCACGGGGTGCTCCTCGGGAGCCACACGCAGCTCATTGTAGAAGGTGTGGTGCCAG  72
            |||||||| ||||||||||| || ||||| |||||||| |||||||||||||||||||||
Sbjct  121  AGCAGCACTGGGTGCTCCTCCGGGGCCACGCGCAGCTCGTTGTAGAAGGTGTGGTGCCAG  62


              c                                a
Query  73   ATTTTCTCCATGTCGTCCCAGTTGGTGACGATGCCGTGCTCAATGGGGTACTTCAGGGTC  132
            || |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  61   ATCTTCTCCATGTCGTCCCAGTTGGTGACGATGCCATGCTCAATGGGGTACTTCAGGGTC  2
```

F-HC7; Score = 170 bits (188), Expect = 7e-48, Identities = 105/111 (95%), Gaps =

1/111 (1%), Strand=Plus/Plus.

```
                     a                       c        c     c  g
Query  20    ATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGTGTGGCTCCCGAGGAGCAC  78
             |||||||| ||||||||||||||||||||||| |||||||| ||||| || |||||||||
 Sbjct  51   ATGGAGAAGATCTGGCACCACACCTTCTACAACGAGCTGCGCGTGGCCCCGGAGGAGCAC  110
```

*R-HC7:*Score = 187 bits (206), Expect = 1e-52, Identities = 114/121 (94%), Gaps = 0/121 (0%), Strand=Plus/Minus.

```
                     t           c  g     g        g
Query  13    AGCAGCACGGGGTGCTCCTCGGGAGCCACACGCAGCTCATTGTAGAAGGTGTGGTGCCAG  72
             |||||||| ||||||||||| || ||||| |||||||| |||||||||||||||||||||
Sbjct  121   AGCAGCACTGGGTGCTCCTCCGGGGCCACGCGCAGCTCGTTGTAGAAGGTGTGGTGCCAG  62
               c                                a
Query  73    ATTTTCTCCATGTCGTCCCAGTTGGTGACGATGCCGTGCTCAATGGGGTACTTCAGGGTC  132
             || |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  61    ATCTTCTCCATGTCGTCCCAGTTGGTGACGATGCCATGCTCAATGGGGTACTTCAGGGTC  2
```

F-ALS1: Score = 188 bits (208), Expect = 3e-53,Identities = 116/121 (96%), Gaps = 2/121 (2%), Strand=Plus/Plus.

```
                   -  -                      c              c
Query  12    CTGGGACGACATGGAGAAGATCTGGCACCACACCTTCTACAATGAGCTGCGCGTGGCTCC  69
             |||||| || |||||||||||||||||||||||||||||||| |||||||||||||| ||
Sbjct  41    CTGGGACGACATGGAGAAGATCTGGCACCACACCTTCTACAACGAGCTGCGCGTGGCCCC  100
```

R-ALS1: Score = 179 bits (198), Expect = 6e-50, Identities = 114/121 (94%), Gaps = 2/121 (2%), Strand=Plus/Minus.

```
                   -          -     a     a        g
Query  11    AGCAGCACTGGGTGCTCCTCCGGGGCCACACGCAGCTCATTGTAGAAGGTGTGGTGCCAG  68
```

```
             |||||| |||||||||| ||||| ||||| |||||||| |||||||||||||||||||||
Sbjct  121   AGCAGCACTGGGTGCTCCTCCGGGGCCACGCGCAGCTCGTTGTAGAAGGTGTGGTGCCAG  62

               t                                  a
Query  69    ATCTTCTCCATGTCGTCCCAGTTGGTGACGATGCCGTGCTCAATGGGGTACTTCAGGGTC  128
             || |||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  61    ATCTTCTCCATGTCGTCCCAGTTGGTGACGATGCCATGCTCAATGGGGTACTTCAGGGTC  2
```

In 9 of the 10 samples analyzed, the cytosine in position 350 of the *ACTG1* gene, in exon 3, is replaced by a thymine (c.350C-to-T):

TCTGGCACCACACCTTCTACAA[C/**T**]GAGCTGCGCGTGGCCCCGGAGGAGCAC

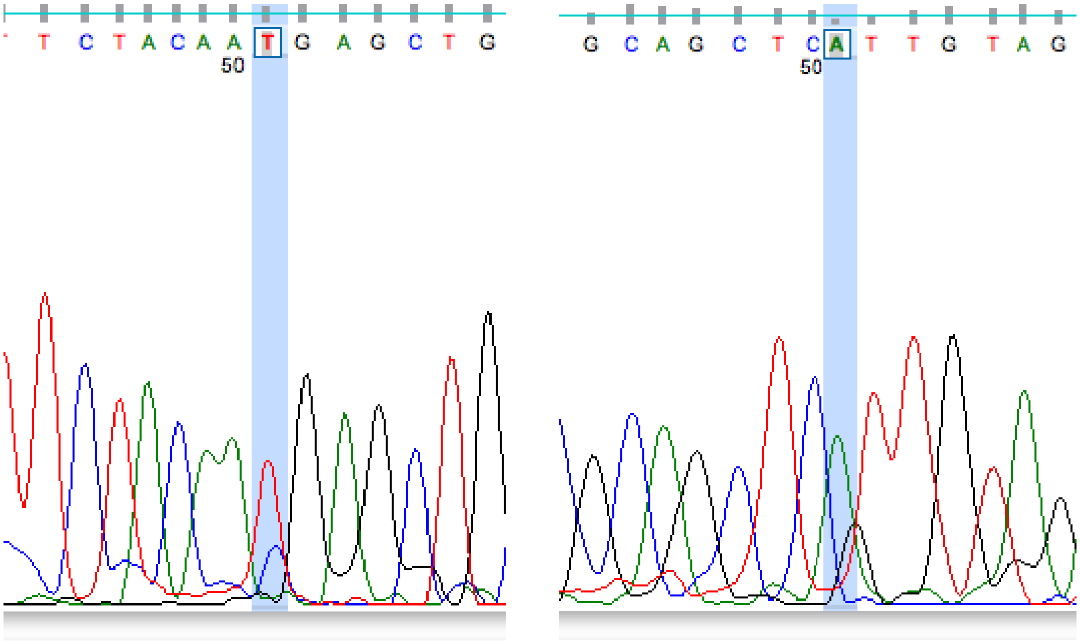
To our knowledge, the SNP has not been reported in the literature. It cannot be associated with the disease state since is present in both sample classes. To validate the amino acid change (ACG→ATG equivalent to THR→MET), if present, further protein assays are necessary, for which we do not have materials at present. Selected screenshots showing the presence of this mutation are given in Figure 3.2.

The results for the rest of the genes, by the respective exons (see Table 3.2) are presented below. The colors follow the scheme described above.

FTH1, with the target sequence:

CCCCATAGCCGTGGGGTGACTTCCCTGGTCACCAAGGCAGTGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCCACTTAAGTTCTTTGATTTGTACCATTCCTTCAAATAAAGAAATTTGGTACCC**AGGTGTTGTCTTTGAGGTCTTGGG**

For most of the obtained sequences, the identity with the target was between 96 - 99%, of which 6 were perfect matches (100% identity); thus no SNP was present.



(a) (b)

Figure 3.2 Electropherograms showing the presence of a novel mutation (the highlighted section) in forward (a) and reversed sequencing (b) for sample HC7.

F-HC7: Score = 232 bits (256), Expect = 3e-66, Identities = 129/130 (99%), Gaps = 0/130 (0%), Strand=Plus/Plus.

```
                                 -

Query  13    TGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCCACTTAA  72
             |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||
Sbjct  41    TGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCCACTTAA  100
```

R-HC7: Score = 208 bits (230), Expect = 3e-59, Identities = 122/126 (97%), Gaps = 1/126 (1%), Strand=Plus/Minus.

```
               N    N-      N

Query  18    TGAAGGAATGGTACAAATCAAAGAACTTAAGTGGATGTTTTGGTACAACTTATAGAAAAG  76
             || ||||  |||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  126   TGAAGGAATGGTACAAATCAAAGAACTTAAGTGGATGTTTTGGTACAACTTATAGAAAAG  67
```

F-ALS1: Score = 242 bits (268), Expect = 6e-69, Identities = 135/136 (99%), Gaps = 0/136 (0%), Strand=Plus/Plus.

```
               N

Query  9     AGGCAGTGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCC  68
             || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  35    AGGCAGTGCATGCATGTTGGGGTTTCCTTTACCTTTTCTATAAGTTGTACCAAAACATCC  94
```

R-ALS1: Score = 226 bits (250), Expect = 5e-64, Identities = 131/134 (98%), Gaps = 2/134 (1%), Strand=Plus/Minus.

```
                          ga

Query  9     TCTTTATTTGAGG--ATGGTACAAATCAAAGAACTTAAGTGGATGTTTTGGTACAACTTA  66
             ||||||||||| |  |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  134   TCTTTATTTGAAGGAATGGTACAAATCAAAGAACTTAAGTGGATGTTTTGGTACAACTTA  75
```

For *B2M*-exon 2, the target sequence that includes the Agilent probe is:

GTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGAGAGAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTACACTGAATTCACCCCCACTGAAAAAGATGAGT**ATGCCTGCCGTGTGAACCATGT**

This sequence alignment with the oligonucleotide obtained after sequencing shows a very good identity (96 - 99 %) with the exon, and the visual inspection of the electropherogram did not suggest the presence of any sequence variants. Below we present selected examples.

F-HC12: Score = 219 bits (242), Expect = 2e-62, Identities = 128/130 (98%), Gaps = 2/130 (2%), Strand=Plus/Plus.

```
                        -       -

Query  16    ATGGAGAGAGAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTT  73
             ||||||||||| ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  47    ATGGAGAGAGAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTT  106
```

R-HC12: Score = 244 bits (270), Expect = 1e-69, Identities = 139/140 (99%), Gaps = 1/140 (1%), Strand=Plus/Minus.

```
                       -
Query  9     GTGGGGGTGAATTCAGTGTAGTACAAGAGATAGAAAGACCAGTCCTTGCTGAAAGACAAG  67
             |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  140   GTGGGGGTGAATTCAGTGTAGTACAAGAGATAGAAAGACCAGTCCTTGCTGAAAGACAAG  81
```

F-ALS8: Score = 219 bits (242), Expect = 2e-62, Identities = 128/130 (98%), Gaps = 2/130 (2%), Strand=Plus/Plus.

```
                        -       -
Query  15    ATGGAGAGAGAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTT  72
             ||||||||||| ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  47    ATGGAGAGAGAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTT  106
```

R-ALS8: Score = 208 bits (230), Expect = 8e-59, Identities = 132/138 (96%), Gaps = 4/138 (3%), Strand=Plus/Minus.

```
                    C                    -    -                 -
Query  8     GGGGGTGAATTCAGTGTAGTACAAGAGATAGAAAGACCAGTCCTTGCTGAAAGACAAGTC  64
             ||||||| |||||||||||||||||||| |||| ||||||||||||||||| ||||||||
Sbjct  138   GGGGGTGAATTCAGTGTAGTACAAGAGATAGAAAGACCAGTCCTTGCTGAAAGACAAGTC  79
```

For *ACTG1*- exon 5, with SNP ID: rs28999112 and corresponding sequence:

GGAATCTTGCGGCATCCACGAGACCA**[C/T]**CTTCAACTCCATCATGAAGTGTGAC;

the target sequence was designed to incorporate the mutation:

GTATGGAATCTTGCGGCATCCACGAGACCA**C**CTTCAACTCCATCATGAAGTGTGACGTGGACATCCGCAAAGACCTGTACGCCAACACGGTGCTGTCGGGCGGCACCACCATGTACCCGGGCATTGCCGACA**GGATGCAGAAGGAGATCACC**

In our samples no polymorphism was seen as all of them contained the ancient allele (C) and 97-100 % identities were observed. Selected examples are given here.

F-HC3: Score = 185 bits (204), Expect = 3e-52, Identities = 109/112 (97%), Gaps =

1/112 (1%), Strand=Plus/Plus.

```
                    a
Query  9    CATCATGA-GTGTGACGTGGACATCCGCAAAGACCTGTACGCCAACACGGTGCTGTCGGG  67
            |||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  41   CATCATGAAGTGTGACGTGGACATCCGCAAAGACCTGTACGCCAACACGGTGCTGTCGGG  100

                           c           c
Query  68   CGGCACCACCATGTATCCGGGCATTGCTGACAGGATGCAGAAGGAGATCACC  119
            ||||||||||||||| ||||||||||| ||||||||||||||||||||||||
Sbjct  101  CGGCACCACCATGTACCCGGGCATTGCCGACAGGATGCAGAAGGAGATCACC  152
```

R-HC3: Score = 187 bits (206), Expect = 8e-53, Identities = 108/110 (98%), Gaps = 1/110 (1%), Strand=Plus/Minus

```
              N
Query  11   GGTGGTGC-GCCCGACAGCACCGTGTTGGCGTACAGGTCTTTGCGGATGTCCACGTCACA  69
            || ||||| |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  110  GGTGGTGCCGCCCGACAGCACCGTGTTGGCGTACAGGTCTTTGCGGATGTCCACGTCACA  51


Query  70   CTTCATGATGGAGTTGAAGGTGGTCTCGTGGATGCCGCAAGATTCCATAC  119
            ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  50   CTTCATGATGGAGTTGAAGGTGGTCTCGTGGATGCCGCAAGATTCCATAC  1
```

F-ALS2: Score = 190 bits (210), Expect = 1e-53, Identities = 110/112 (98%), Gaps = 1/112 (1%), Strand=Plus/Plus.

```
                    -
Query  10   CATCATGAAGTGTGACGTGGACATCCGCAAAGACCTGTACGCCAACACGGTGCTGTCGGG  68
            |||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  41   CATCATGAAGTGTGACGTGGACATCCGCAAAGACCTGTACGCCAACACGGTGCTGTCGGG  100

                           c
Query  69   CGGCACCACCATGTATCCGGGCATTGCCGACAGGATGCAGAAGGAGATCACC  120
            ||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  101  CGGCACCACCATGTACCCGGGCATTGCCGACAGGATGCAGAAGGAGATCACC  152
```

R-ALS2: Score = 192 bits (212), Expect = 5e-54, Identities = 110/111 (99%), Gaps =

1/111 (1%), Strand=Plus/Minus.

```
                     -
Query  11   TGGTGGTGCCGCCCGACAGCACCGTGTTGGCGTACAGGTCTTTGCGGATGTCCACGTCAC  69
            ||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  111  TGGTGGTGCCGCCCGACAGCACCGTGTTGGCGTACAGGTCTTTGCGGATGTCCACGTCAC  52


Query  70   ACTTCATGATGGAGTTGAAGGTGGTCTCGTGGATGCCGCAAGATTCCATAC  120
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  51   ACTTCATGATGGAGTTGAAGGTGGTCTCGTGGATGCCGCAAGATTCCATAC  1
```

ACTG1 was also screened for the SNP reported in exon 6, rs104894547. Therefore, we designed the target sequence to include this mutation.

TGAGGCTAGCATGAGGTGTGTGCATTTGCCAGGGGCAAATTTCTATTCTCAATTAACCCATGCAGCAAATGCTACGCATCTGCTGAGTCCGTTTAGAAGCATTTGCGGTGG**A**CGATGGAGGGGCCCGACTCGTCGTACTCCTGCTTGCT**AATCCACATCTGCTGGAAGG**

F-HC14: Score = 188 bits (208), Expect = 8e-53, Identities = 119/128 (93%), Gaps = 4/128 (3%), Strand=Plus/Plus.

```
                        a-
Query  13   TCTATTCTCATT--ACCCATGCAGCAAATGCTACGCATCTGCTGAGTCCGTTTAGAANAN  70
            |||||||||| |  |||||||||||||||||||||||||||||||||||||||||||   
Sbjct  42   TCTATTCTCAATTAACCCATGCAGCAAATGCTACGCATCTGCTGAGTCCGTTTAGAAGCA  101


Query  71   T--GCGGTGGACGATGGAGGGGCCCGACTCGTCGTACTCCTGCTTGCTAATCCACATCTG  128
            |  |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  102  TTTGCGGTGGACGATGGAGGGGCCCGACTCGTCGTACTCCTGCTTGCTAATCCACATCTG  161
```

R-HC14: Score = 230 bits (254), Expect = 2e-65, Identities = 132/135 (98%), Gaps = 1/135 (1%), Strand=Plus/Minus.

```
                 N N
```

```
Query  9     GACGAAGTCGGGCCCCTCCATCGTCCACCGCAAATGCTTCTAAACGGACTCAGCAGATGC  68
             ||||| | ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  134   GACGA-GTCGGGCCCCTCCATCGTCCACCGCAAATGCTTCTAAACGGACTCAGCAGATGC  76
```

F-ALS8: Score = 223 bits (246), Expect = 4e-63, Identities = 129/132 (98%), Gaps = 1/132 (1%), Strand=Plus/Plus.

```
                 N         -
Query  10    AATTTCTATTCTCAATTAACCCATGCAGCAAATGCTACGCATCTGCTGAGTCCGTTTAGA  68
             |||| ||||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  38    AATTTCTATTCTCAATTAACCCATGCAGCAAATGCTACGCATCTGCTGAGTCCGTTTAGA  97

                           g
Query  69    AGCATTTGCGGTGGACGATGGAGGGGCCCGACTCGTCGTACTCCTGCTTGCTAATCCACA  128
             |||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  98    AGCATTTGCGGTGGCCGATGGAGGGGCCCGACTCGTCGTACTCCTGCTTGCTAATCCACA  157
```

R-ALS8: Score = 219 bits (242), Expect = 2e-62, Identities = 128/131 (98%), Gaps = 1/131 (1%), Strand=Plus/Minus.

```
                     c                    -
Query  11    GAGTCGGGGCCCTCCATCGTCCACCGCAAATGCTTCTAAACGGACTCAGCAGATGCGTAG  69
             |||||||| |||||||||| ||||||||| ||||||||||||||||||||||||||||||
Sbjct  131   GAGTCGGGCCCCTCCATCGGCCACCGCAAATGCTTCTAAACGGACTCAGCAGATGCGTAG  72
```

For *RPL21*, the target sequence was designed around the Agilent probe sequence, where no SNPs were previously reported.

AGTTGTTCCTTTGGCCACATATATGCGAATCTATAAGAAAGGTGATATTGTAGACATCAAGGGAATGGGTACTGTTCAAAAAGGAATGCCCCACAAGTGTTACCATGGCAAAACTGGAAGAGTCTACAATGTTACCCAGCA**TGCTGTTGGCATTGTTGTAAA**

Sequencing assays did not show any new sequence variant, the two-sequence alignment providing identities in 90 – 100 % interval. Selections for two samples, in both forward and reversed directions, are presented.

F-HC10: Score = 214 bits (236), Expect = 7e-61, Identities = 122/123 (99%), Gaps =

1/123 (1%), Strand=Plus/Plus.

```
Query  10    AGGTGATATTGTAGACATCAAGGGAATGGGTACTGTTCAAAAAGGAATGCCCCACAAGTG  68
             |||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  40    AGGTGATATTGTAGACATCAAGGGAATGGGTACTGTTCAAAAAGGAATGCCCCACAAGTG  99
```

R-HC10: Score = 208 bits (230), Expect = 3e-59, Identities = 119/120 (99%), Gaps = 1/120 (1%), Strand=Plus/Minus.

```
                  c
Query  14    TCTTC-AGTTTTGCCATGGTAACACTTGTGGGGCATTCCTTTTTGAACAGTACCCATTCC  72
             ||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   TCTTCCAGTTTTGCCATGGTAACACTTGTGGGGCATTCCTTTTTGAACAGTACCCATTCC  62
```

F-ALS6: Score = 201 bits (222), Expect = 4e-57, Identities = 115/116 (99%), Gaps = 1/116 (1%), Strand=Plus/Plus.

