Applying Gene Expression, Proteomics and SNP Analysis for Complex Trait Gene Identification


* The Jackson Laboratory, Bar Harbor, ME 04609
† Applied Biosystems, Framingham, MA 01701, USA

¹ Present address: University of Pennsylvania, School of Medicine, Institute for Translational Medicine and Therapeutics, Philadelphia, PA 19104-6160
² Present address: 454 Life Sciences, Branford, CT 06405

Microarray submission, GEO database series reference: GSE4765.
Short title (running head): Mining for QTL genes

Keywords: QTL, complex traits, gallstones, obesity, haplotype, microarray, HDL.

3 Corresponding Author:

Beverly Paigen
The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04609, USA
Phone: 207-288-6388
Fax: 207-288-6078
Email: bev.paigen@jax.org
ABSTRACT

Previous quantitative trait locus (QTL) analysis of an intercross involving the inbred mouse strains NZB/BlNJ and SM/J revealed QTL for a variety of complex traits. Many QTL have large intervals containing hundreds of genes, and methods are needed to rapidly sort through these genes for probable candidates. We chose nine QTL; the three most significant for HDL-cholesterol, gallstone formation, and obesity. We searched for candidate genes using three different approaches; mRNA microarray gene expression technology to assess over 45,000 transcripts, publicly available SNPs to locate genes that are not identical by descent and that contain non-synonymous coding differences, and a mass-spectrometry-based proteomics technology to interrogate nearly 1000 proteins for differential expression in the liver of the two parental inbred strains. This systematic approach reduced the number of candidate genes within each QTL from hundreds to a manageable list. Each of the three approaches selected candidates that the other two approaches missed. For example, candidate genes such as Apoa2 and Acads had differential protein levels although the mRNA levels were similar. We conclude that all three approaches are important and that focusing on a single approach such as mRNA expression may fail to identify a QTL gene.
INTRODUCTION

The use of animal models and quantitative trait locus (QTL) mapping has greatly aided the identification of genes that affect complex traits (KORSTANJE and PAIGEN 2002). However, identifying the causal gene within a QTL remains a non-trivial matter. Whole transcriptome profiling has advanced the gene discovery process by identifying causal genes on the basis of differential mRNA levels (AITMAN et al. 1999; JOHANNESSON et al. 2005). Nonetheless, a QTL may be caused by differences that are not identifiable at the mRNA level, such as protein concentration (DOOLITTLE et al. 1990) or protein-function caused by changes in the amino acid sequence (THEUNS et al. 2000).

In this study we combine mRNA expression with proteomics and SNP analysis to show that using a global mRNA approach alone is insufficient for identifying QTL candidate genes. Proteomic technologies are still in their infancy relative to mRNA gene expression technologies; nonetheless the ability to interrogate the proteome is improving (PILCH and MANN 2006), and as we report here, valuable data can be gathered and applied to the complex disease gene discovery process. In addition, mouse strains have been extensively genotyped; indeed, many of the SNPs between strains are known and are publicly available. SNP data can be used to focus on genes in genomic regions that are not identical by descent (IBD) (PETERS et al. 2007) and to identify amino acid changes in the protein-coding region. This allows the selection of candidates that may affect the QTL through functional changes rather than quantitative differences in mRNA and protein levels.

Typically, large-scale protein interrogation studies involve either two-dimensional (2D) gel electrophoresis followed by mass spectrometry (MS) or large scale immunoblotting (BD Biosciences, CA). However, these technologies have limitations. The proteins that can be
interrogated on 2D-gels are limited in size and nature and require considerable effort to identify. Immunoblotting is limited by the availability and cost of antibodies. For the identification of larger scale complex proteomes, a technology for proteome analysis by MS has emerged (AEBERSOLD and MANN 2003). This method can be extended by coupling liquid chromatography with matrix assisted laser induced ionization tandem time of flight mass spectrometry (MALDI-TOF/TOF MS). Relative and absolute quantitation of proteins by MS is possible with either labeled (ROSS et al. 2004; ZHEN et al. 2004) or label-free (ONO et al. 2006) technologies.

The maturation of both microarray and proteomic technologies has led to a small number of whole transcriptome-proteome comparisons (GYGI et al. 1999; McREDMOND et al. 2004; MIJALSKI et al. 2005; PASTORELLI et al. 2006; RIBEIRO et al. 2004; RUSE et al. 2004; VALENZUELA et al. 2003). However, limited quantitative analysis of protein and transcript abundance could be performed as the starting point of these proteomics studies was 2D electrophoresis followed by single protein-by-protein identification using MS. In this study, to complement the microarray data, we employed a quantitative proteomics approach using isobaric tags for relative and absolute quantitation (iTRAQ™ reagents) in combination with liquid chromatography-MALDI TOF/TOF technology for extensive protein analysis and quantitation of two inbred mouse strains. We were able to directly identify and simultaneously quantify a total of 943 liver proteins in a gel-free MS analysis. To our knowledge this is the first attempt to correlate differentially regulated proteins with QTL in order to identify candidate genes for complex quantitative phenotypes.

A systematic search for candidate genes must also include a search for non-synonymous coding changes. This is possible for 16 inbred mouse strains with the combination of the public sequencing of C57BL/6, the sequencing of A, DBA/2 and 129 by Celera (MURAL et al. 2002)
and the re-sequencing of 15 strains by Perlegen (http://mouse.perlegen.com/mouse/). Combining these databases will identify a large proportion of the coding differences. SM/J (SM) and NZB/B1NJ (NZB) are currently not on the list of such strains; however, the SNP data can be used to examine the haplotypes, and coding differences between SM and NZB can be inferred since mouse inbred strains share common ancestry (Wade et al. 2002). The SNP data identified many candidate genes that had neither mRNA nor protein abundance differences, yet differed at SNPs in the coding region and hence might affect protein function.

The mouse strains ‘Small’ (SM) (Chai 1956) and ‘New Zealand Black’ (NZB) (Bielschowsky and Goodall 1970) differ in a number of metabolic phenotypes (http://www.jax.org/phenome) (Bogue et al. 2007). These were used in a large intercross to yield QTL for a number of complex traits including lipid levels (Korstanje et al. 2004), body composition and size (Stylianou et al. 2006), and gallstone susceptibility (Lyons et al. 2005). In this study we evaluate genes within nine QTL for these traits using all three approaches; mRNA levels, protein levels, and amino acid-changing SNPs in the coding region. The number of candidate genes in each QTL was reduced from hundreds down to a list ranging from 10 to 88 genes depending on the QTL. Some genes were identified by more than one method, but each approach yielded unique genes showing that it is the combination of methods that provides the best list of candidate genes for further testing.

MATERIALS AND METHODS

**Mice, crosses and tissue collection:** SM and NZB inbred mice were obtained from The Jackson Laboratory, Bar Harbor, ME. Further details relating to the NZBxSM F2 cross and husbandry conditions have been published (Stylianou et al. 2006). All the QTL identified
between SM and NZB are present in both sexes except for two QTL that are female specific (Table S1); thus only females were interrogated on the microarray and proteomic platforms.

Liver samples from the large lobe of the liver of three female mice for each strain, NZB and SM, were dissected under conditions suitable for RNA. Following a 4hr morning fast, samples were collected and flash frozen in liquid nitrogen. Two liver samples from each mouse were taken and stored separately to be used for gene and protein analysis. Microarray experiments were performed using all three samples of each strain, but proteomic analysis was performed on two samples of each strain. Additional mice of each sex from both strains were dissected under the same conditions for Western blot analysis. All experiments were performed on 8 week old mice maintained on a standard chow diet and procedures were approved by The Jackson Laboratory’s Animal Care and Use Committee.

**SNP and haplotype analysis:** The 138,000 SNPs (~1SNP/24Kb) from the Broad Institute database were updated to NCBI Build 36 and used to compare the NZB and SM strains over the 9 QTL regions as defined by the 95% CI. Where the strains shared at least three common adjacent SNPs, we inferred that the region might be in a haplotype block and be identical by descent. The SNP feature of the Mouse Phenome Database was used to search these same QTL regions for all genes carrying a coding region difference that resulted in an amino acid change. These genes were then compared to the haplotype data, and those genes in regions identical between NZB and SM were eliminated from the list. Examining the candidate genes for association to their respective traits in published literature was performed as previously described (STYLIANOU et al. 2006).

**Microarrays:** RNA was extracted from frozen samples using TRIzol® as previously described (STYLIANOU et al. 2005), and the quality of the RNA was assessed using a 2100
Bioanalyzer instrument and RNA 6000 Nano LabChip assay (Agilent Technologies, Palo Alto, CA). Following reverse transcription with an oligo(dT)-T7 primer (Affymetrix, CA, USA), double-stranded cDNA was synthesized with the Superscript double-stranded cDNA synthesis custom kit (Invitrogen, USA). In an *in vitro* transcription (IVT) reaction with T7 RNA polymerase, the cDNA was linearly amplified and labeled with biotinylated nucleotides (Enzo Diagnostics, NY, USA). Fifteen micrograms of biotin-labeled and fragmented cRNA was then hybridized onto Mouse Genome 430 2.0 GeneChip® arrays (Affymetrix) for 16 hours at 45°C. Post-hybridization staining and washing were performed according to the manufacturer's protocols using the Fluidics Station 450 instrument (Affymetrix). Finally, the arrays were scanned with a GeneChip® Scanner 3000 laser confocal slide scanner. The data have been deposited in the GEO database (series reference: GSE4765).