```
Query  15    ATTGTAGACATCAAGGGAATGGGTACTGTTCAAAAAGGAATGCCCCACAAGTGTTACCAT  73
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  47    ATTGTAGACATCAAGGGAATGGGTACTGTTCAAAAAGGAATGCCCCACAAGTGTTACCAT  106
```

R-ALS6: Score = 179 bits (198), Expect = 2e-50, Identities = 112/118 (95%), Gaps = 2/118 (2%), Strand=Plus/Minus.

```
                           at
Query  22    CAGTTTTGCCATGGGGAACACTTGTGGGGGCATTCCTTTTTGAACAGTACCCATTCCCCT  81
             ||||||||||||||  ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  116   CAGTTTTGCCATGGT-AACACTTGTGGGG-CATTCCTTTTTGAACAGTACCCATTCCCTT  59



             g                        t
Query  82    GATGTCTACAATATCACCTTTCTTAGAGATTCGCATATATGTGGCCAAAGGAACAACT  139
             ||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  58    GATGTCTACAATATCACCTTTCTTATAGATTCGCATATATGTGGCCAAAGGAACAACT  1
```

For *RPS10*, the target sequence comprising the Agilent probe is:

CTCACAAGAGGGGAAGCTGACAGAGATACCTACAGACGGAGTGCTGTGCCACCTGGTGCCGACAAGAAAGCCGAGGCTGGGGCTGGGTCAGCAACCGAATTCCAGTTTAGAGGCGGATTTGGTCGTGGACG**TGGTCAGCCACCTCAGTAAA**

Although the presence of any sequence variant was not confirmed, we selected one of each HC / sALS sample to exemplify the sequence alignment query-to-target obtained with bl2seq. In general, 96 – 100% identities were found.

F-HC14: Score = 188 bits (208), Expect = 2e-53, Identities = 109/111 (98%), Gaps = 1/111 (1%), Strand=Plus/Plus.

```
                     a                  a
Query  11   GTGCTGTGCCACCTGGTGCCGACAAGAA-GCCGAGGCTGGGGCTGGGTCAGCAACCGAAT  69
            ||||||||| |||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  41   GTGCTGTGCCACCTGGTGCCGACAAGAAAGCCGAGGCTGGGGCTGGGTCAGCAACCGAAT  100
```

R-HC14: Score = 187 bits (206), Expect = 9e-53, Identities = 111/114 (97%), Gaps = 2/114 (2%), Strand=Plus/Minus.

```
              N     a
Query  8    GCCTCTAA-CTGGA-TTCGGTTGCTGACCCAGCCCCAGCCTCGGCTTTCTTGTCGGCACC  65
            || ||||| ||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  114  GCCTCTAAACTGGAATTCGGTTGCTGACCCAGCCCCAGCCTCGGCTTTCTTGTCGGCACC  55
```

F-ALS8: Score = 181 bits (200), Expect = 3e-51, Identities = 108/111 (97%), Gaps = 2/111 (2%), Strand=Plus/Plus.

```
                     N-                 -
Query  13   GTGCTGTGCCACCTGGTGCCGACAAGAAAGCCGAGGCTGGGGCTGGGTCAGCAACCGAAT  70
            |||||||||  ||||||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  41   GTGCTGTGCCACCTGGTGCCGACAAGAAAGCCGAGGCTGGGGCTGGGTCAGCAACCGAAT  100
```

R-ALS8: Score = 183 bits (202), Expect = 1e-51, Identities = 105/106 (99%), Gaps = 1/106 (1%), Strand=Plus/Minus.

```
                  a
Query  15   ACTGGA-TTCGGTTGCTGACCCAGCCCCAGCCTCGGCTTTCTTGTCGGCACCAGGTGGCA  73
            |||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  106  ACTGGAATTCGGTTGCTGACCCAGCCCCAGCCTCGGCTTTCTTGTCGGCACCAGGTGGCA  47
```

The gene in which we had the most interest, *SLC25A3* was searched for the presence of SNPs in exon 3A (a previously reported mutation) and also in exon 8 where the Agilent probe used in microarray analysis was located. In regards to exon 3A, after repeated trials, where no probe signal was received from the DNA Analyzer, we concluded that this particular exon is not expressed in PBLs, but most probably is tissue specific. This result is concordant with work of Mayr *et al*. [102] and Shah *et al*.[108], who showed that exon 3A is expressed only in muscle, heart and thyroid tissues.

For exon 8, only 80% of the sequencing reactions were successful, as for four of the samples no signal was detected. Since other sequencing from these samples were successful this may also indicate variant isoforms. In the sequences present no polymorphisms were detected. The designed target sequence is:

CTCCGTGAAGGTCTACTTCAGACTTCCTCGCCCTCCTCCACCCGAGATGCCAGAGTCTCTGAAGAAGAAGCTTGGGTTAACTCAGTAGTTAGATCAAAGCAAATGTGGACTGAATCTGCTTGTTGATCAG**TGTTGAAGAAAGTGCAAAAGGA**

A single example from a diseased sample is shown, since the control samples did not yield usable data.

F-ALS1: Score = 179 bits (198), Expect = 1e-50, Identities = 111/115 (97%), Gaps = 3/115 (3%), Strand=Plus/Plus.

```
                  -                              g
Query  8    CCCGAGGATGCCAGAGTCTCTGAAGAAGAAGCTTGGG-TTAACTCAGTAGTTAGATCAAA  67
            |||||| |||||||||||||||||||||||||||||| ||||||||||||||||||||||
Sbjct  41   CCCGAG-ATGCCAGAGTCTCTGAAGAAGAAGCTTGGG-TTAACTCAGTAGTTAGATCAAA  98
```

```
                     cc
Query  68    GCAAATGTGG-GACTGAATCTGCTTGTTGATCAGTGTTGAAGAAAGTGCAAAAGGA  122
             ||||||||  |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  99    GCAAATGTG-GACTGAATCTGCTTGTTGATCAGTGTTGAAGAAAGTGCAAAAGGA  152
```

R-ALS1: Score = 185 bits (204), Expect = 3e-52, Identities = 109/112 (97%), Gaps = 1/112 (1%), Strand=Plus/Minus.

```
                  c                          g
Query  10    CAGTC-ACATTTGCTTTGATCTAACTACTGAGTTAACCCAAGCTTCTTCTTCAGAGACTC  68
             ||||| |||||||||||||||||||||||||| |||||||||||||||||||||||||||
Sbjct  112   CAGTCCACATTTGCTTTGATCTAACTACTGAGTTAACCCAAGCTTCTTCTTCAGAGACTC  53

                                                    g
Query  69    TGGCATCTCGGGTGGAGGAGGGCGAGGAAGTCTGAAGTAAACCTTCACGGAG  120
             ||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  52    TGGCATCTCGGGTGGAGGAGGGCGAGGAAGTCTGAAGTAGACCTTCACGGAG  1
```

PCR amplification failed to yield products for *B2M (*exon 1) and for *TARDBP (*exon 8) in the regions selected. Since the primers and conditions had been reported as successful previously, and since microarray detection was well within the reliable range, this outcome is most probably due to sample degradation.

3.4 Conclusions

Direct sequencing was performed to screen for possible mutations in selected exons of differentially expressed genes determined by the LO-BaFL pipeline, for the five healthy controls and five diseased (sALS) samples for which we had good quality cDNA. In the majority of the products good quality sequence that matched database records was obtained and no polymorphisms were found. In one case we identified a novel mutation in exon 3 of *ACTG1* gene, c.350 C-to-T, as follows:

TCTGGCACCACACCTTCTACAA[**C**/**T**]GAGCTGCGCGTGGCCCCGGAGGAGCAC

Previous studies on this *actin, gamma 1* gene showed the presence of several allelic variants related to hearing loss [109-112] (see Table 3.3).

Table 3.3 Allelic variants for *ACTG1*, from OMIM database.

Number	**Phenotype**	**Mutation**	**dbSNP**
.0001	Deafness, autosomal dominant 20	THR89ILE	rs28999111
.0002	Deafness, autosomal dominant 20	LYS118MET	rs104894544
.0003	Deafness, autosomal dominant 20	PRO332ALA	rs104894545
.0004	Deafness, autosomal dominant 20	PRO264LEU	rs104894546
.0005	Deafness, autosomal dominant 20	THR278ILE	rs28999112
.0006	Deafness, autosomal dominant 20	VAL370ALA	rs104894547

Research on animal models that reduced expression levels of the gene correlate with human muscle weakness and myopathies [98], not surprising given the role of *ACTG1* in skeletal muscle development. However, since we found this variant in what are labeled ‘healthy controls’, it does not correlate with sALS.

Further studies, at the protein level, are needed to shed light into possible protein changes. Since the exons show evidence of differential expression but no local SNP was present, the change is due to turnover or regulation, not to disturbed binding to the DNA probe from an uncharacterized SNP. So it might be worthwhile to sequence each of these genes in their entirety, including the regulatory regions. The changes might be due to changes in the transcription factors that bind those regions - that should be indicated by changes in other genes controlled by those factors. That is one type of pathway analysis, which is covered in the next chapter.

CHAPTER 4: PATHWAY ANALYSIS

4.1 Introduction

As discussed in Chapter 3, Amyotrophic Lateral Sclerosis is a complex disease whose pathology and etiology have not been deciphered. While biomarker discovery can be pursued using patterns of expression-level change, described in Chapter 2, our eventual goal must be to understand the mechanism underlying its biology. Since this is a motor-neuron disease it is unlikely that the circulating blood cells are directly affected by the causative agent, but subsets of these cells respond to signals from decaying cells and these signals may provide clues to the original source of the pathology. Having confirmed increased levels of expression of marker genes, the cause for such increases was investigated. One possibility is expression of specific alleles in the patient. Therefore, a search for sequence variants by direct sequencing was performed, described in Chapter 3; since we could not demonstrate the presence of a specific SNP linked to the disease in the candidate genes, we hypothesize that the observed changes in expression level are due to other sources, e.g. transcription factors or other regulatory molecules elicited in the disease process. Such factors nearly always affect multiple genes, so their presence can be inferred by looking for a concerted suite of effects (not always in the same direction or to the same extent) in pathways sharing the regulatory element. This type of analysis is covered by computational systems biology methods, e.g. pathway and network analysis.

Identification of ‘signature’ networks that co-regulate the genes of interest often provide insights into the bigger picture of processes and development, by transitioning from examining single molecules to global states [113, 114]. Such states may include descriptions of the biologically relevant interactions between genes, proteins and compounds; the interactions are grouped into functional structures (described with specific ontologies) such as metabolic signaling, transcription factor interactions, regulatory networks and functional roles [124-128]. Collectively known as systems biology, analysis on this level is an important tool for discovery because a single phenotype may result from errors in any one of the many component elements of such pathways or networks [115].

Pathway analysis in previous sALS studies using human motor cortex samples emphasized the involvement of defense responses, cytoskeletal development, and mitochondrial and proteosomal dysfunction in ALS pathology [116]. More recently, Kudo *et al*. (2010) performed Gene Ontology analysis using DAVID [117, 118] on tissue microarrays from human postmortem spinal cord tissues from subjects with sALS that revealed associations between the biological processes corresponding to motor neurons and surrounding cells and protein modification/posphorylation, signaling, muscle contraction regulation, stress responses, immune responses and cell communications [119]. We would predict that immune responses, communication and signaling and stress responses would be propagated to the PBLs, the question being whether the responses are disease-specific or a general systems-alert.

Despite increasing evidence that peripheral blood can be a powerful source of biomarkers for neurological diseases, very few studies have attempted to use sALS blood samples. Results from studies of patients with multiple sclerosis [120], Alzheimer's disease [121], Huntington's disease [122], and a few with ALS [119, 123-125] have proved that the peripheral blood transcriptome is a reliable source for biomarker detection.

More specifically for the current discussion, network analysis using weighted gene co-expression method on peripheral blood from ALS by Saris *et al.* (2009) found several significant pathways related to sALS, i.e. post-translational modification, infection mechanism, neurological disorder (Huntington), genetic disorder, skeletal and muscular disorder and inflammatory disease [123]. Zhang *et al.* (2011) show a strong association between aberrant activity of monocytes circulating in peripheral blood from patients with sALS and LPS (plasma endotoxin/lipopolysaccharide system) / TLR4 (toll-like receptor 4) pathways, suggesting that activation of monocytes /macrophages via these signaling pathways would affect the disease progression [126].

However, early studies have used whole blood or peripheral blood mononuclear cells. There is an increased literature support for potential differences between the disease-related signatures due to subpopulations of the cells, i. e. PBLs compared with monocytes and even subpopulations within the PBL grouping [127-129]. Therefore, it was suggested that, in order to detect disease-specific changes in transcription, it is necessary to profile purified leukocyte subsets [92, 127].

Taking this approach, an examination of microarray data from PBLs (a subpopulation of PBMCs) from subjects with sALS [75] showed alterations in the KEGG-designated ALS disease pathway, suggesting the propagation of gene expression changes first induced in brain and spinal cord tissue to cells in the circulating PBLs [116, 130]. We performed a pathway analysis using the genes identified as differentially expressed using the LO-BaFL pipeline (see Chapter 1), which differed from the TM4-based study described in [51] by a number of filters.

Note: since TM4 genes have been already extensively discussed in the pathway context elsewhere [75], we do not replicate those comments here but only highlight the points most relevant to our own findings.

4.2 Materials and Methods

In order to identify the pathways and regulatory elements possibly associated with the genes in our list, we conducted an analysis using the MetaCore™ program, version 6.0. This is composed of a suite of tools for gene set enrichment analysis, multi-experiment comparison, interactome analysis and biological network identification [115]. The canonical pathways and network maps were obtained from the manually curated GeneGo database (GeneGo Inc., St. Joseph, MI) which incorporates, for human cells, protein-protein, protein-DNA, and protein-compound interactions, as well as experimentally verified information on metabolic and signaling pathways [114]. Statistical tests (using hypergeometric distributions) assign to each pathway or network a corresponding *p-valu*e, Z-score and G-score, to assess their change from baseline, depending on the degree of saturation of the modeled set with the objects from the initial gene list [131, 132].

Networks with higher scores are considered more relevant to the phenotype, relative to the context presented in the specific data set. The *p-value* is corrected using the False Discovery Rate algorithm, which represents the probability that a given number of genes from the input list will match a certain number of gene nodes in the network [145, 146]. In this study, only pathways or networks with $p < 0.01$ were considered statistically significant.

4.3 Results and Discussions

As a first step, the entire set of sixty LO-BaFL differentially expressed genes (including those supported by qRT-PCR) with their corresponding intensities (Supplementary Material/ Data post filtering/ DE_genes_12_samples from Chapter 1) was uploaded in MetaCore for building the biological networks. The software returned twelve networks whose involvement met our criteria. The two most strongly supported pathways (G-score=42.8, *p*=1.44e-15) are involved in immune system signaling; the TWEAK gene (TNFSF12) via the TNF (tumor necrosis factor) receptor-associated factors 2 (TRAF2) or 5 (TRAF5). See Table 4.1 for Gene Ontology processes, and scores. To clarify the roles of these genes in the pathways, a GeneGo map of the network with the highlighted pathways is given in Figure 4.1. Other networks in which these genes participate are shown in Appendix A, Figure A1.

The TWEAK (TNFSF12/ Apo-3L) gene codes for a type II transmembrane protein, and is a member of the TNF superfamily. The gene is involved in immune regulation, induced cell death, and hence inflammation [147-151]. Although it is expressed in various tissues, the highest levels of expression have been found in brain tissue skeletal muscle, heart muscle and immune system cells [133]. It has been shown that the TWEAK signaling pathway has roles in apoptosis, proliferation, migration, angiogenesis, and inflammation [134].

Several genes (*ACTG1, DIA1, IRS2, JUNB*) from our list participate in signaling pathways, alone or in combination. A myriad of complex processes can be invoked; we provide some examples that we consider most likely to be relevant to ALS below. *ACTG1* participates in cytoskeleton remodeling by RhoGTPAse regulation of the actin cytoskeleton (Figure 4.2 (a)); in combination with *DIA1* it is a component in immune responses via *CCR3* signaling in the eosinophil pathway (Figure 4.2 (b)); in combination with *IRS2*, it has been shown to interact with alpha-6/beta-4 integrins in carcinoma progression (Figure 4.2 (c)) and it contributes to the regulation by growth factors of transport macropinocytosis (Figure 4.2 (d)). *JUNB* and *IRS2* are part of developmental growth hormone signaling pathways, working through the PI3K/AKT and MAPK cascades (Figure 4.2 (e)).