Array images were quantified using GeneChip® Operating Software (GCOS) v1.2. Probe level data were imported into the R software environment and expression values were summarized using the RMA (Robust MultiChip Average) function (Irizarry *et al.* 2003a) in the R/affy package in Bioconductor (version 1.3.25, Irizarry, Gautier, and Bolstad). Using the R/maanova package (Wu *et al.* 2003), an analysis of variance (ANOVA) model was applied to the data. The model:

$$Y_i = \mu + \text{STRAIN} + \varepsilon_i$$  \[1\]

was used to fit the log-transformed gene expression measures $Y_i$, where $\mu$ is the mean for each array, STRAIN is the strain effect and $\varepsilon_i$ captures random error. Strain differences were tested using $F_s$, a modified F-statistic incorporating shrinkage estimates of variance components (Cui *et al.* 2005). The proportion of false positives generated was estimated using a FDR adjustment described previously (Storey and Tibshirani 2003). Differences between strains were
determined at the level FDR < 0.01, which corresponded to permutation p-values less than 2.6 x 10^{-3}.

**TaqMan assays:** mRNA differences of a number of genes that were identified by microarrays were confirmed by TaqMan quantitative real-time PCR assays. Briefly, total RNA (1µg) was reverse transcribed using the High Capacity cDNA Archive kit (Applied Biosystems, CA, USA) according to manufacturer’s protocol. A portion of the cDNA was then used in a PCR reaction containing TaqMan Universal PCR Master Mix (Applied Biosystems) which includes AmpliTaq Gold DNA Polymerase, AmpErase UNG, dNTPs with dUTP, Passive Reference 1 and other buffer components. The gene-specific primers and probe sets were obtained from Applied Biosystems Assay on Demand service and used according to manufacturer’s protocols. Real-Time PCR was performed in an ABI PRISM® 7900HT Sequence Detection System (Applied Biosystems) with the standard protocol of 95°C for 10 minutes to activate the DNA polymerase followed by 40 cycles of amplification. The threshold cycle (Ct) was determined using the Sequence Detection System software (SDS2.2) and relative fold change ratios were calculated.

**Mass spectrometry:** Two samples for each strain were prepared in parallel. Liver lobes were diced with a sterile scalpel and placed into extraction buffer containing 500mM TEAB, 6M GuHCl, 0.1% Triton X-100, pH 8.5. DNA and tissue was sheared by using a 3cc syringe fitted with a 20-gauge needle, then sonicated for 20 minutes. Protein was separated from the buffer by the addition of 6X volume of ice-cold acetone. Samples were then incubated at -20°C for 2 hours, and then briefly spun before removing the acetone. The protein was solubilized in iTRAQ reagent dissolution buffer with SDS. Protein determination was done using bicinchoninic acid (BCA) (Pierce). 100µg of each sample was labeled with iTRAQ reagent. Reduction and cysteine
blocking was done according to standard protocol. Trypsin digestion was done in two steps, half of the recommended amount of trypsin was added to each sample and incubated for 2 hours at 37°C, the second half was then added and left overnight at 37°C. Labeling: the iTRAQ reagents were added to each of the samples as follows: SM No.1 with 114 (S1), SM No.2 with 115 (S2) NZB No.1 with 116 (N1), and NZB No.2 with 117 (N2). Labeling was performed according to the standard protocol. After labeling the samples were pooled and dried on a Speed-Vac. HPLC: iTRAQ Reagent labeled sample were diluted in Cation Exchange Loading Buffer to reduce the salt concentration. The pH was adjusted to 3.0 with 1N phosphoric acid. The sample was loaded onto o-Poly LC polysulfoethyl A 4.5x100mm, 5µm, 200 Å column using the Agilent 1100. UV at Å214 was measured. Samples were loaded at 0.5ml/min and the gradient was run at 1.0ml/min. The gradient conditions were 0% Buffer B (10mM sodium phosphate, 25% acetonitrile, 500mM KCl, pH 3.0) to 10% Buffer B 2 min, 20% Buffer B to 45% 2 min. 45% Buffer B to 100% 5 min. One ml fractions were collected and dried to completeness on Speed-Vac. Fractions were then subjected to reverse phase chromatography, combined with MALDI-matrix and spotted on a MALDI target plate. MS and MS/MS spectra were acquired using a 4800 Proteomics Analyzer (AB1) MALDI-TOF/TOF mass spectrometer.