Our results are in good agreement with the work of Lederer *et al.* (2007), which showed the involvement of similar biological processes based on candidate genes derived from motor cortex samples from patients with sALS. The concurrence of gene expression results from PBLs to those of presumably more directly involved tissues is also suggested by the studies of Mougeot *et al.* (2011), perhaps not surprising where sensitive signaling cascades are involved. It does suggest that the microarray assay of PBLs, or derivative qPCR assays, are quite sensitive to the system-wide changes and that further cell sorting is not required.

Examination of the cross-section of functional networks (Table 4.1) by co-expression networks (Figure A1) shows the ribosomal proteins *RPS10, RPS15A* which are involved in translational regulation, but on further inspection, in Figure A1 (ii), these genes are in a set (*RPS10, RPS25, RPS15A, RPL21*), that, with the solute carrier transporter, *SLC25A3,* are part of an immune-response network, with the *HLA-A, HLA-B, HLA-C* proteins. Using a wider lens, members of the LO-BaFL gene set *PABC1, PTMA, SFRS3, DIA1, LAMR1, FTH1, B2M, SLC25A3, CCDC6, OAZ1* also interact to regulate immune responses, and other types of cellular responses including apoptosis, which is discussed below. Some of their known interactions are with integrin-type proteins having roles in cell adhesion and cell-surface mediated signaling, and specifically with Myosin II and similar smooth muscle-specific genes whose degeneration is characteristic of ALS. That is, the genes converge in pathways that play important roles in sALS pathology.

Some genes balance cell processes, especially ubiquitin. Ubiquitin helps up-regulate the subset of genes including *PTMA, JUNB, SCP1*, and *OAZ1*, seen in network 5 (Figure 2A (iv)), and *IRS2, JUNB* seen in network 9 (see Table 4.1; graphics are not shown), whose functions include essential metabolic processes and cellular development. But upon interaction with *PABPC1* and *JUNB,* ubiquitin leads to apoptosis, seen in network 10 (Table 4.1). Aberrant forms of the ubiquitin protein and the resultant apoptosis are present in patients with neurodegenerative diseases, e.g. Alzheimer's disease or Down's syndrome [135]. Although this effect is well documented in motor neurons it is not reported for circulating cells, however the indication that it is induced is reinforced by the upregulation of *IRS2, SPON2,* and *DIA1* shown in network 8 (see Table 4.1).

Table 4.1 List of biological networks formed by selection of candidate genes

No.	Network name	GO process	p-value	zScore	gScore
1.	DIA1, PSMA1, TWEAK(TNFSF12), Mindin, Actin cytoplasmic 2	nucleosome assembly (26.0%; 8.535e-19), chromatin assembly (26.0%; 1.612e-18), nucleosome organization (26.0%; 4.354e-18), chromatin assembly or disassembly (26.0%; 7.625e-18), protein-DNA complex assembly (26.0%; 7.625e-18)	1.44e-15	40.28	42.78
2.	PABPC1, PTMA, SFRS3, DIA1, Mindin	immune system process (56.0%; 3.046e-18), cellular response to organic substance (46.0%; 1.556e-15), leukocyte migration (26.0%; 6.161e-15), cellular response to chemical stimulus (48.0%; 4.130e-14), regulation of immune system process (40.0%; 4.884e-14)	1.44e-15	40.28	40.28
3.	RPS10, LAMR1, SLC25A3, SP1, mRNA intracellular	antigen processing and presentation of peptide antigen via MHC class I (29.2%; 4.791e-28), antigen processing and presentation of peptide antigen (29.2%; 7.266e-25), immune system process (62.5%; 2.174e-21), antigen processing and presentation (29.2%; 8.387e-21), regulation of immune response (45.8%; 1.071e-20)	4.06e-13	34.86	34.86
4.	Beta-2-microglobulin, FTH1, CCDC6, RELT, OAZ1	positive regulation of biological process (79.2%; 6.338e-21), regulation of response to stimulus (64.6%; 2.161e-18), regulation of immune system process (47.9%; 4.304e-18), regulation of cell death (54.2%; 6.116e-18), immune system process (56.2%; 1.045e-17)	4.61e-13	34.50	34.50
5.	Ubiquitin, PTMA, JunB, SCP1, OAZ1	positive regulation of macromolecule metabolic process (83.7%; 6.743e-37), positive regulation of metabolic process (83.7%; 2.406e-35), positive regulation of cellular metabolic process (79.6%; 6.578e-33), positive regulation of gene expression (69.4%; 8.569e-32), positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (69.4%; 1.226e-30)	1.20e-10	28.72	28.72
6.	PABPC1, SFRS3, Serglycin, PRR13, SCP1	response to organic substance (72.9%; 5.224e-24), response to endogenous stimulus (58.3%; 5.785e-22), system development (79.2%; 4.992e-20), response to chemical stimulus (79.2%; 9.925e-20), developmental process (85.4%; 1.048e-19)	1.20e-10	28.72	28.72

Table 4.1 (continued)

7.	CCDC6, SCP1, JunB, PTMA, p21	positive regulation of transcription from RNA polymerase II promoter (61.2%; 1.505e-31), regulation of transcription from RNA polymerase II promoter (69.4%; 2.128e-30), regulation of gene expression (91.8%; 1.360e-29), positive regulation of RNA metabolic process (65.3%; 4.858e-29), positive regulation of macromolecule metabolic process (73.5%; 7.358e-29)	2.45e-08	22.94	22.94
8.	IRS-2, Mindin, DIA1, SP1, ERK1/2	regulation of programmed cell death (71.4%; 3.975e-30), regulation of cell death (71.4%; 9.906e-30), regulation of apoptosis (69.4%; 9.656e-29), positive regulation of apoptosis (57.1%; 1.019e-28), positive regulation of programmed cell death (57.1%; 1.270e-28)	3.80e-06	17.16	17.16
9.	Ubiquitin, JunB, IRS-2, Bcl-2, TGF-beta 1	organ development (89.6%; 1.072e-32), positive regulation of cellular process (91.7%; 1.734e-31), positive regulation of biological process (93.8%; 2.355e-31), developmental process (97.9%; 4.718e-29), multicellular organismal development (95.8%; 7.379e-29)	3.80e-06	17.16	17.16
10.	Ubiquitin, PABPC1, JunB, Androgen receptor, iNOS	regulation of cell death (71.4%; 9.906e-30), positive regulation of cellular process (87.8%; 5.893e-29), regulation of apoptosis (69.4%; 9.656e-29), regulation of programmed cell death (69.4%; 1.285e-28), multi-organism process (67.3%; 6.064e-28)	3.80e-06	17.16	17.16
11.	JunB, OAZ1, Galanin, SET, PLAT (TPA)	response to organic substance (65.3%; 9.446e-20), positive regulation of cellular process (69.4%; 3.361e-17), positive regulation of biological process (71.4%; 6.569e-17), response to chemical stimulus (73.5%; 6.628e-17), response to steroid hormone stimulus (34.7%; 1.044e-14)	4.20e-04	11.39	11.39
12.	JunB, VEGF-A, NF-kB p50/p65, Androgen receptor, NF-kB	positive regulation of transcription from RNA polymerase II promoter (75.0%; 6.435e-22), positive regulation of macromolecule biosynthetic process (83.3%; 2.498e-21), positive regulation of transcription, DNA-dependent (79.2%; 1.079e-20), positive regulation of cellular biosynthetic process (83.3%; 1.683e-20), positive regulation of RNA metabolic process (79.2%; 2.343e-20)	1.49e-02	8.05	8.05

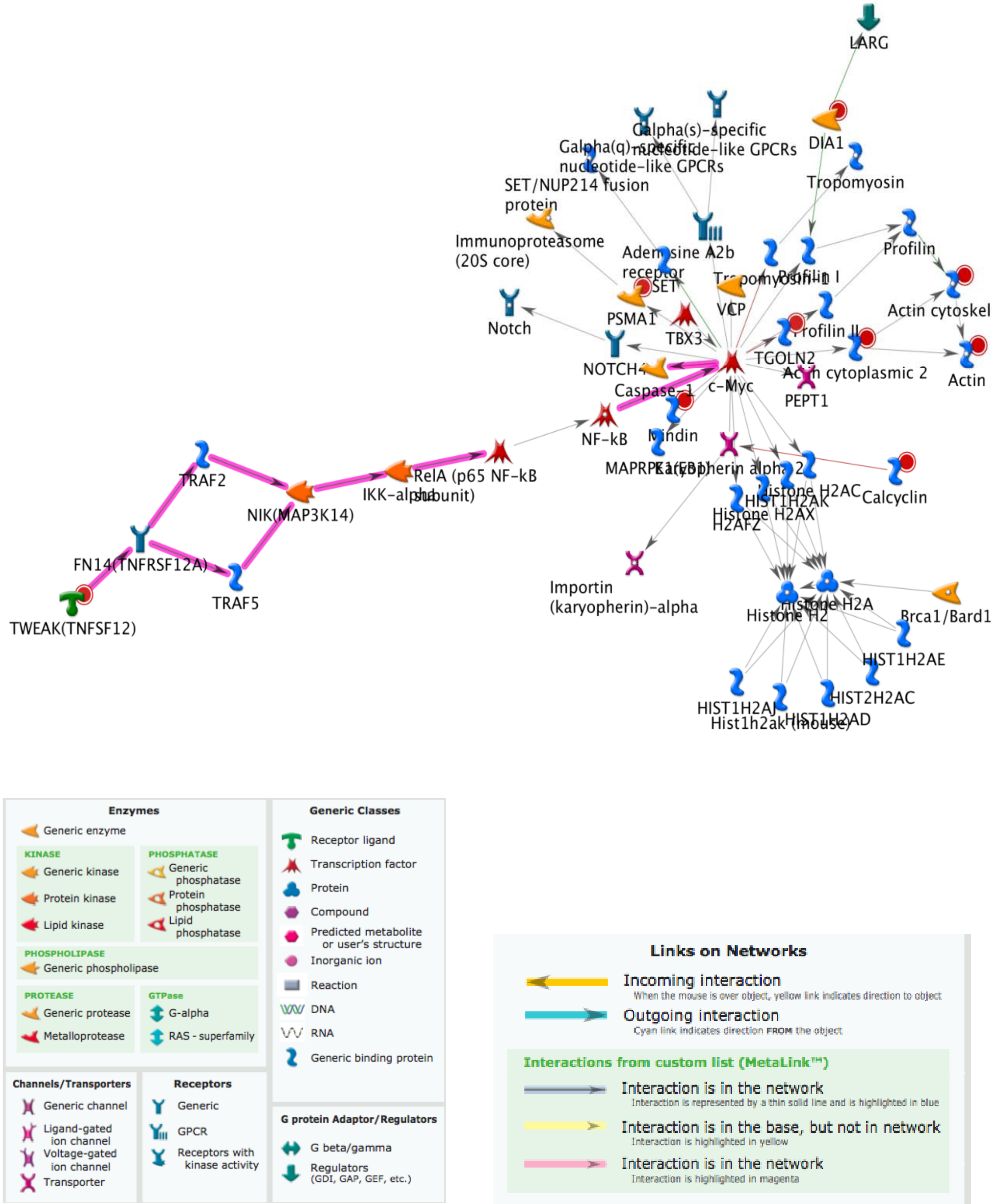
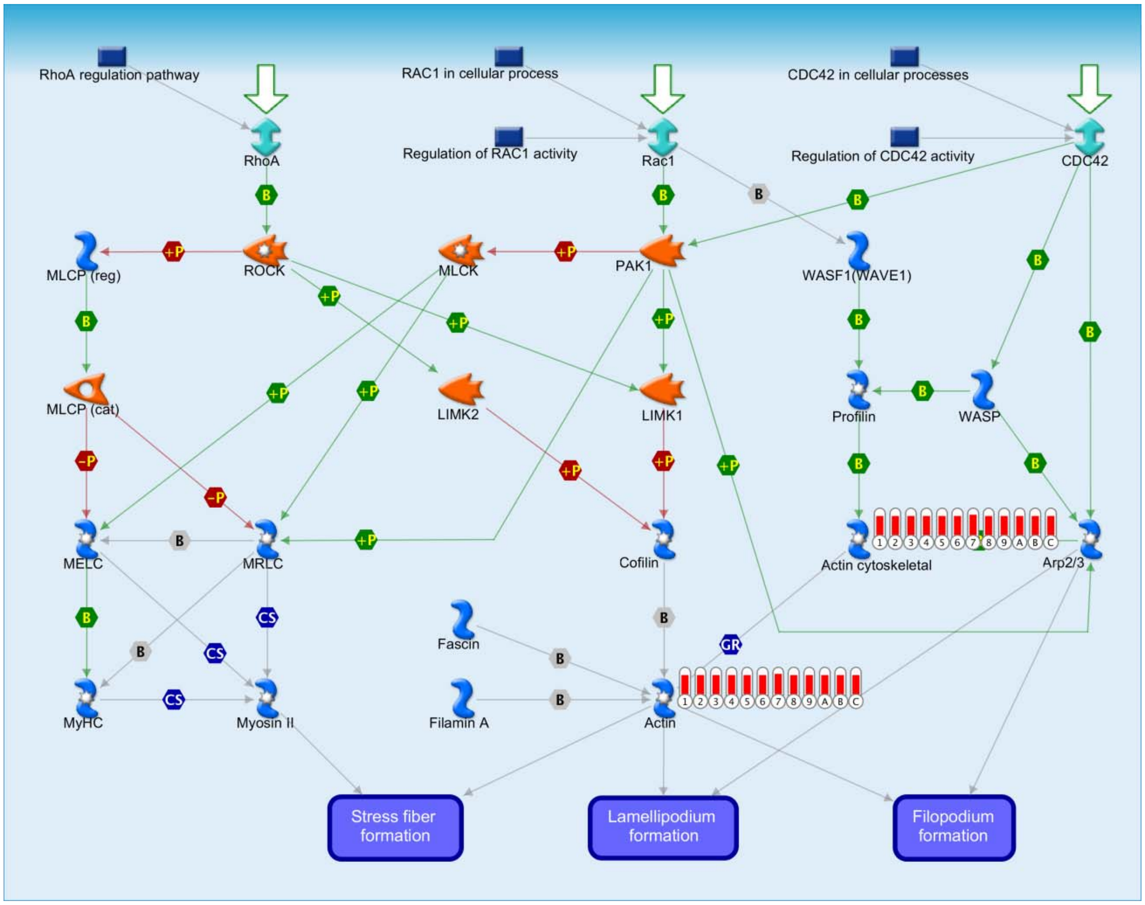
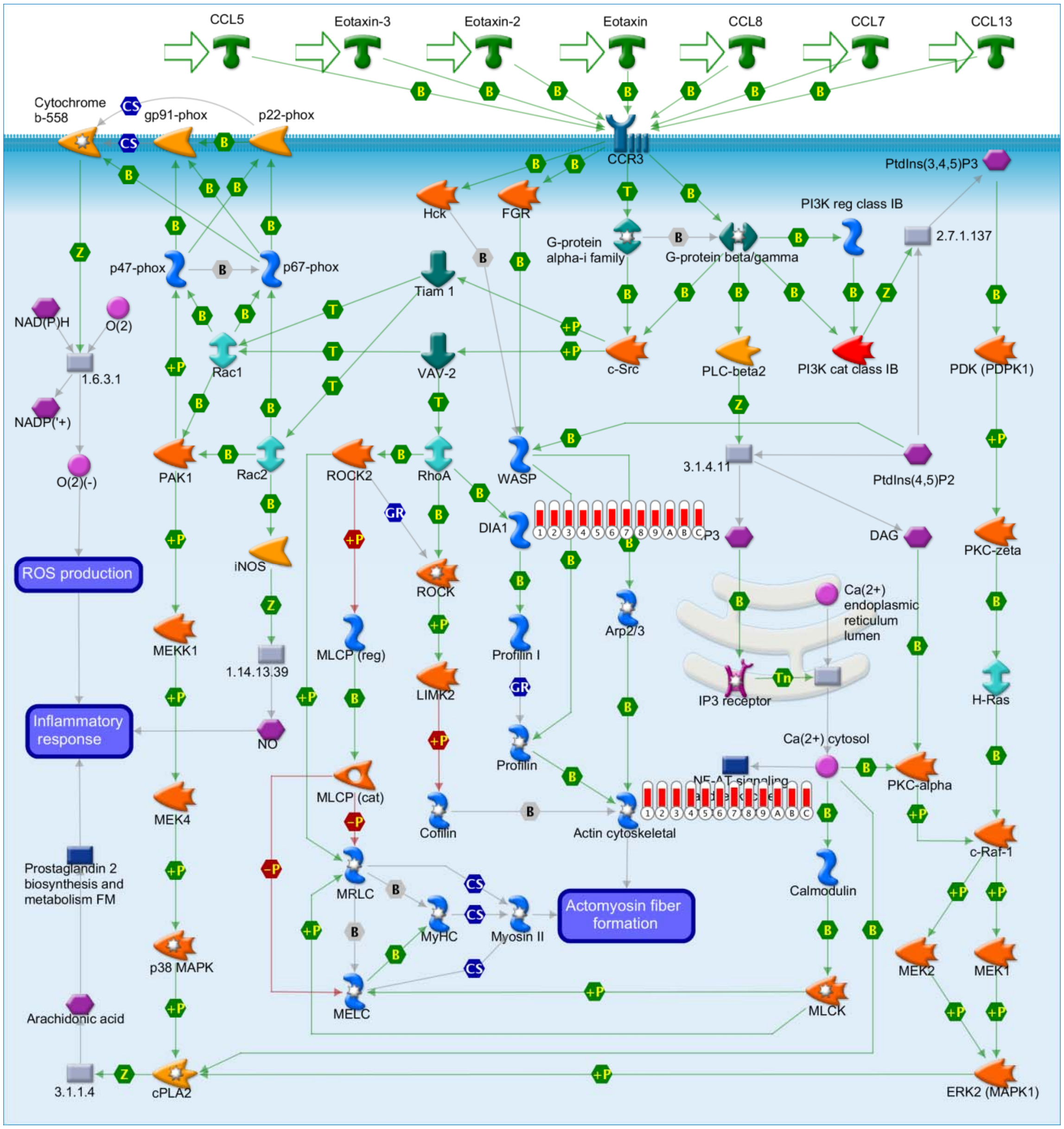


Figure 4.1 (*upper*) GenoGo map of networks with known interactions (highlighted lines). Red circles denote the genes from input list; (*lower*) Symbol legend (selections from MetaCore full legend).



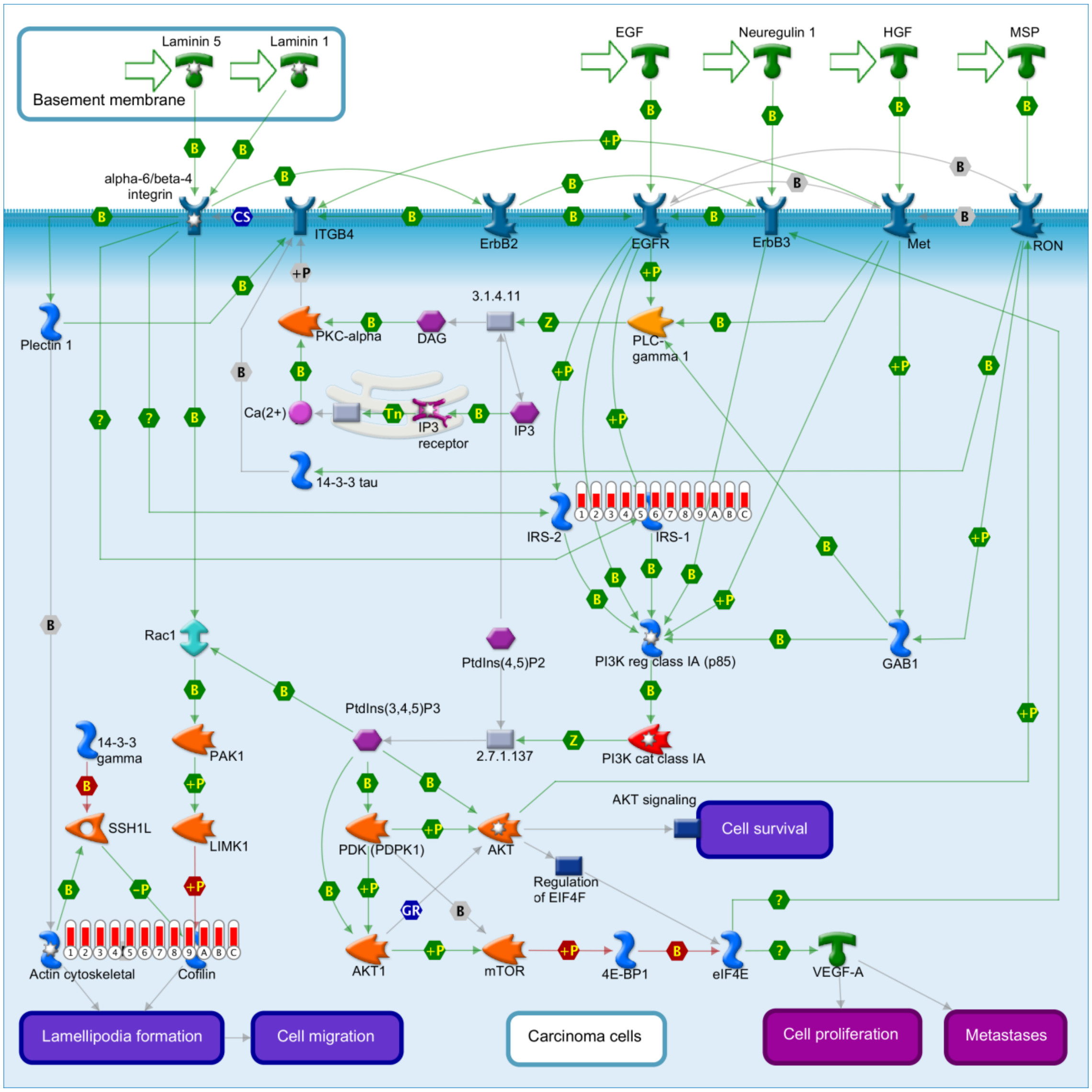
(a)

Figure 4.2. MetaCore analysis of signaling pathways in PBLs from sALS samples. Symbols are explained in the Figure 4.1 key. (a) cytoskeleton remodeling regulation of actin cytoskeleton by RhoGTPase; (b) immune responses via *CCR3* signaling in eosinophils; (c) role of alpha-6/beta-4 integrins in carcinoma progression; (d) transport macropinocytosis regulation by growth factors; (e) development growth hormone signaling, via PI3K/AKT and MAPK cascades. Numbers in the red bars refer to the experiment in which the target was quantified. The letter for the mechanism involved is written inside the colored hexagon annotating the interaction arrow, abbreviations indicate the following: CF = complex formation; Cm = covalent modification; Tr = transcription regulation; B = binding; +P = phosphorylation; -P = dephosphorylation; Z = catalysis; Tn = transport; CS = complex subunit; GR = group relation.



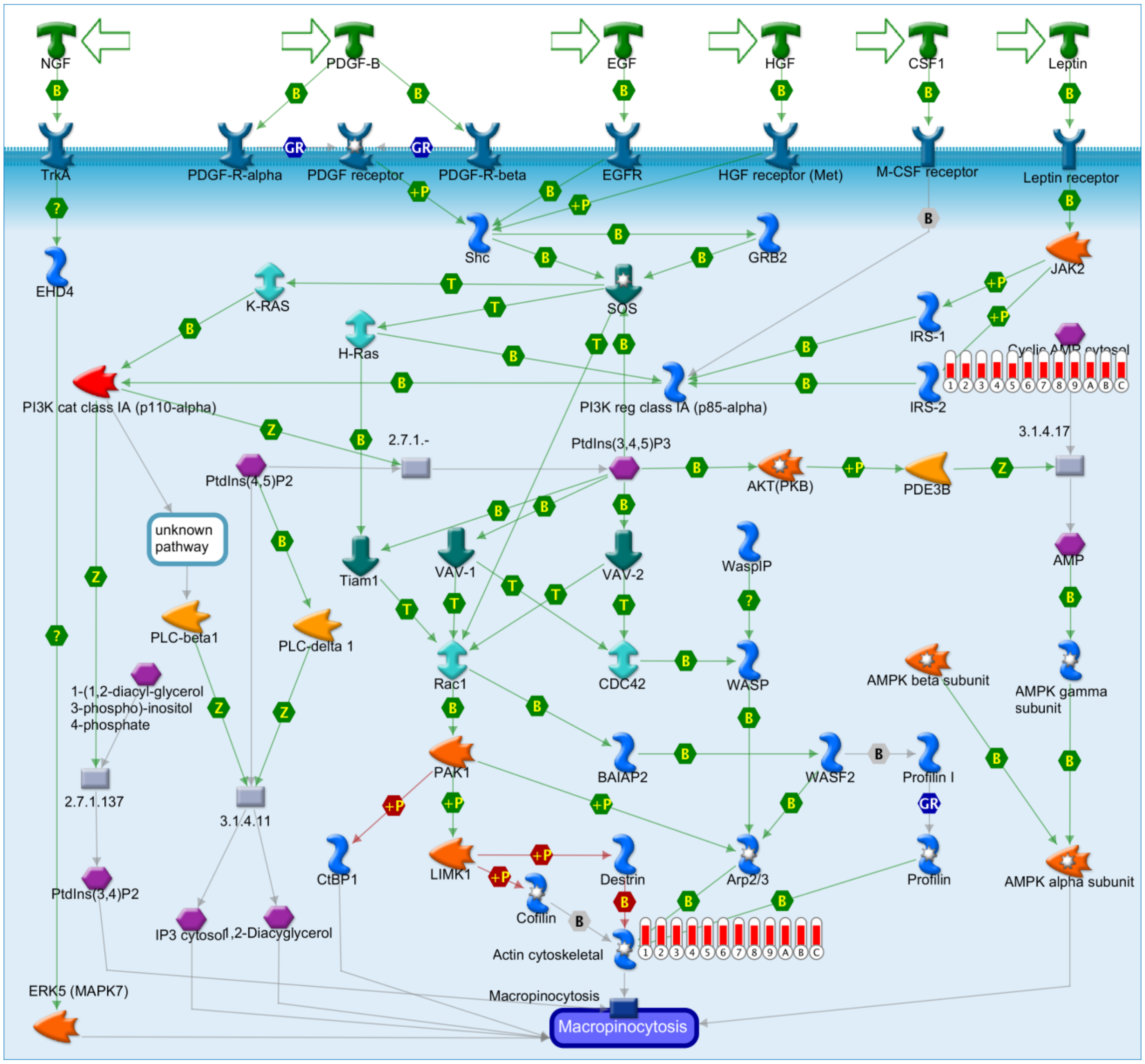
(b)

Figure 4.2 (continued).



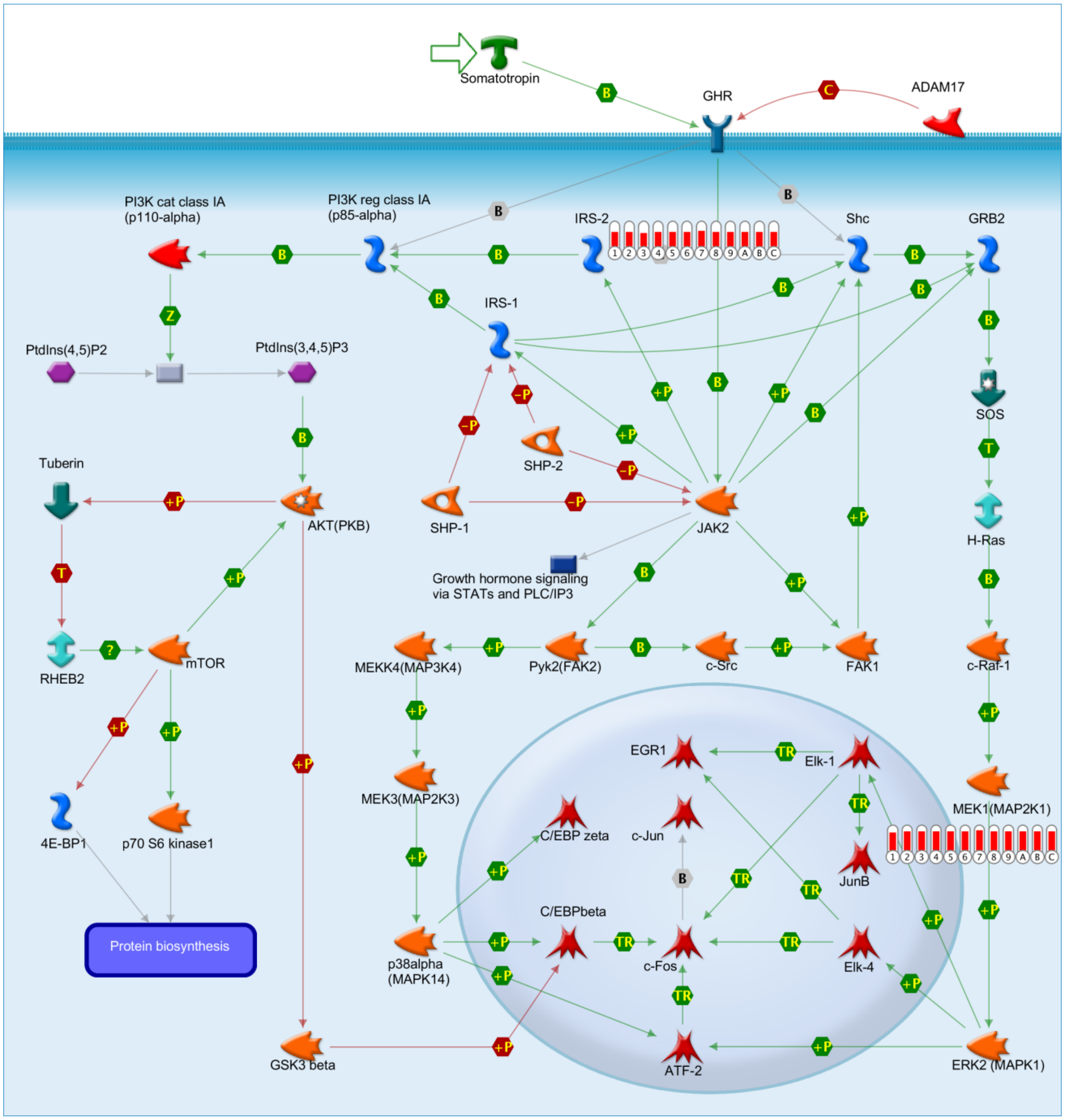
(c)

Figure 4.2 (continued).



(d)

Figure 4.2 (continued).



(e)

Figure 4.2 (continued).

The last two networks in Table 4.1, involving *JUNB* and/ or *OAZ1* are predicted to play roles in general cell responses to chemical stimulus and subsequent transcriptional regulation.

Regulation of cell processes over time is the product of many partially independent networks, from epigenetic marks to RNA polymerases whose activity is regulated by transcription factors to splicing, translation and protein modification. We next looked for evidence that specific sets of transcription factors (TFs, which are DNA binding proteins that regulate the transcription of their target genes by binding in the promoter region) were involved in the altered expression of the DE genes. Transcription factors are often expressed in very small amounts and any changes in their concentration are difficult to quantify, so TF networks are usually inferred based on enrichment of their binding sites, relative to randomly selected genes, in the DE genes [136]. The MetaCore software prioritizes the most statistically significant transcription factor networks, providing visualization with GeneGo maps, such as those shown in Figure 4.3. The prediction of activation versus deactivation is indicated by the arrow direction. For our set of genes, 29 significant networks are predicted; of these 9 have been selected for discussion, based on their statistical importance (by *p-value*, Z-score and G-score): the scores and GO process annotations are given in Table 4.2. Only network 8 includes a known regulatory pathway (see Figure 4.3 (a)), discussed in detail after the individual TFs have been described.

Only seven TFs (SP1, SP1/SP3, c-Myc, p63, ESR2, SREPB1 and STAT3) have the most significant influence patterns on our DE genes set, combining both activation and inhibition. These are discussed in the following paragraphs.

SP1 (specificity protein 1) causes cells to respond to various physiological or pathological stimuli. It is both an activator and repressor, affecting the transcription of many genes involved in cell growth, immune responses, and apoptosis [137]. When the

SP1/SP3 complex is dissected the proteins have been shown to carry a wide range of post-translational modifications, including phosphorylation, glycosylation, and proteolytic cleavage [138]. Previous studies have shown an association of SP1 regulated genes with neurodegenerative diseases, including Alzheimer's [139] and Prion diseases [140]. SP1 positively regulates *S100A6* (Calcyclin)*, IRS2*, *SLC25A3* and *UBB* as indicated in networks 1, and 5-8 (Figure 4.3 (a-b, f-h)). The mode of regulation for *RPS10, OAZ1*, *PABC1, LAMR1, SCP1, and PSMA* is unknown and likely depends on what other TFs are present. SP1 is known to interact with the TFs E2F3 (network 1, Figure 4.3 (b)) and FB1-1 (network 5, Figure 4.3 (f)), which in turn are connected with *JUNB* through an unknown mode of action. SP1 activates the TF p63 that in turn inhibits *JUNB* and activates *B2M* (network 6, Figure 4.3 (g)). Through the p63 interaction it affects STAT3 (network 6, Figure 4.3 (g)) and HNF3 (network 9, Figure 4.3 (i)), the latter being responsible for *JUNB* upregulation; SP1 activates *SLC 25A3*, leading to a c-Myc connection, that may also by modulated by several genes whose mechanism of interaction is currently unknown, including *PABC1, RPS10, OAZ1, PSMA1* (network 8, Figure 4.3 (a)). SP1 is inhibited by RIPK1 kinase following positive regulation initiated by the TWEAK (TNFSF12) and TNFRSF12A receptors (Figure 4.3 (b, f-h)).

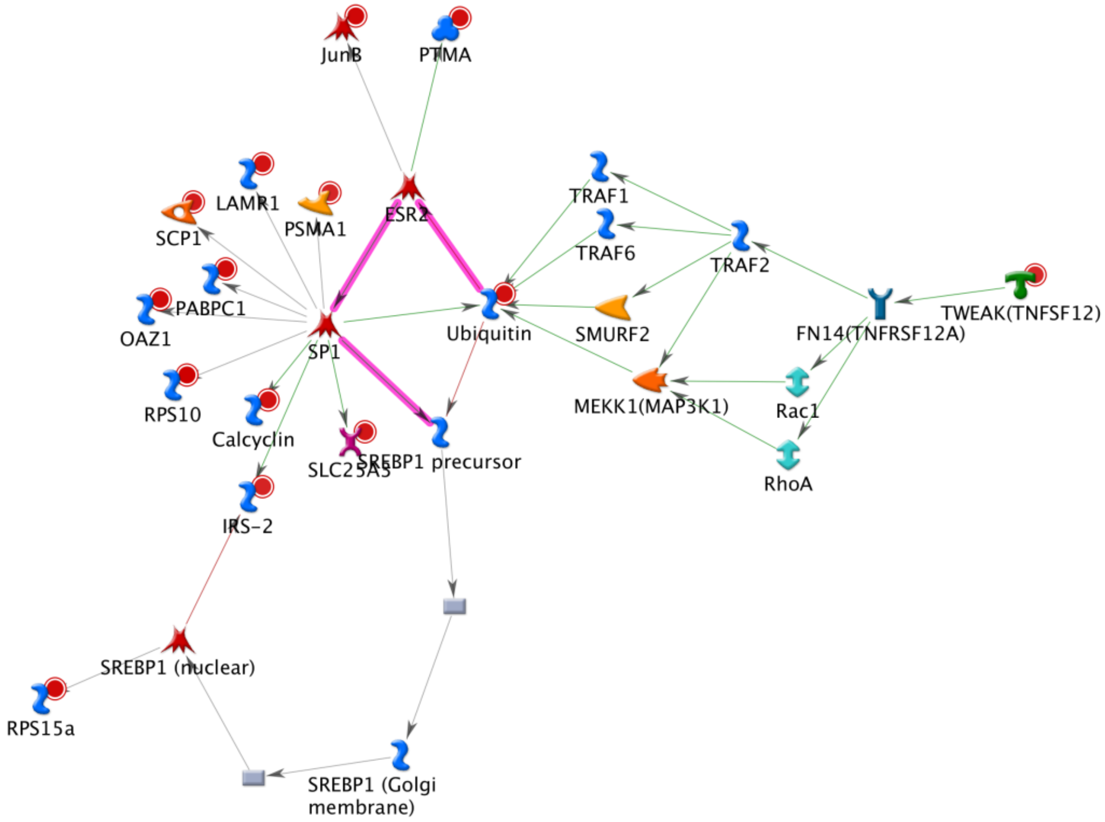
The SP1/SP3 complex, shown in network 7 (Figure 4.3 (h)), has a more restricted set of functions than SP1 alone. It positively regulates *UBB, SLC25A3, S100A6* and *IRS2*, while its regulation of *PSMA1, OAZ1, SCP1, PABC1, RPS10* is more nuanced. The complex connects indirectly to the *JUNB* pathway through inhibition of the activated HMGA2 binding protein.

Table 4.2 Transcription factors networks for current selection of genes.

No.	Network name	GO process	p-value	gScore
1.	E2F3, DR3(TNFRSF12), FN14(TNFRSF12A)	viral genome expression (19.2%; 6.509e-08), viral transcription (19.2%; 6.509e-08), translational termination (19.2%; 8.521e-08), cellular protein complex disassembly (19.2%; 1.847e-07), protein complex disassembly (19.2%; 2.107e-07)	9.02e-59	152.12
2.	HNF1-alpha, FN14(TNFRSF12A)	endocrine pancreas development (21.1%; 6.042e-06), pancreas development (21.1%; 1.600e-05), RNA metabolic process (57.9%; 4.021e-05), viral reproduction (26.3%; 4.409e-05), cellular component disassembly at cellular level (21.1%; 4.778e-05)	5.30e-47	140.48
3.	HOXB4, FN14(TNFRSF12A)	system development (65.0%; 4.707e-06), developmental process (70.0%; 1.164e-05), anatomical structure development (65.0%; 1.474e-05), multicellular organismal development (65.0%; 3.048e-05), anatomical structure morphogenesis (45.0%; 5.066e-05)	2.12e-46	136.92
4.	ZNF206, FN14(TNFRSF12A)	viral reproduction (33.3%; 1.569e-06), cellular component disassembly at cellular level (22.2%; 3.804e-05), cellular component disassembly (22.2%; 3.992e-05), ribosomal small subunit assembly (11.1%; 4.862e-05), viral transcription (16.7%; 5.270e-05)	1.16e-43	134.71
5.	FBI-1 (Pokemon), DR3(TNFRSF12)	positive regulation of cellular process (62.5%; 2.514e-05), viral transcription (18.8%; 3.639e-05), viral genome expression (18.8%; 3.639e-05), ribosomal small subunit assembly (12.5%; 3.816e-05), embryonic process involved in female pregnancy (12.5%; 3.816e-05)	4.65e-37	122.46
6.	p63, DR3(TNFRSF12), FN14(TNFRSF12A)	positive regulation of biological process (77.3%; 9.566e-10), positive regulation of cellular process (72.7%; 3.039e-09), apoptosis (40.9%; 2.005e-07), programmed cell death (40.9%; 2.357e-07), induction of apoptosis by extracellular signals (22.7%; 2.639e-07)	5.01e-38	113.12
7.	HMGA2, DR3(TNFRSF12), FN14(TNFRSF12A)	positive regulation of cellular process (70.0%; 6.703e-08), embryonic process involved in female pregnancy (15.0%; 1.802e-07), positive regulation of biological process (70.0%; 2.428e-07), positive regulation of macromolecule metabolic process (50.0%; 1.260e-06), positive regulation of cellular metabolic process (50.0%; 1.772e-06)	3.22e-35	109.52

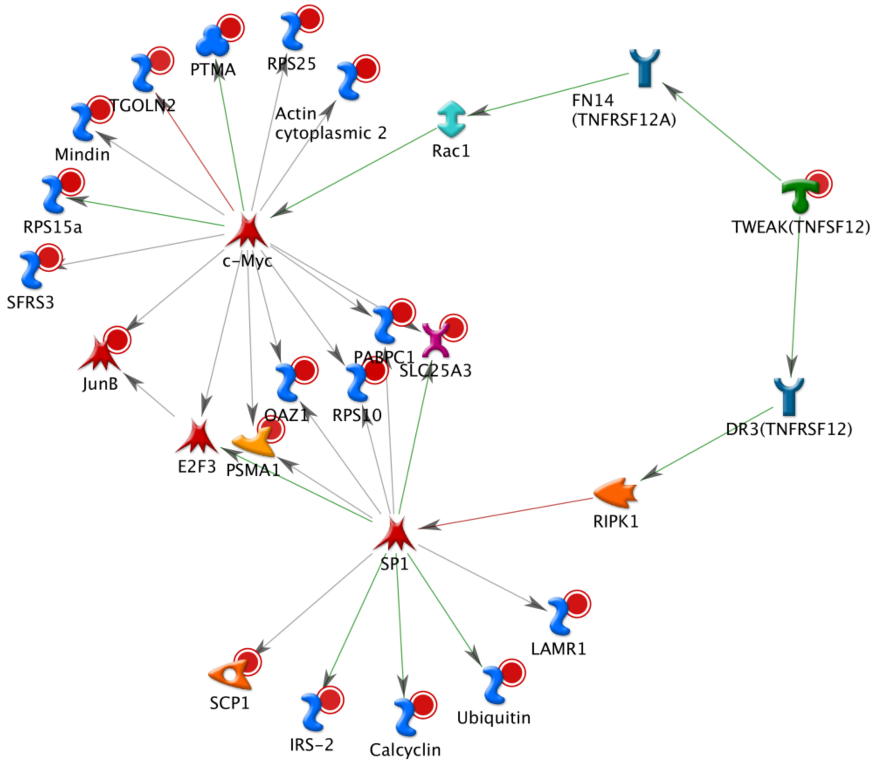
Table 4.2 (continued)

8.	SREBP1 (nuclear), FN14(TNFRSF12A)	positive regulation of cellular process (72.0%; 3.944e-10), positive regulation of biological process (72.0%; 2.087e-09), regulation of apoptosis (48.0%; 2.235e-08), regulation of programmed cell death (48.0%; 2.460e-08), regulation of cell death (48.0%; 3.308e-08)	1.52e-39	107.97
9.	HNF3, DR3(TNFRSF12), FN14(TNFRSF12A)	positive regulation of cellular process (77.3%; 1.926e-10), positive regulation of biological process (77.3%; 9.566e-10), positive regulation of macromolecule metabolic process (59.1%; 2.017e-09), positive regulation of cellular metabolic process (59.1%; 3.177e-09), lung epithelial cell differentiation (18.2%; 3.530e-09)	1.65e-34	104.41

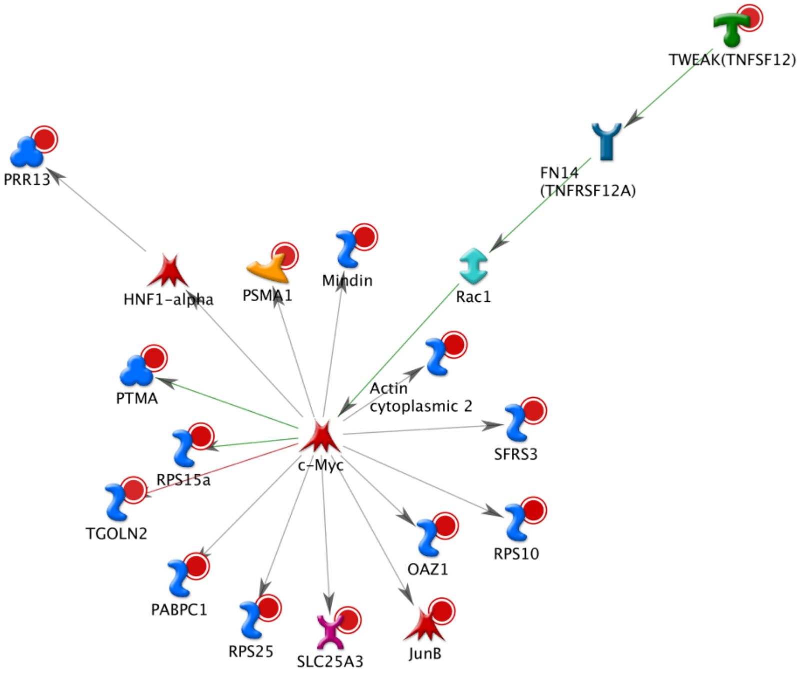


(a)

Figure 4.3 GeneGo maps of transcription factors interaction networks (corresponding numbers are found in Table 4.2). Some symbols are explained in the Figure 4.1 key. (a) network 8; (b) network 1; (c) network 2; (d) network 3; (e) network 4; **(**f) network 5; (g) network 6; (h) network 7; (i) network 9. (j) Symbol legend from MetaCore.

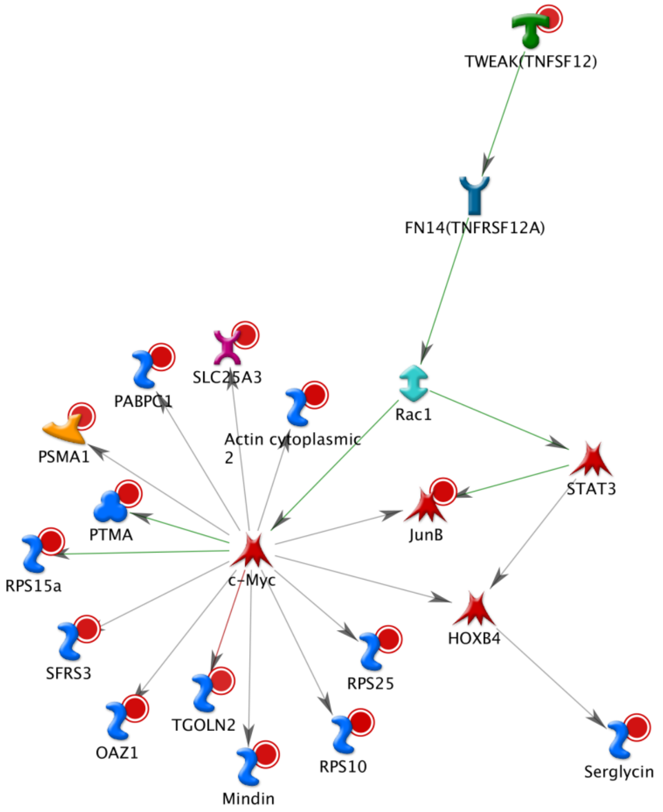


(b)

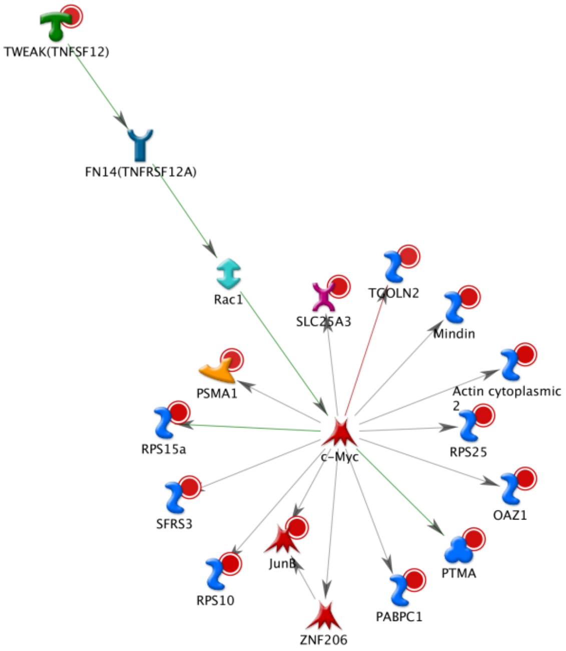


(c)

Figure 4.3 (continued)



(d)



(e)

Figure 4.3 (continued)

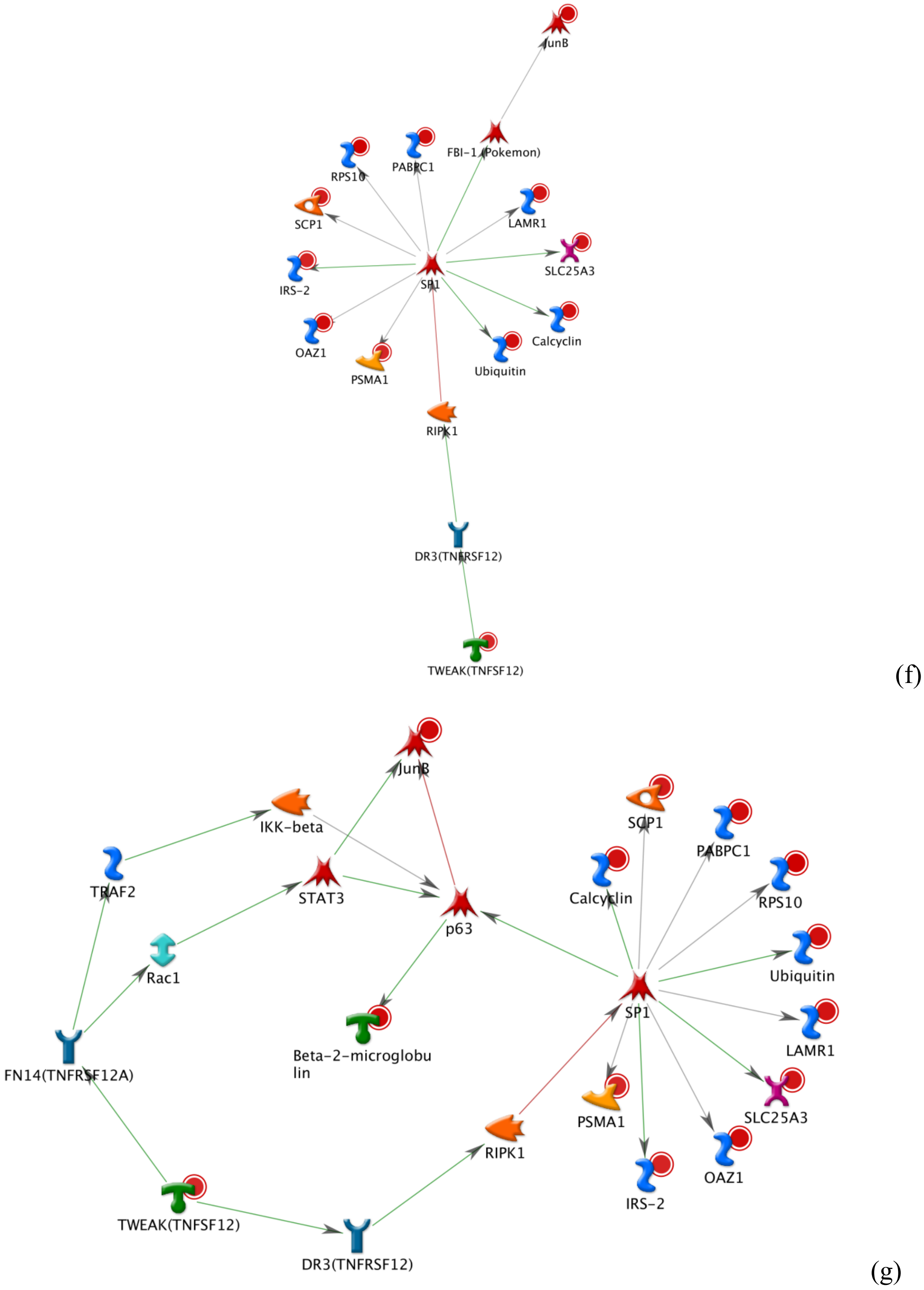


Figure 4.3 (continued).

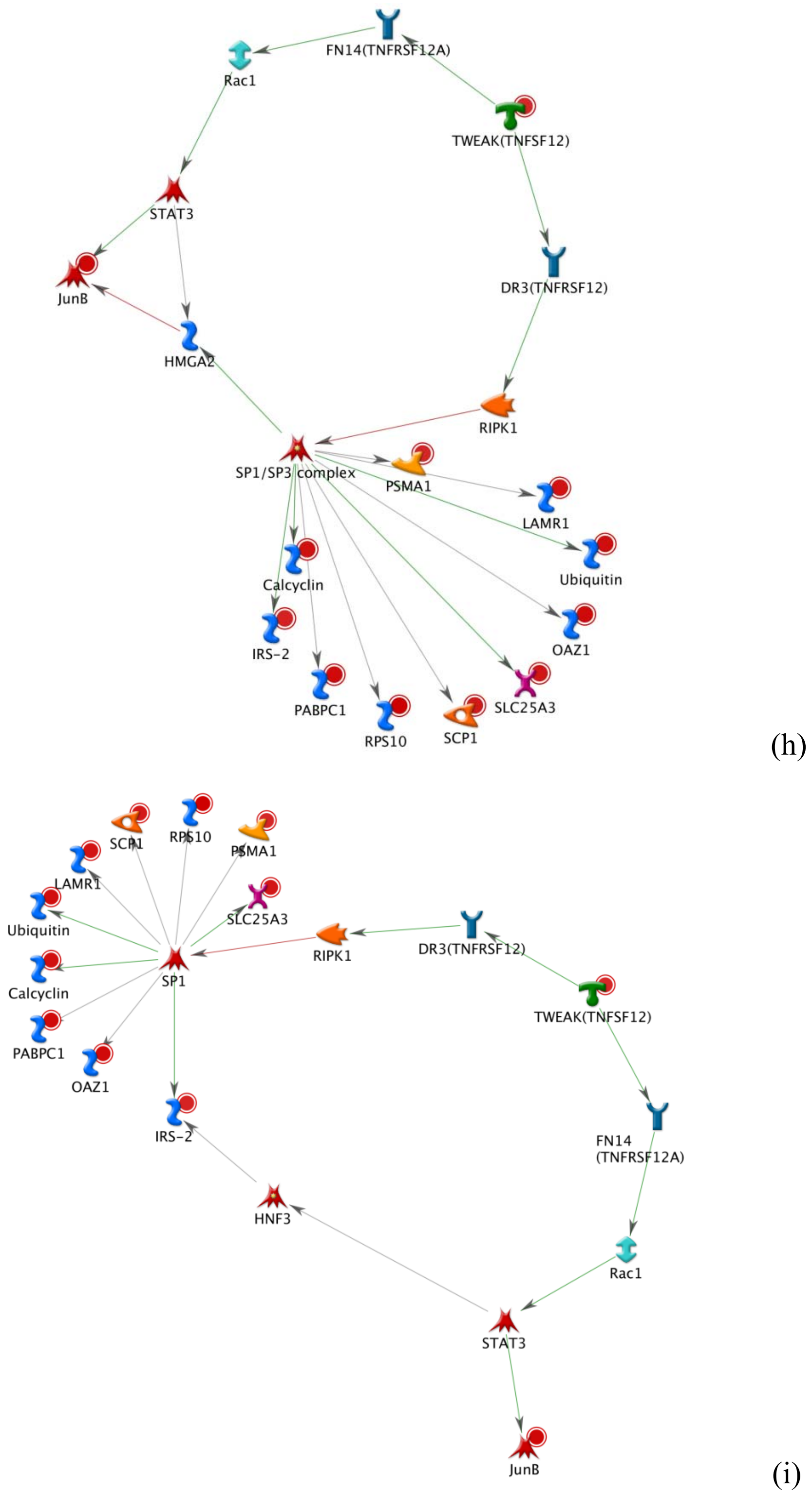
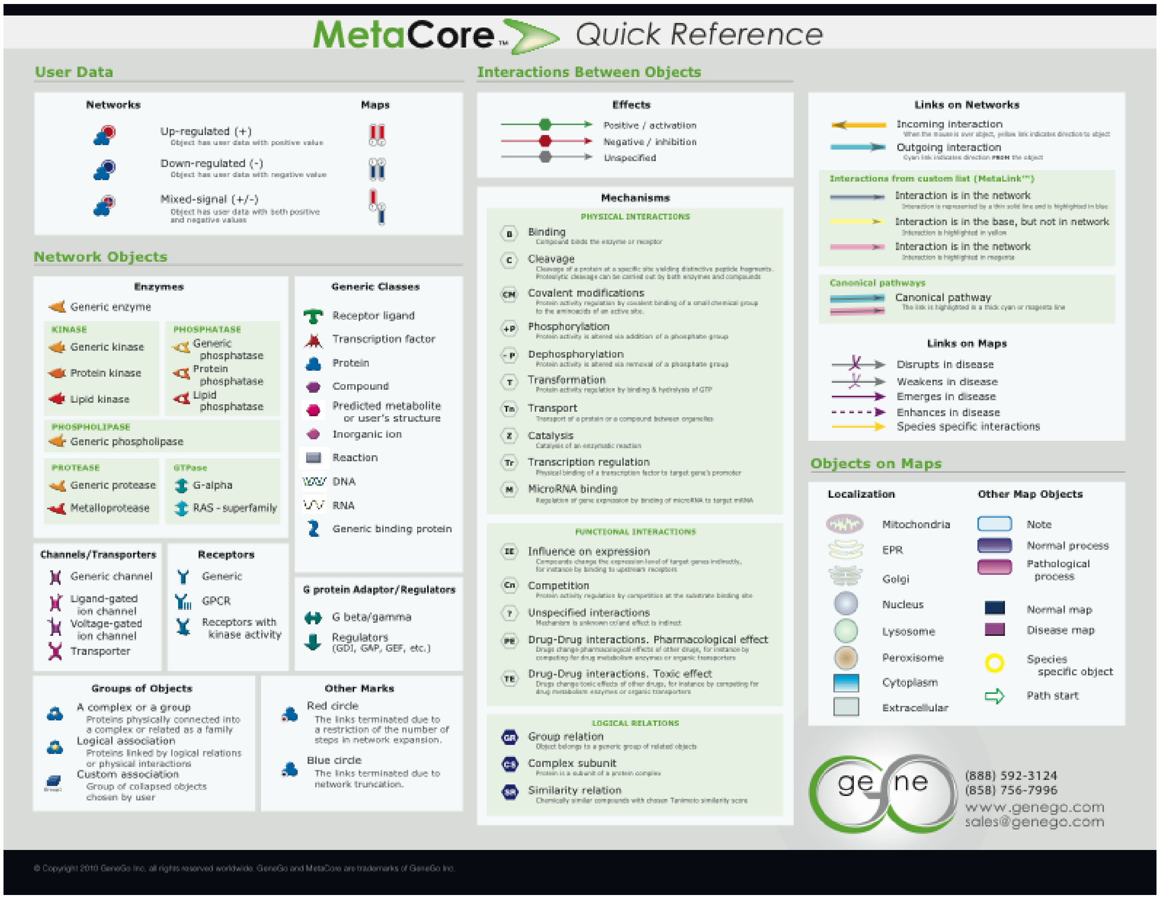


Figure 4.3 (continued).



(j)

Figure 4.3 (continued)

Appearing in four of the nine TFs networks, c-Myc or the c-Myc proto-oncogene, is known for its wide range of functions as transcription factor and particularly for its importance in various tumors, leukemias and lymphomas [141-143]. It is known to activate *RPS15A* and *PTMA1*, shown in networks 1-4 (Figure 4.3 (b-e)), but it inhibits *TGOLN2*. Its mode(s) of influence on *PABC1*, *SLC25A3*, *RPS10*, *RPS25*, *OAZ1*, *PSMA1*, *JUNB*, *S100A6*, *ACTG1*, *SFRS3* is not currently known. c-Myc itself is activated by small GTPase Rac1, a member of RAS superfamily, following positive regulation from

TWEAK (TNFSF12) via the FN14 receptor. Transcription factors are themselves regulated by transcription factors, and c-Myc transcription is regulated by SP1 (network 1), discussed above, HNF1-alpha and *JUNB* (network 2, Figure 4.3 (c)), STAT3 via *JUNB* (network 3, Figure 4.3 (d)) and ZNF206 (network 4, Figure 4.3 (e)).

The p63 gene encodes a member of p53 family of transcription factors; it has known roles in development and the maintenance of stratified epithelial tissues [144]. As described above and shown in network 6 (Figure 4.3 (g)), p63 interacts directly with the TFs SP1 and STAT3 and regulates two of the DE genes, activating *B2M* and inhibiting *JUNB*. Similar to the TFs SP1 and c-Myc, p63 is activated through the pathway connecting the surface receptor TWEAK→FN14→Rac1→STAT3→p63.

STAT3 is one member of the large STAT family, that play key roles in many cellular processes, i.e. cell growth and apoptosis [145]. As seen in network 6, STAT3 activates *JUNB* and p63, and is induced by the small GTPase Rac1.

The final two TFs, ESR2 (estrogen receptor 2) and SREPB1 (sterol regulatory element binding TF1) are part of a large network with complex regulatory interactions. Shown in network 8, Figure 4.3 (a)), the pathways also are modulated by SP1 and *JUNB*. The highlighted interactions show the positive regulation leading from Ubiquitin→ESR2→SP1→SREBP1 precursor, but Ubiquitin itself is activated by interactions that start with the TWEAK (TNFSF12) ligand receptor. This highly connected network is the reason that such high-level GO processes are involved, i.e. cell maintenance and cell death.

4.4 Conclusions

From a set of 60 differentially expressed genes, assayed from circulating white blood cells and selected using the LO-BaFL pipeline we used several forms of pathway analysis to see whether they are more likely connected to specific sALS pathology or to general disease responses. We input the entire list to derive the meaningful biological interactions and transcription factors networks from GeneGo maps of MetaCore. 12 statistically significant networks involving several of the selected genes were obtained; the most highly ranked was TWEAK (TNFSF12) via TNF receptor-associated factors 2 (TRAF2) or 5 (TRAF5). Of the constituent genes, *ACTG1, IRS2, DIA1* and *JUNB* in a variety of associations promote processes such as cytoskeleton remodeling, regulation of the actin cytoskeleton by RhoGTPase, immune responses via *CCR3* signaling in eosinophils, the role of alpha-6/beta-4 integrins in carcinoma progression, transport macropinocytosis regulation by growth factors, development growth hormone signaling, via PI3K/AKT and MAPK cascades, with many other roles in apoptosis, proliferation, migration, angiogenesis, and inflammation [134]. These are in good agreement with previous reports [116, 119, 123].

As would be expected in a disease in which cell death occurs, many of the genes are part of immune response pathways (HLA-type), including *RPS10, RPS15A, RPS25, RPL21, LAMR1* and *SLC25A3,* other genes participate in regulation of the immune response or responses to other stimuli including *PABC1, PTMA, SFRS3, DIA1, LAMR1, FTH1, B2M, SLC25A3, CCDC6, OAZ1*. Responses more specific to a disease involving muscle-neuron interactions were seen for integrin-type proteins with roles in cell adhesion and cell-surface mediated signaling, and with Myosin II and myosin smooth muscle specific genes characteristic of ALS.

Programmed cell death and cellular responses to the products of those events are prominent through networks that include ubiquitin, *UBB*. Examples of affected genes include *PTMA, JUNB, SCP1, OAZ1, IRS2 or PABC1*, while *IRS2, SPON2, DIA1* are nodes in one network that regulates apoptosis. Other networks including *JUNB* and / or *OAZ1* are predicted to play roles in cell responses to chemical stimulus. Although we do not show here the direct affect of these genes on Ubiquitin/Proteasome System that has been proved to perturb the ALS pathway [92], the presence of Ubiquitin in several significant interaction networks might suggest a similar trend.

MetaCore also derived 29 statistically significant transcription factors networks, for which we summarized the 9 most significant. SP1, SP1/SP3, c-Myc, p63, STAT3, ESR2 and SREPB1 were identified although not all of the interactions are currently defined. For instance, *S100A6, IRS2, UBB* and *SLC25A3* are activated by SP1 or the SP1/SP3 complex, but the way they regulate *RPS10, PABC1, LAMR1, SCP1, OAZ1, PSMA* is not known. c-Myc activates *RPS15A* and *PTMA*, while it inhibits *TGOLN2* and the mode of action is not known for *PABC1, SLC25A3, RPS10, OAZ1, PSMA1, JUNB, S100A6, RPS25, ACTG1* and *SFRS3*. *JUNB*, in direct connection with STAT3 and p63, is activated by the first and inhibited by the latter, which in turn stimulates *B2M* activity. *PTMA* activity is shown to be stimulated by *ESR2*, while *IRS2* is inactivated by *SREPB1*. Because the targeted cells were of several types, an effect in one cell type may mask an opposing effect in another cell type, particularly when complexes of TFs have different effects. For instance, p63 in Figure 4.3 (g) is activated by two other TFs: SP1 and STAT3, as a result of which *B2M* becomes activated while *JUNB* is repressed.

It is interesting that a large subset of the DE genes in our list (*ACTG1, IRS2, DIA1, JUNB, PABC1, PTMA, SFRS3, LAMR1, FTH1, B2M, SLC25A3, CCDC6, OAZ1, UBB, RPS10, RPS15A, RPS25, RPL21, PSMA1, S100A6, TGOLN2, SFRS3, SCP1*) are regulatory and while many are engaged in normal cell processes that are perhaps ramped up to accommodate a higher than normal cell turnover, the immune response signature and apoptosis and responses to chemical stimulus are likely more specific to sALS and are good candidates for a simplified blood-based biomarker signature for its presence than have been yielded by previous studies. The mechanism of action and their exact role in sALS pathology is still to be determined by future work.

SUMMARY

Amyotrophic Lateral Sclerosis is a heterogeneous, complex disease whose etiology is poorly understood, despite many studies performed over many sample types, from biopsies to biofluids, and DNA, RNA and proteins. The goal of the present study was to find diagnostic markers that will help determine who has ALS, hopefully as early as possible and in a readily obtainable medium, in this case blood, which is drawn during most routine physicals. This has immediate benefits in the clinic, since with a 18 month-5 year life expectancy even the gain of 2-3 months from Riluzol is significant. We were looking for markers that are clearly present in all patients of a particular group and clearly distinct from individuals in the contrast group, whether as present/absent expression of genes, or expression levels that are completely distinct.

The approach taken in the first stage of this research was to revise the array design for a set of microarray experiments in order to remove design errors and compare the subsequent predictions of differential expression to the standard method. As a control, an independent experiment performed on the same platform for which independent assays had been performed to test the microarray predictions was analyzed, despite not being focused on the disease phenotype of interest (CAD versus sALS). The novel pipeline, LO-BaFL, was developed to correct for errors that arise in microarray design, i.e. cross-hybridization, loss of binding site, miss-assignment of particular probes. Because of its strict filtering, LO-BaFL improves the power and discrimination of identifying the differentially expressed genes but also it eliminates genes that are not problematic in specific populations.

Comparison of the responses from a similar set of Normal samples in an independent study using the same Agilent platform revealed a good correlation (R=0.81) between Healthy Controls in sALS and CAD studies, giving us confidence in the disease responses as well.

LO-BaFL pipeline, and SAM and TM4 as comparative methods, were used to cleanse data and analyze microarray data from sALS study. LO-BaFL revealed a subset of 87 DE genes, versus 209 of SAM and 264 of TM4. Of particular note was that by combining TM4, which predicts a TARDBP expression change, with LO-BaFL, which indicates that 4 genes all contribute signal to the probe mapped to TARDBP, it became clear that several genes had to be tested in the follow-up assay, including ILKAP.

After comparing the three lists of DE genes identified by LO-BaFL, SAM and TM4, we selected the top genes for validation with qRT-PCR assay, an independent method. Such validation is recommended as a follow-up for microarray predictions. The selected genes are described in details in Chapter 2.

The biomolecular assays have been performed on RNA - derived PBLs samples from subjects with sALS. Testing of 12 genes with qRT-PCR, using the samples that passed the quality assessment (RIN > 5.5) confirmed the microarray observations and most of our computational predictions when applying LO-BaFL and comparative methods for microarray analysis: *ACTG1, SKIV2L2, C12orf35, B2M, DYNLT1, ILKAP, TARDBP* were found to have higher expression ratio in patients with sALS vs. Healthy Controls. With respect to the genes listed, the corresponding expression ratio values are: 48.5; 37.3; 22.4; 18.2; 17.4; 8.8; 5.6. This confirms the results of previous and more recent studies

[8, 12, 75, 92], with additional new candidate biomarkers in the genes *ACTG1, B2M, ILKAP*. Also importantly, qRT-PCR results confirmed *TARDBP* is among the DE genes for sALS. However, the results show less differential expression than the microarray predicted. This is due to increased expression of two genes that cross-hybridize (*TARDBP* and *ILKAP*) measured by the same Agilent probe.

Following up on selected DE transcripts, we searched for the presence of sequence variants, e.g. SNPs, by performing Sanger sequencing assays, as described in Chapter 3.

Direct sequencing was performed to screen for possible mutations in selected exons of DE genes determined by the LO-BaFL pipeline, for the five Healthy Controls and five sALS samples that passed quality control step. No sequence variant that consistently segregated with the sALS samples was found. In 9 out of 10 samples we identified a novel mutation in exon 3 of the *ACTG1* gene, c.350 C-to-T, as follows:

TCTGGCACCACACCTTCTACAA[**C**/**T**]GAGCTGCGCGTGGCCCCGGAGGAGCAC

However, since we found this variant in what are labeled ‘Healthy Controls’, it does not correlate with sALS.

For the other DE genes tested, since the exons do show differential expression, but no sequence variant was found, we assume that changes could have other causes: more distant structural changes or regulatory changes due to presence of transcription factors or other modulators. This is one type of pathway analysis we performed, described in Chapter 4.

We input the entire LO-BaFL DE genes list to derive the meaningful biological interactions and transcription factors networks from GeneGo maps of MetaCore. Twelve statistically significant networks involving several of the selected genes were obtained; the most highly ranked was TWEAK (TNFSF12) via TNF receptor-associated factors 2 (TRAF2) or 5 (TRAF5), with roles in apoptosis, proliferation, angiogenesis and inflammation. Of the constituent genes, *ACTG1*, *IRS2*, *DIA1* and *JUNB* in a variety of associations promote processes such as cytoskeleton remodeling, immune responses, playing roles in carcinoma progression, transport macropinocytosis regulation by growth factors, development growth hormone signaling, via PI3K/AKT and MAPK cascades, with many other roles in apoptosis, proliferation, migration, angiogenesis, and inflammation [134]. These are in good agreement with previous reports [116, 119, 123].

Many of the genes are part of immune response pathways (HLA-type), including *RPS10, RPS15A, RPS25, RPL21, LAMR1* and *SLC25A3,* other genes participate in regulation of the immune response or responses to other stimuli including *PABC1, PTMA, SFRS3, DIA1, LAMR1, FTH1, B2M, SLC25A3, CCDC6, OAZ1*. Responses more specific to a disease involving muscle-neuron interactions were seen for integrin-type proteins with roles in cell adhesion and cell-surface mediated signaling, and with Myosin II and myosin smooth muscle specific genes characteristic of ALS.

MetaCore also derived 29 statistically significant transcription factor networks, for which we summarized the 9 most significant. SP1, SP1/SP3, c-Myc, p63, STAT3, ESR2 and SREPB1 were identified although not all of the interactions are currently defined. A large selection of genes from our input list (*ACTG1, IRS2, DIA1, JUNB, PABC1, PTMA, SFRS3, LAMR1, FTH1, B2M, SLC25A3, CCDC6, OAZ1*, *UBB*, *RPS10, RPS15A, RPS25, RPL21, PSMA1, S100A6, TGOLN2, SFRS3, SCP1*), regulated by these transcription factors, are regulatory, and while many are engaged in normal cell processes that are perhaps ramped up to accommodate a higher than normal cell turnover, the immune response signature and apoptosis and responses to chemical stimulus are likely more specific to sALS and are good candidates for a simplified blood-based biomarker.

REFERENCES

1. Rothstein JD: Current Hypotheses for the Underlying Biology of Amyotrophic Lateral Sclerosis. *Ann Neurol* 2009, 65(suppl):S3-S9.

2. Lagier-Tourenne C, Cleveland DW: Rethinking ALS: The FUS about TDP-43. *Cell* 2009, 136:1001-1004.

3. Mougeot JL, Milazi SR, Brooks BR: Whole-genome association studies of sporadic lateral sclerosis: are retroelements involved? *Trends in Molecular Medicine* 2009, 15(14):148-158.

4. Corcia P, Meininger V: Management of amyotrophic lateral sclerosis. *Drugs* 2008, 68(8):1037-1048.

5. Valdmanis PN, Rouleau GA: Genetics of familial amyotrophic lateral sclerosis. *Neurology* 2008, 70(2):144-152.

6. Brown RH, Jr.: SOD1 aggregates in ALS: cause, correlate or consequence? *Nat Med* 1998, 4(12):1362-1364.

7. Boillee S, Velde CV, Cleveland DW: ALS: a disease of motor neurons and their noneuronal neighbors. *Neuron* 2006, 52:39-59.

8. Daoud H, Valdmanis PN, Kabashi E, Dion P, Dupre N, Camu W, Meininger V, Rouleau GA: Contribution of TARDBP mutations tp sporadic amyotrophic lateral sclerosis. *J Med Genet* 2008, 124:649-658.

9. Gitcho MA, Baloh RH, Chakraverty S, Mayo K, *et al.*: TDP-43 A315T mutation in familial motor neuron disease. *Ann Neurol* 2008, 63:535-538.

10. Rutherford NJ, Zhang Y-J, Baker M, *et al.*: Novel Mutations in TARDBP (TDP-43) in Patients with Familial Amyotrophic Lateral Sclerosis. *PLoS Genet* 2008, 4(9):e1000193.

11. Van Deerlin VM, Leverenz JB, Bekris LM, *et al.*: TARDBP mutations in amyotrophic lateral sclerosis with TDP-43 neuropathology: a genetic and histopathological analysis. *Lancet Neurol* 2008, 7:409-416.

12. Gitcho MA, Bigio EH, Mishra M, *et al.*: TARDBP 3'-UTR variant in autopsy-confirmed frontotemporal lobar degeneration with TDP-43 proteinopathy. *Acta Neuropathol* 2009.

13. Valdmanis PN, Daoud H, Dion PA, Rouleau GA: Recent Advances in the Genetics of Amyotrophic Lateral Sclerosis. *Current Neurology and Neuroscience Reports* 2009, 9:198-205.

14. Kasperaviciute D, Weale ME, Shianna KV, *et al.*: Large-scale pathways-based association study in amyotrophic lateral sclerosis. *Brain* 2007, 130:2292-2301.

15. Schymick JC, Scholz SW, Fung HC, *et al.*: Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2007, 6:322-328.

16. Dunckley T, Huentelman MJ, Craig DW, *et al.*: Whole-Genome Analysis of Sporadic Amyotrophic Lateral Sclerosis. *N Engl J Med* 2007, 357.

17. van Es MA, van Vught PW, Blauw HM, *et al.*: Genetic Variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. *Nature Genetics* 2008, 40(1):29-31.

18. van Es MA, Van Vught PW, Blauw HM, *et al.*: ITPR2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: a genome-wide association study. *Lancet Neurol* 2007, 6:869-877.

19. Del Bo R, Ghezzi S, Corti S: DPP6 gene variability confers increased risk of developing sporadic amyotrophic lateral sclerosis in Italian patients. *J Neurol Neurosurg Psychiatry* 2008, 79:1085.

20. Cronin S, Tomik B, Bradley DG, *et al.*: Screening for replication of genome-wide SNP association in sporadic ALS. *Eur J of Hum Genet* 2009, 17:213-218.

21. Blauw HM, Veldink JH, van Es MA, *et al.*: Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. *Lancet Neurol* 2008, 7:319-326.

22. Cronin S, Blauw HM, Veldink JH: Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. *Hum Mol Genet* 2008, 17:3392-3398.

23. Cluskey S, Ramsden DB: Mechanisms of neurodegeneration in amyotrophic lateral sclerosis. *Molecular Pathology* 2001, 54:386-392.

24. Shaw PJ: Molecular and cellular pathways of neurodegeneration in motor neuron disease. *J Neurol Neurosurg Psychiatry* 2005, 76(8):1046-1057.

25. Ryberg H, Bowser R: Protein biomarkers for Amyotrophic Lateral Sclerosis. *Expert Rev Proteomics* 2008, 5(2):249-262.

26. Leparc G, Tuchler T, G S, Bayer K, Sykacek P, GHofacker IL, Kreil DP: Model-based probe set optimization for high-performance microarrays. *NAR* 2008:1-12.

27. Binder H, Kirsten T, Loeffler M, Stadler PF: Sensitivity of Microarray Oligonucleotide Probes: Variability and Effect of Base Composition. *J Phys Chem B* 2004, 108(46):18003-18014.

28. Mathews DH, Burkard ME, Freier SM, Wyatt JR, Turner DH: Predicting oligonucleotide affinity to nucleic acid targets. *RNA* 1999, 5:1458-1469.

29. Rouillard JM, Zuker M, Gulari E: Oligoarray 2.0: design of ologonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic Acids Res* 2003, 31:3057-3062.

30. The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851-862.

31. Emmert Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: Laser capture microdissection. *Science* 1996, 274(5289):998–1001.

32. Thompson K, Deshmukh H, Solva J, Weller JW: A white-box approach to microarray probe response characterization: the BaFL pipeline. *BMC Bioinformatics* 2009, 10:449.

33. Kumari S, Verma L, Weller JW: AffyMAPSDetector: a software tool to characterize Affymetrix GeneChip expression arrays with respect to SNPs. *BMC Bioinformatics* 2007, 8:276.

34. Rouchka EC, Phatak AW, Singh AV: Effect of single nucleotide polymorphisms on Affymetrix(R) match-mismatch probe pairs. *Bioinformatics* 2008, 2(9):405-411.

35. Wang M, Hu X, Li G, Leach LJ, Potokina E, Druka A, Waugh R, Kearsey MJ, Luo Z: Robust detection and genotyping of single feature polymorphisms from gene expression data. *PLoS Comput Biol* 2009, 5(3):e1000317.

36. Deshmukh H: Modeling the Physical Parameters Affecting the Measurements from Microarrays. Fairfax: George Mason University; 2006.

37. Ratushna VG, Weller JW, Gibas CJ: Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* 2005, 8(6):31.

38. Thompson K: An Adenocarcinoma Case Study of the BaFL Protocol: Biological Probe Filtering for Robust Microarray Analysis. Fairfax: George Mason University; 2009.

39. Bengtsson H, Irizarry R, Carvalho B, Speed TP: Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 2008, 24:759-767.

40. Bengtsson H, Simpson K, Bullard J, Hansen K: aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. In: *Tech Report #745*. Department of Statistics, University of California, Berkeley; February 2008.

41. Bengtsson H, Wirapati P, Speed TP: A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6. *Bioinformatics* 2009.

42. Bengtsson H, Jönsson G, Vallon-Christersson J: Calibration and assessment of channel-specific biases in microarray data with extended dynamical range. *BMC Bioinformatics* 2004, 5:177.

43. Shi L, Tong W, Su Z, Han T, Han J, Puri RK, Fang H, Frueh FW, Goodsaid FM, Guo L *et al*: Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics* 2005, 6(Suppl 2):S11.

44. Liu P, Hwang GJT: Quick calculation for sample size while controlling false discovery rate with application to microarray analysis *Bioinformatics* 2007, 6:739-746.

45. Wingrove JA, Daniels SE, Sehnert AJ, Tingley W, Elashoff MR, Rosenberg S, Buellesfeld L, Grube E, Newby LK, Ginsburg GS *et al*: Correlation of Peripheral-Blood Gene Expression With the Extent of Coronary Artery Stenosis. *Circ Cardiovasc Genet* 2008, 1:31-38.

46. Giorgi FM, Bolger AM, Lohse M, Usadel B: Algorithm-driven Artifacts in median polish summarization of Microarray data. *BMC Bioinformatics* 2010, 11:553.

47. Saeed A, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M *et al*: TM4: A Free, Open-Source System for Microarray Data Analysis. *Bio Techniques* 2003, 34:374-378.

48. Mieczkowski J, Tyburczy ME, Dabrowski M, Pokarowski P: Probe set filtering increases correlation between Affymetrix GeneChip and qRT-PCR expression measurements. *BMC Bioinformatics* 2010, 11:104.

49. Dheda K, Hugget JF, Bustin SA, Johnson MA, Rook G, Zumla A: Validation of housekeeping genes for normalizing RNA expression in real-time PCR. *BioTechniques* 2004, 37(1):112-119.

50. Sioson AA, Mane SP, Li P, Sha W, Heath LS, Bohnert HJ, Grene R: The statistics of identifying differentially expressed genes in Expresso and TM4: a comparison. *BMC Bioinformatics* 2006, 7:215.

51. King N: Methods in Molecular Biology: RT-PCR Protocols, 2nd edn. New York Dordrecht Heidelberg London: Humana Press; 2010.

52. Stonebraker LAR M, Hirohama M: The Design of POSTGRES. In: *IEEE Transactions on Knowledge and Data Engineering* 1986, 8.0.3 edn.

53. Rossum G: Python. In: *Pythonorg.*

54. PGAdmin III http://www.pgadmin.org/

55. R Development Core Team: R: A Language and Environment for Statistical Computing. In: *R Foundation for Statistical Computing.* Vienna, Austria; 2009.

56. TM4: http://www.tm4.org/.

57. Storey JD, Tibshirani R: SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: *The Analysis of Gene Expression Data: Methods and Software.* Edited by Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. New York: Springer; 2003.

58. Agilent: http://www.home.agilent.com/.

59. Kerr KF, Serikawa KA, Wei C, Peters MA, Bumgarner RE: What Is the Best Reference RNA? And Other Questions Regarding the Design and Analysis of Two-color Microarray Experiments. *OMICS* 2007, 11(2):152-165.

60. Cogenics a Division of Clinical Data Inc.: http://www.clda.com.

61. TimeLogic-Decypher system: http://www.timelogic.com/.

62. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ: Assessment of sensitivity and specificity of oligonucletide(50mer) microarrays. *NAR* 2000, 28:4542-4557.

63. dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/.

64. Bevilacqua PC, SantaLucia JJ: The biophysics of RNA. *ACS Chem Biol* 2007, 2(7):440-444.

65. SantaLucia JJ, Allawi HT, Seneviratne PA: Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 1996, 35(11):3555-3562.

66. SantaLucia JJ, Hicks D: The thermodynamics of DNA structural motifs. *Annu Rev Biophys Biomol Struct* 2004, 33:415-440.

67. Levy A, Sela N, Ast G: TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Res* 2007:1-6.

68. Royston P: An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics* 1982, 31:115-124.

69. Royston P: Algorithm AS 181: The W test for Normality. *Applied Statistics* 1982, 31:176-180.

70. Royston P: Remark AS R94: A remark on Algorithm AS 181: The W test for normality. *Applied Statistics* 1995, 44:547-551.

71. Bauer DF: Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* 1972, 67:687-690.

72. Myles Hollander, Wolfe DA: Nonparametric Statistical Methods. In: *Nonparametric Statistical Methods.* Edited by Sons JW. New York: John Wiley & Sons; 1973: 27-33 (one-sample), 68-75 (two-sample).

73. Storey JD, Tibshirani R: Statistical significance for genome-wide studies. *PNAS* 2003, 100(16):9440-9445.

74. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995, 57(1):125-133.

75. Mougeot JLC, Li Z, Price AE, Wright FA, Brooks BR: Microarray analysis of peripheral blood lymphocytes from ALS patients and the SAFE detection of the KEGG ALS pathway. *BMC Medical Genomics* 2011, 4(74).

76. Flikka K, Yadetie F, Laegreid A, Jonassen I: XHM: a system for detection of potential cross hybridizations in DNA microarrays. *BMC Bioinformatics* 2004, 5:1117.

77. Wren JD, Kulkarni A, Joslin J, Butow RA, Garner HR: Cross-hybridization on PCR-spotted microarrays. *IEEE Eng Med Biol Mag* 2002, 21(2):71-75.

78. Abdi H: Bonferroni and Sidak corrections for multiple comparisons. In: *Encyclopedia of Measurement and Statistics.* Thousand Oaks, CA: Sage: N.J. Salkind (ed.); 2007.

79. Reiner A, Yekutieli D, Benjamini Y: Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003, 19(3):368-375.

80. BIO-RAD: http://www.bio-rad.com/.

81. Pfaffl MW: A new mathematical model for relative quantification in real-time RT- PCR. *NAR* 2001, 29:e45.

82. ThermoScientific: http://www.nanodrop.com/.

83. Qiagen: http://www.qiagen.com.

84. System Biosciences: http://www.systembio.com/.

85. Sambrook, Russel: Molecular Cloning: a laboratory manual, 3rd edn. New York: Cold Spring Harbor; 2001.

86. Invitrogen: http://www.invitrogen.com/.

87. BioLabs: http://www.neb.com/.

88. AppliedBiosystems: http://www.appliedbiosystems.com/absite/us/en/home.html.

89. Primer3: http://frodo.wi.mit.edu/primer3/.

90. Primer-BLAST: http://www.ncbi.nlm.nih.gov/tools/primer-blast/.

91. Eurofins mwg | operon: http://www.operon.com/.

92. Mougeot JL, Price EA, Wright AF, Brooks BR: Microarray analysis of peripheral blood lymphocytes from ALS patients and the SAFE detection of the KEGG ALS pathway. *BMC Medical Genomics* 2011, 4:74.

93. Moffatt MF, *et al*.: Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 2007, 448:470-473.

94. Kingsmore SF, *et al.*: Genome-wide association studies: Progress and potential for drug discovery and development. *Nature Reviews Drug Discovery* 2008, 7:221-230.

95. Sha Q, Zhang Z, Schymick JC, Traynor BJ, Zhang S: Genome-wide association reveals three SNPs associated with sporadic amyotrophic lateral sclerosis through a two-locus analysis. *BMC Medical Genetics* 2009, 10:86.

96. Manolio TA, Brooks DL, Collins SF: A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008, 118:1590-1605.

97. Taylor JG, Choi EH, Foster CB, Chanock SJ: Using genetic varation to study human disease. *Trends Mol Med* 2001, 7(11):507-512.

98. Sonnemann KJ, Fitzsimons DP, Patel JR, Liu Y, Schneider MF, Moss RL, Ervasti JM: Cytoplasmic gamma-actin is not required for skeletal muscle development but its absence leads to a progressive myopathy. . *Dev Cell* 2006, 11:387-397.

99. Cunningham BA, Wang JL, Berggard I, Peterson PA: The complete amino acid sequence of beta-2-microglobulin. *Biochem* 1973, 12:4811-4821.

100. Fischer P, Götz ME, Danielczyk W, Gsell W, Riederer P: Blood transferrin and ferritin in Alzheimer's disease. *Life Sci* 1997, 60(25):2273-2278.

101. Ohta E, Takiyama Y: MRI Findings in Neuroferritinopathy. *Neurology Research International* 2011, 2012.

102. Mayr JA, Merkel O, Kohlwein SD, Gebhardt BR, Bohles H, Fotschl U, Koch J, Jaksch M, Lochmuller H, Horvath R *et al*: Mitochondrial phosphate-carrier deficiency: a novel disorder of oxidative phosphorylation. *Am J Hum Genet* 2007, 80:478-484.

103. Trotti D, Aoki M, Pasinelli P, Berger UV, Danbolt NC, Brown RH, Jr., Hediger MA: Amyotrophic lateral sclerosis-linked glutamate transporter mutant has impaired glutamate clearance capacity. . *J Biol Chem* 2001, 276:576-582.

104. Rothstein JD, Martin LJ, Kuncl RW: Decreased glutamate transport by the brain and spinal cord in amyotrophic lateral sclerosis. *New Eng J Med* 1992, 326:1464-1468.

105. Rothstein JD, Van Kammen M, Levey AI, Martin LJ, Kuncl RW: Selective loss of glial glutamate transporter GLT-1 in amyotrophic lateral sclerosis. . *Ann Neurol* 1995, 38:73-84.

106. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Nati Acad Sci USA* 1977, 74(12):5463-5467.

107. FinchTV: http://www.geospiza.com/Products/finchtv.shtml.

108. Shah HS, Pallas AJ: Identifying differential exon splicing using linear models and correlation coefficients. *BMC Bioinformatics* 2009, 10:26.

109. Morin M, Bryan KE, Mayo-Merino F, Goodyear R, Mencia A, Modamio-Hoybjor S, del Castillo I, Cabalka JM, Richardson G, Moreno F *et al*: In vivo and in vitro effects of two novel gamma-actin (ACTG1) mutations that cause DFNA20/26 hearing impairment. *Hum Molec Genet* 2009, 18:3075-3089.

110. Rendtorff ND, Zhu M, Fagerheim T, Antal TL, Jones M, Teslovich TM, Gillanders EM, Barmada M, Teig E, Trent JM *et al*: A novel missense mutation in ACTG1 causes dominant deafness in a Norwegian DFNA20/26 family, but ACTG1 mutations are not frequent among families with hereditary hearing impairment. *Europ J Hum Genet* 2006, 14:1097-1105.

111. DeWan AT, Parrado AR, Leal SM: A second kindred linked to DFNA20 (17q25.3) reduces the genetic interval. . *Clin Genet* 2003, 63:39-45.

112. Zhu M, Yang T, Wei S, DeWan AT, Morell RJ, Elfenbein JL, Fisher RA, Leal SM, Smith RJH, Friderici KH: Mutations in the gamma-actin gene (ACTG1) are associated with dominant progressive deafness (DFNA20/26). *Am J Hum Genet* 2003, 73:1082-1091.

113. Bugrim A, Nikolskaya T, Nikolsky Y: Early prediction of drug metabolism and toxicity: systems biology approach and modeling. . *Drug DiscovToday* 2004, 9:127–135.

114. Hassan SS, Romero R, Tarca LA, Draghici S, Pinels B, Bugrim A, Khalek N, Camacho N, Mittal P, Yoon BH *et al*: Signature pathways identified from gene expression profiles in the human uterine cervix before and after spontaneous term parturition *Am J Obstet Gynecol* 2007, 197(3): 250.e251–250.e257.

115. Nikolsky Y, Kirillov E, Zuev R, Rakhmatulin E, Nikolskaya T: Functional Analysis of OMICs Data and Small Molecule Compounds in an Integrated ‘‘Knowledge-Based’’ Platform. In: *Protein Network and Pathway Analysis.* Edited by Nikolsky Y, Bryant J, vol. 563: Humana Press; 2009: 177-196.

116. Lederer WC, Torrisi A, Pantelidou M, Santama N, Cavallaro S: Pathways and genes differentially expressed in the motor cortex of patients with sporadic amyotrophic lateral sclerosis. *BMC Genomics* 2007, 8:26.

117. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane HC, Lempicki R: David: Database for annotation, visualization, and integrated discovery. *Genome Biology* 2003, 4(5):3.This is the first version of this article to be made available publicly. A peer-reviewed and modified version is now available in full at http://genomebiology.com/2003/2004/2009/R2060.

118. Huang DW, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols* 2008, 4:44-57.

119. Kudo CL, Parfenova L, Vi N, Lau K, Pomakian J, Valdmanis P, Rouleau AG, Vinters VH, Wiedau-Pazos M, Karsten LS: Integrative gene-tissue microarray based approach for identification of human disease biomarkers: application to amyotrophic lateral sclerosis. *Hum Mol Genet* 2010, 19(16):3233-3253.

120. Gandhi KS, McKay FC, Cox M, Riveros C, Armstrong N, Heard RN, Vucic S, Williams DW, Stankovich J: The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. *Hum Mol Genet* 2010:10.1093/hmg/ddq1090.

121. Grunblatt E, Bartl J, Zehetmayer S, Ringel TM, Bauer P, Riederer P, Jacob CP: Gene expression as peripheral biomarkers for sporadic Alzheimer's disease. *J Alzheimers Dis* 2009, 16:627-634.

122. Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV *et al*: Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *PNAS* 2005, 102(31):11023-11028.

123. Saris CG, Horvath S, van Vught PW, van Es MA, Blauw HM, Fuller TF, Langfelder P, DeYoung J, Wokke JH, Veldink JH *et al*: Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients. *BMC Genomics* 2009, 10:405.

124. Lincecum JM, Vieira FG, Wang MZ, Thompson K, De Zutter GS, Kidd J, Moreno A, Sanchez R, Carrion IJ, Levine BA *et al*: From transcriptome analysis to therapeutic anti-CD40L treatment in the SOD1 model of amyotrophic lateral sclerosis. *Nat Genet* 2010, 42:392-399.

125. Lin J, Diamanduros A, Chowdhury SA, Scelsa S, Latov N, Sadiq SA: Specific electron transport chain abnormalities in amyotrophic lateral sclerosis. *J Neurol* 2009, 256(5):774-782.

126. Zhang R, Hadlock KG, Do H, Yu S, Honrada R, Champion S, Forshew D, Madison C, Katz J, Miller RG *et al*: Gene expression profiling in peripheral blood mononuclear cells from patients with sporadic amyotrophic lateral sclerosis (sALS). . *J Neuroimmunol* 2011, 230(1-2):114-123.

127. Lyons PA, Koukoulaki M, Hatton A, Doggett K, Woffendin HB, Chaudhry AN, Smith KG: Microarray analysis of human leucocyte subsets: the advantages of positive selection and rapid purification. *BMC Genomics* 2007, 8:64.

128. McKinney EF, Lyons PA, Carr EJ, Hollis JL, Jayne DR, Willcocks LC, Koukoulaki M, Brazma A, Jovanovic V, Kemeny DM *et al*: A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nat Med* 2010, 16(5):586-591.

129. Lyons PA, McKinney EF, Rayner TF, Hatton A, Woffendin HB, Koukoulaki M, Freeman TC, Jayne DR, Chaudhry AN, Smith KG: Novel expression signatures

identified by transcriptional analysis of separated leucocyte subsets in systemic lupus erythematosus and vasculitis. *Ann Rheum Dis* 2010, 69(6):1208-1213.

130. Gagliardi S, Cova E, Davin A, Guareschi S, Abel K, Alvisi E, Laforenza U, Ghidoni R, Cashman JR, Ceroni M *et al*: SOD1 mRNA expression in sporadic amyotrophic lateral sclerosis. *Neurobiol Dis* 2010, 39(2):198-203.

131. Pemov A, Park C, Reilly MK, Stewart RD: Evidence of perturbations of cell cycle and DNA repair pathways as a consequence of human and murine NF1-haploinsufficiency. *BMC Genomics* 2010, 11:194.

132. Ekins et al.: Pathway Mapping Tools for Analysis of High Content Data. In: *High content screening: a powerful approach to systems cell bilogy and drug discovery.* Edited by Taylor DL, Haskins RJ, Giuliano KA. Totowa, New Jersey: Humana Press; 2007: 322-349.

133. Chicheportiche Y, Bourdon PR, Xu H, Hsu YM, Scott H, Hession C, Garcia I, Browning JL: TWEAK, a new secreted ligand in the tumor necrosis factor family that weakly induces apoptosis. *J Biol Chem* 1997, 272(51):32401-32410.

134. Wiley SR, Winkles JA: TWEAK, a member of the TNF superfamily, is a multifunctional cytokine that binds the TweakR/Fn14 receptor. *Cytokine Growth Factor Rev* 2003, 14(3-4):241-249.

135. Seo H, Isacson O: The hAPP-YAC transgenic model has elevated UPS activity in the frontal cortex similar to Alzheimer's disease and Down's syndrome. *J Neurochem* 2010, 114(6):1819-1826.

136. Blais A, Dynlacht DB: Constructing transcriptional regulatory networks. *Genes Dev* 2005, 19:1499-1511.

137. Iacobazzi V, Infantino V, Costanzo P, Izzo P, Palmieri F: Functional analysis of the promoter of the mitochondrial phosphate carrier human gene: identification of activator and repressor elements and their transcription factors. *Biochem* 2005, 391:613-621.

138. Danko GC, Pertsov MA: Identification of gene co-regulatory modules and associated cis-elements involved in degenerative heart disease. *BMC Medical Genetics* 2009, 2:31.

139. Santpere G, Nieto M, Puig B, Ferrer I: Abnormal Sp1 transcription factor expression in Alzheimer disease and tauopathies. *Neurosci Lett* 2006, 397(1-2):30-34.

140. Bellingham SA, Coleman LA, Masters CL, Camakaris J, Hill AF: Regulation of prion gene expression by transcription factors SP1 and metal transcription factor-1. *J Biol Chem* 2009, 284(9):129-301.

141. Larramendy ML, Niini T, Elonen E, Nagy B, Ollila J, Vihinen M, Knuutila S: Overexpression of translocation-associated fusion genes of FGFRI, MYC, NPMI, and DEK, but absence of the translocations in acute myeloid leukemia. A microarray analysis. *Haematologica* 2002, 87(6):569-567.

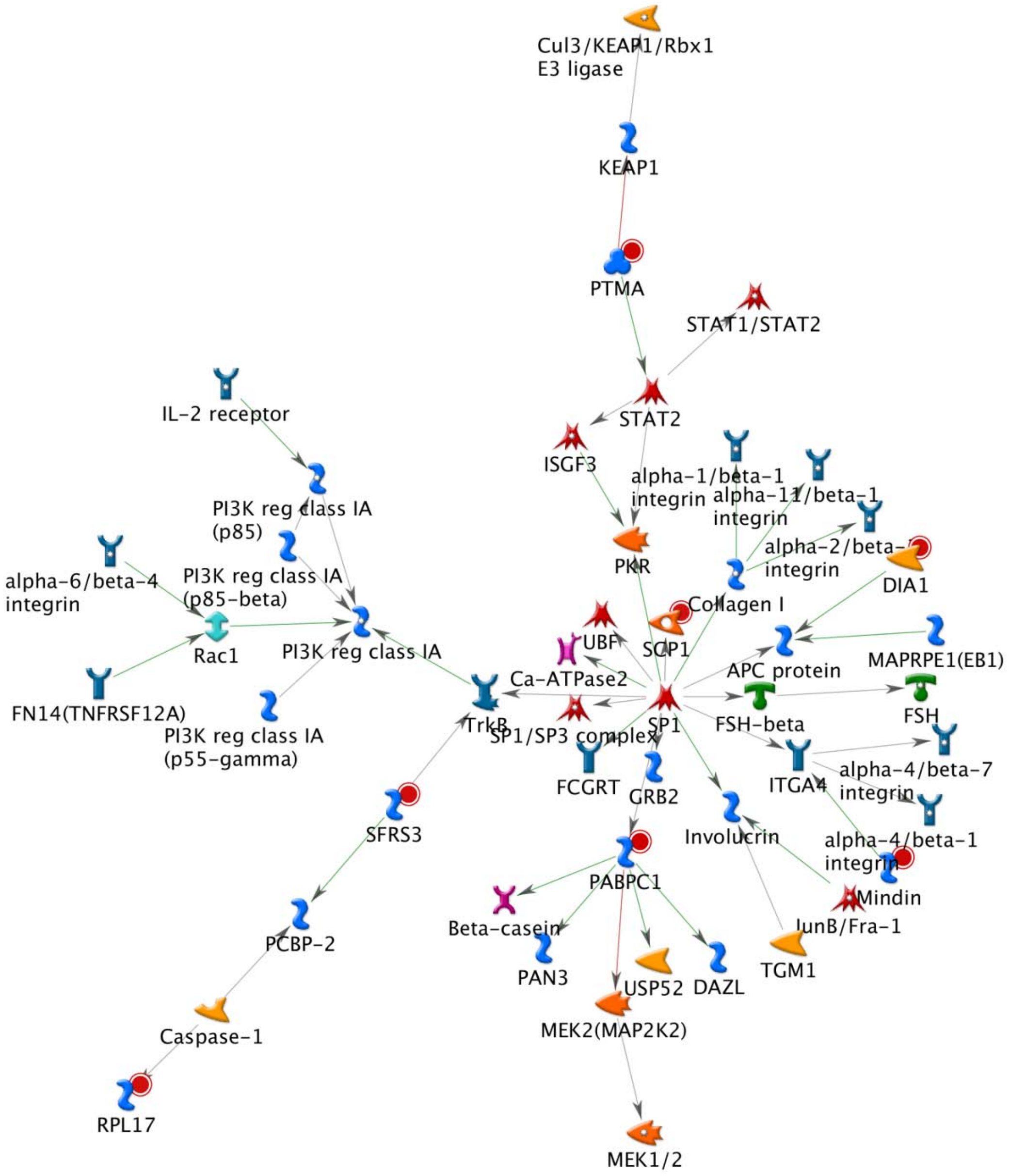
142. Blancato J, Singh B, Liu A, Liao DJ, Dickson RB: Correlation of amplification and overexpression of the c-myc oncogene in high-grade breast cancer: FISH, in situ hybridisation and immunohistochemical analyses. *Br J Cancer* 2004, 90(8):1612-1619.

143. Ikeguchi M, Hirooka Y: Expression of c-myc mRNA in hepatocellular carcinomas, noncancerous livers, and normal livers. *Pathobiology* 2004, 71(15):281-286.

144. Chikh A, Matin RN, Senatore V, Hufbauer M, Lavery D, Raimondi C, Ostano P, Mello-Grand M, Ghimenti C, Bahta A *et al*: iASPP/p63 autoregulatory feedback loop is required for the homeostasis of stratified epithelia. *EMBO J* 2011, 30(20):4261-4273.

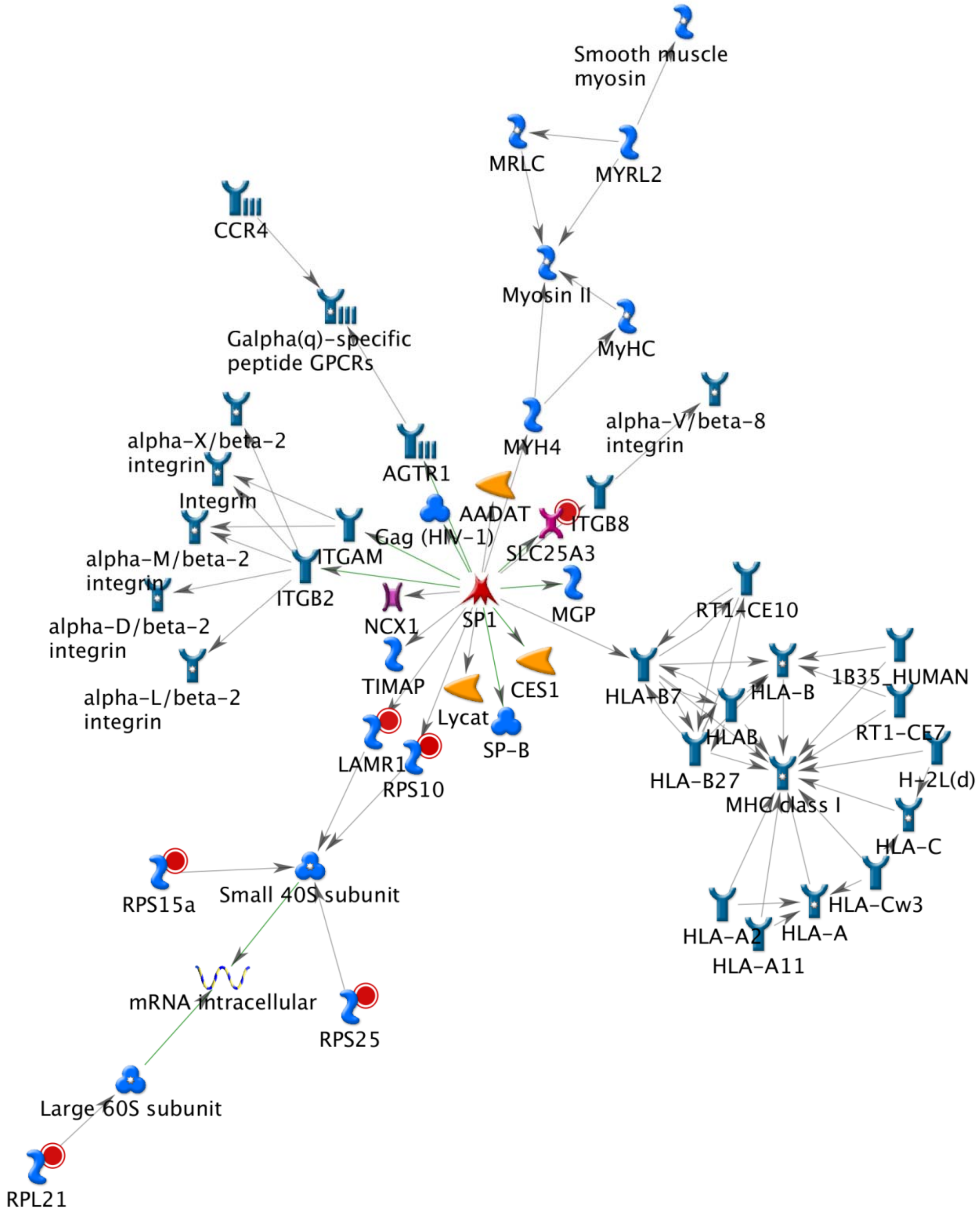
145. Zammarchi F, de Stanchina E, Bournazou E, Supakorndej T, Martires K, Riedel E, Corben AD, Bromberg JF, Cartegni L: Antitumorigenic potential of STAT3 alternative splicing modulation. *Proc Nati Acad Sci USA* 2011, 108(43):17779-17784.

APPENDIX A: ADDITIONAL NETWORK INTERACTIONS



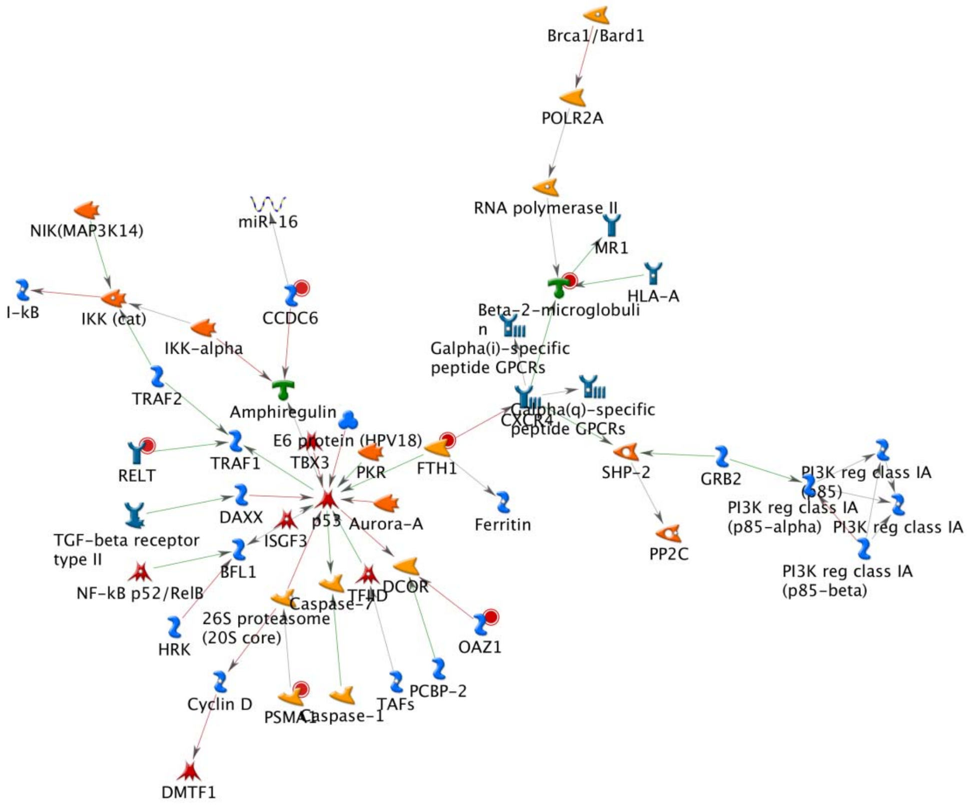
(i)

Figure A1. Other relevant networks (from Table 4.1) for our selection of genes. (i) network 2, (ii) network 3;(iii) network 4; (iv) network 5; (v) network 6.

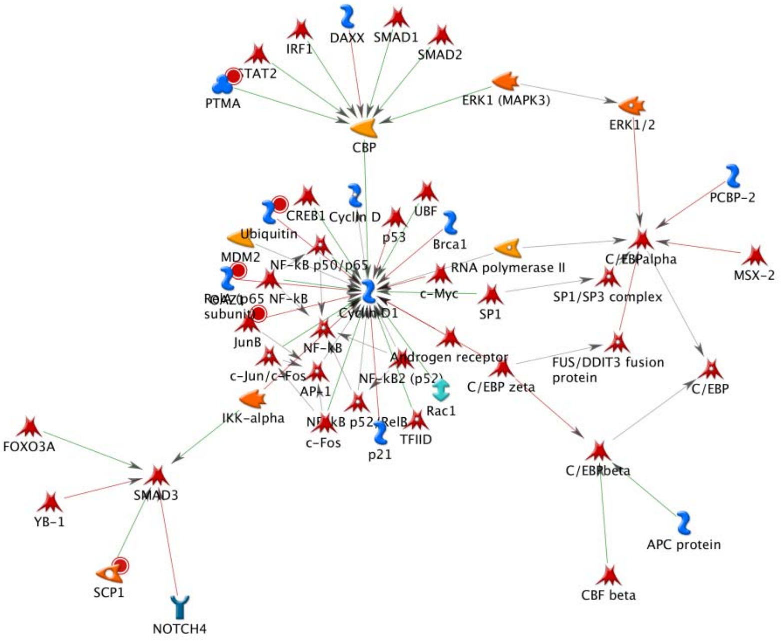


(ii)

Figure A1 (continued).



(iii)



(iv)

Figure A1 (continued).

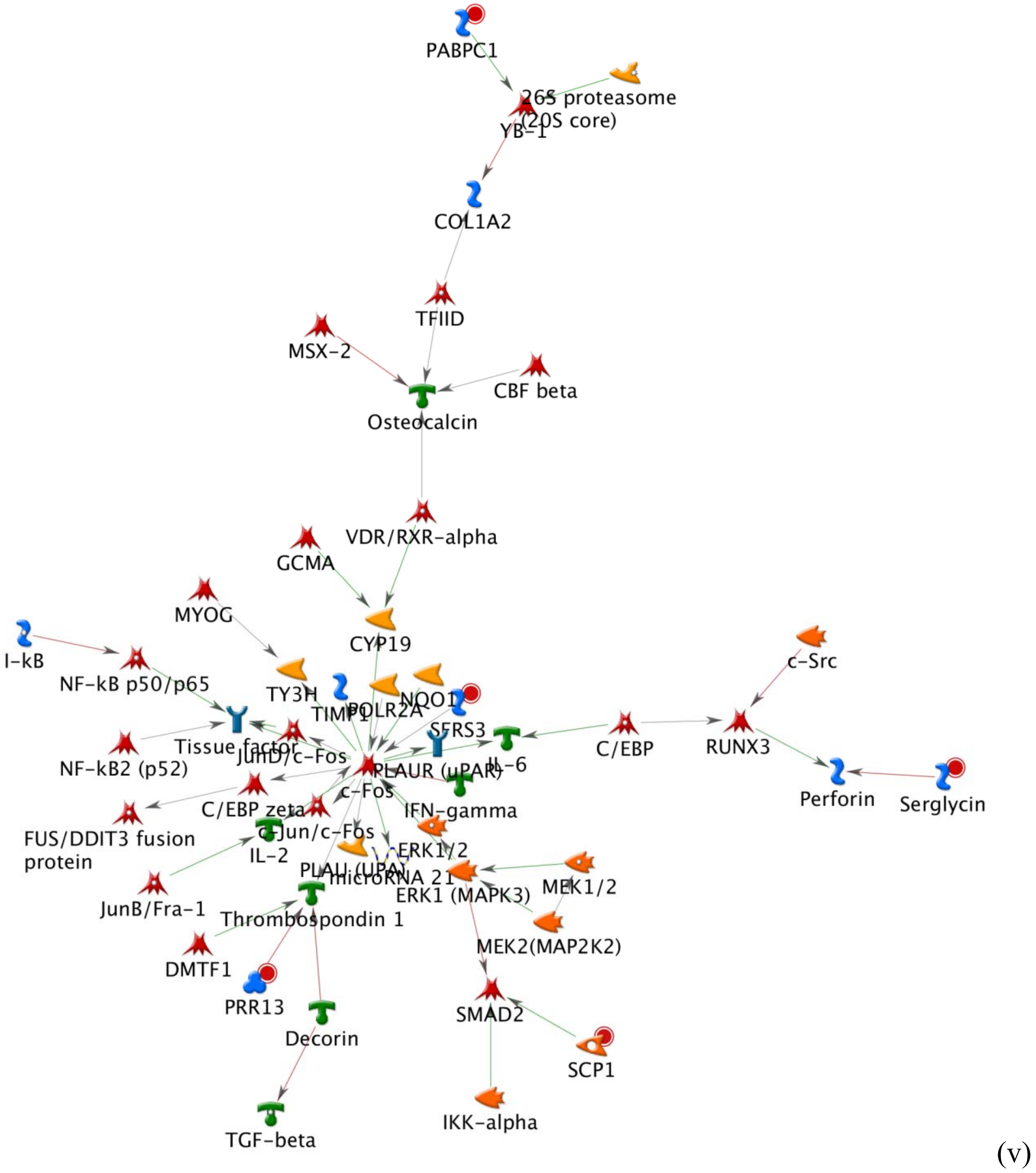


Figure A1 (continued).

VITA

Name: Cristina Baciu

Place of Birth: Buhusi, Romania

Year of Birth: 1966

Education

- *University of North Carolina - Charlotte*

Ph.D. candidate, Bioinformatics and Genomics (2007–present)

- *University of Windsor, Windsor, ON, Canada*

M. S. Biochemistry, 2005

- *Technical University Ghe. Asachi, IASI, Romania*

B. S. Chemical Engineering, 1993

Academic Experience

2007- present Graduate Research Assistant

UNC Charlotte, Department of Bioinformatics and Genomics

- Working on dissertation: Bioinformatics and Biomolecular Tools for the discovery of disease-onset and progression biomarkers in peripheral blood lymphocytes from patients with Sporadic Amyotrophic Lateral Sclerosis.

Sept. – Dec. 2009 Teaching Assistant

UNC Charlotte, Department of Bioinformatics and Genomics

- Design and Implementation of Bioinformatics Databases.

2001-2004 Graduate Research Assistant

University of Windsor, Windsor, ON, Canada

- Advisor: Dr. J. W. Gauld

Thesis**:** Computational Investigations on S-nitrosothiols (RSNOs)

Ab initio calculations and Density Functional Theory (DFT) methods have been used to do an assessment of theoretical methods for obtaining accurate structures and RS-NO homolytic Bond Dissociation Energy (BDE) of a variety of S-nitrosothiols.

2001-2004 Teaching Assistant

University of Windsor, Windsor, ON, Canada

- General Chemistry Laboratory, Organic Chemistry Laboratory
- Assisted students with labs and course material.
- Demonstrated lab techniques (wet chemistry) and marked exams.
- Assisted students with tutorials.

Engineering experience

2005- 2007 Quality Control Engineer

Talhin/T, Oldcastle, ON, Canada

- Inspect manufactured plastic products for defects and conformance to specifications and quality standards, visually or using instruments.
- Maintain inspection records and complete inspection reports on products inspected.

1993-2000 Production Engineer

Stofe Buhusi S.A., Buhusi, Romania

- Supervising the production activity of the chemical finishing operations of textile fabrics, from pre-washing, washing, dyeing, thermostability assurance and softening operations to quality control of finished materials.

- Working in close contact with customers, in order to meet their quality exigencies.

Research interest

- Identifying bioinformatics methods and pipelines for correcting errors that may occur in microarray analysis.
- Biomarker discovery of different diseases.
- Relationa databases.
- Bioinformatics study associated with Single Nucleotide Polymorfism (SNPs)
- High-throughput genomic research.
- Pathway analysis.

Computational Skills

- Extensive experience in using Linux/Unix/Windows operating systems on Mac OSX and PC.
- Proficient in R, dabase PostgreSQL, SQL scripting.
- Pathway and protein network analysis (MetaCore, GeneGo, STRING)
- Protein modeling using FIRST (Floppy Inclusion and Rigid Structure Topology) application.
- Proficient with computational chemistry software applications: Gaussian, Jaguar, Molekel, Spartan.
- Familiar with network administration.

Laboratory Skills

- General molecular biology lab techniques (RNA/DNA extraction and purification, PCR, qRT-PCR, gel electrophoresis, primer design, etc).

- Sanger sequencing.
- Next generation sequencing (IonTorrent Technology).
- General chemistry techniques.

Publications

- Robinet, J. J.; Baciu, C.; Cho, K. B.; Gauld, J. W. "A Computational Study on The Interaction of Nitric Oxide Ions NO^+ and NO^- With Aromatic Amino Acids" J. Phys. Chem. A, 2007, *111 (10),* 1981-1989.

- Baciu, C.; Cho, K-B.; Gauld, J. W. "Influence of Cu^+ on the RS-NO Bond Dissociation Energy of S-nitrosothiols" *J. Phys. Chem. B.* 2005, *109*, 1334-1336.

- Baciu, C.; Cho, K-B.; Gauld, J. W. "Ring Complexes of S-nitrosothiols (RSNOs) with Cu^+ : A Density Functional Study" *Eur. J. Mass Spectrom.* 2004**,** *10*, 941-948.

- Baciu, C.; Gauld, W. J. " An Assessment of Theoretical Methods for the Calculation of Accurate Structures and S-N Bond Dissociation Energies of S-nitrosothiols" *J. Phys. Chem. A*, 2003, *107*, 9946-9952.

Manuscripts in preparation

- *Cristina Baciu*, Kevin J Thompson , Jean-Luc Mougeot and Jennifer W Weller "The LO-BaFL pipeline for microarray expression analysis"; target paper: BMC Bioinformatics

- *Cristina Baciu*, Kevin J Thompson and Jennifer W Weller: "Background estimation and corrections in microarray analysis"; target paper: BMC Bioinformatics.

Scientific Presentations at Conferences

- *Cristina Baciu*, Jean-Luc Mogeot and Jennifer W Weller "The LO-BaFL pipeline for microarray expression analysis"-Poster presentation at ISMB in Boston, 2010.

- A. Carr, C. Overall, *C. Baciu* and J. Weller "SNPs: The Study of Single Nucleotide Polymorfisms from a computational perspective using Affymetrix's SNP 6.0 Platform"-Poster presentation at inauguration of Department of Bioinformatics and Genomics, UNCC, 2008.

- *C. Baciu*, M.Henry, J. Robinet, Q. Jin_and J. W. Gauld "Interaction of NO and Its Ions With Aromatic Biomolecules"-Poster presentation at IUPAC and CSC Conference, London, ON, 2004.

- *C. Baciu* and J. W. Gauld "Computational Studies on Copper-Nitrosothiol Complexes" - Poster presentation at IUPAC and CSC Conference, Ottawa, ON, 2003.

- *C. Baciu* and J. W. Gauld "Theoretical Studies of S-nitrosothiols" -Oral presentation, Chemical Biology Discussion Weekend, University of Windsor, ON, 2003.

- *C. Baciu* and J. W. Gauld "Theoretical Studies of S-nitrosothiols" - Poster presentation at CSC Conference, Vancouver, BC, 2002.

Professional Affiliations & Memberships

- International Society for Computational Biology (2010-present)

- Canadian Society for Chemistry (2001-2005)

- Chemical Institute of Canada (2001-2005)

Awards and Scholarships

- John & Anne Cristescu Memorial Scholarship, for scholar year 2001-2002.