Protein mass spectroscopy experiments were performed in duplicate and converted to ratios between samples (i.e., SM2:SM1, NZB1:SM1, NZB2:SM1). Protein measures (analogous to the transcript measures above) were derived from these ratios by first arbitrarily setting the SM1 value equal to one, calculating the corresponding values from the ratio relationships, log transforming each value (SM1, SM2, NZB1, NZB2) and re-centering the data by subtracting the column means. The statistical tests for differences in protein levels between strains were performed in the same manner as transcriptional differences except that the significance level
(determined by permutation) was set at p< 0.1. No significant differences would have been found using an FDR as for mRNA due to the limitations of the method (only 2 biological replicates).

**Western analysis for liver proteins:** Western analysis was performed to confirm differentially abundant candidate proteins identified from the MS analysis. Frozen livers were crushed on a stainless steel block while on dry ice and transferred to a glass homogenizer with a Teflon pestle containing lysis buffer (50mM Tris pH 8.0, 150mM NaCl, 1mM EDTA, 1% igepal CA-630 detergent, 1mM sodium fluoride, 1mM sodium vanadate and 1X Sigma protease inhibitor cocktail P2714). Following centrifugation at 16,000g for 15 minutes, the protein concentrations of the supernatants were evaluated using a BCA assay (Pierce, IL, USA).

Electrophoresis of 10µg of each lysate was performed using 4-12% Bis-Tris SDS-PAGE gels with MOPS buffer (Invitrogen). Proteins were transferred to 0.45µm polyvinylidene fluoride membranes (Invitrogen) using Invitrogen’s transfer buffer containing 10% methanol for 1hr at 30V and 25C. Membranes were blocked for 1hr in Tris-Buffered Saline Tween-20 with 10% skim milk. **ACADS:** Rabbit anti-ACADS serum (kindly provided by Dr. Vockley’s Lab from the Children’s Hospital of Pittsburg, PA.) was used at the dilution of 1:1000 with 5% skim milk for the primary immunostain. Goat anti-rabbit IgG horse radish peroxidase (HRP) 0.4mg/ml was used at 1:50,000 for the secondary (Pierce). **NDUFA9:** Mouse monoclonal against purified mitochondrial complex I of cow 200µg/ml (Santa Cruz Biotechnology, Inc, CA, USA) was used at the dilution of 1:500 with 5% skim milk for the primary immunostain. Goat anti-mouse IgG HRP 10µg/ml was used at 1:1,000 for the secondary stain (Pierce). **FH1:** Goat polyclonal IgG (Santa Cruz Biotechnology) was used at the dilution of 1:200 with 5% skim milk for the primary immunostain. Donkey anti-goat IgG H+L HRP 1mg/ml was used 1:5000 for the secondary stain.
HSD11B1: Rabbit polyclonal antibody (Abcam, MA, USA) to human sequence was used at 1:300 for the primary immunostain. Goat anti-rabbit IgG (H+L) HRP 0.4mg/ml was used at 1:50,000 dilution for the secondary stain. For a loading control, we used rabbit anti-B-TUBULIN conjugated with HRP at 1:200 dilution of 200mg/ml (Santa Cruz Biotechnology).

For each blot, 3 samples each from a different mouse were run for both strains (SM and NZB) and both sexes resulting in a total of 12 samples with an N=3 within sex-strain comparison and an N=6 within strain comparison. The immunoblots were immersed in Pierce’s SuperSignal West Dura HRP substrate and imaged using a FujiLas CCD camera. Quantitation was performed using ImageJ (Abramoff et al. 2004).

RESULTS

QTL analysis: Previous analysis of the SM x NZB intercross population identified 38 main-effect QTL for the phenotypes of HDL-cholesterol (Korstanje et al. 2004), gallstone susceptibility (Lyons et al. 2005), body weight, adiposity and obesity (Stylianou et al. 2006) including two HDL QTL affected by sex (Supplementary Table S1). In this report we focused on the three most significant QTL for each of these traits: Hdlq20, 1 and 24 on Chrs 1, 5 and 6 respectively for HDL; Lith17, 19 and 21 on Chrs 5, 8 and 10 respectively for gallstones; and Obwq3, 4, and 5 on Chrs 6, 17 and 19 respectively for obesity. We re-analyzed the QTL data from the original publications with updated marker positions based on mouse genome build 36 (Table 1). To obtain the 95% confidence intervals, we used posterior probability distribution as described previously (Sen and Churchill 2001).
Use of SNP data to reduce the QTL interval and identify genes with coding changes:

QTL regions were reduced by eliminating regions within the 95% confidence interval (CI) that appeared to be identical by descent (IBD) between the two parental strains; these regions are unlikely to contain the causal polymorphisms unless the mutation arose in the last 100 years since the strains were inbred. It is estimated that 97% of mutations are ancestral (Wade et al. 2002). Thus, 3% of mutations, and by inference 3% of QTL, may be due to a recent mutation unique to a single strain. If a mutation is recent, then the QTL is unlikely to be found except in a cross with that particular strain; thus, if the same QTL is found in more than one intercross, the mutation is more likely to be ancestral. For all QTL in this report except one (Lith19 on Chr 8), the QTL have been reported in other crosses thus increasing greatly the probability that the causal mutation is ancestral and shared among strains. We inferred the IBD regions by directly examining the common polymorphisms between the parental strains using 130,000 SNPs for SM and NZB from the Broad Institute (http://www.broad.mit.edu/mouse/snp_xls.html). These data were used to exclude large regions within each QTL and reduced the QTL genomic space by between 40-72%.

To identify genes that differed within the coding region of the gene, we utilized all the Perlegen SNP data in combination with all SNPs for SM and NZB available through the Mouse Phenome Database (www.jax.org/phenome) (Bogue et al. 2007). We inferred coding differences between SM and NZB for the nine QTL (Table 2). This resulted in a list of genes that may be causal for each QTL due to coding and hence functional differences (Supplementary Tables S2-S10).
Transcriptome analysis: Liver mRNA profiles were compared between three SM and three NZB female mice using the Affymetrix Mouse Genome 430 2.0 oligonucleotide GeneChip® arrays. Statistical diagnostic analysis indicates that the array data generated are highly replicable and hence robust. Genes that are not differentially expressed between the two strains have a characteristically even distribution of non-significant p-values (Figure 1A). Scatter plots of mRNA gene expression measures from the six arrays derived from robust multi-chip analysis (RMA) (IRIZARRY et al. 2003b) show that within strain comparisons are tightly distributed and between strain distributions show differential expression (Fig 1B). To validate the microarray results, we performed real-time PCR using TaqMan assays on 10 differentially expressed genes that had conservative absolute fold changes of less than 4.0 and false discovery rates (FDR) less than 1x10^{-4}. TaqMan assays were consistent with microarray data for all 10 genes; that is, all genes were differentially expressed in the expected direction and the differences were significant after correcting for multiple testing (Supplementary Table S11). Using the stringent FDR of less than 1x10^{-4}, we identified 831 differentially expressed (DE) probe sets.

Candidates based on differences in mRNA abundance: A number of differentially expressed genes were identified from the microarray analysis (Figure 2, Supplementary Tables S2-S10, S12). Several of these are of particular interest because there is evidence that they are involved with the target traits (Table 3). The cholesterol scavenger receptor class B member 1 (Scarb1) is a known receptor for HDL (ACTON et al. 1996; TRIGATTI et al. 2003), and therefore an obvious candidate for Hdlq1 on Chr 5; adiponectin receptor 2 (Adipor2) is a functional
candidate for Hdlq24 and Obwq3 on Chr 6. For gallstone QTL, the connective tissue growth factor (Ctgf), shown to have a 40-fold mRNA differential expression in the gallbladders of human cholecystolithiasis patients compared to controls (KONINGER et al. 2005), is a candidate for Lith21 on Chr10; also Scl10a2 on Chr 8 is essential for efficient intestinal absorption of bile acids and therefore may be responsible for the Lith19 QTL (DAWSON et al. 2003). Finally, for the obesity QTL on Chr 19, attractin like 1 (Atrnl1) has a 3.4 up-regulation in NZB relative to SM. Attractin is part of the agouti pathway that leads to obesity due to aberrant ectopic protein expression (HE et al. 2001), and Atrnl1 may similarly be involved in this pathway.

Mass-spectrometric proteomic analysis: We examined liver protein expression differences between the strains using mass-spectrometry with iTRAQ reagent chemistry labeling technology (ROSS et al. 2004). Samples were used from two of the three biological replicates used for the microarray study. iTRAQ labels are essentially chemical compounds with two parts. The first part, retained by the peptide after scattering, is available at four different masses of 114, 115, 116 or 117 Daltons. The second part of the label is not retained; this part of the molecule is designed to balance the mass of the four labels. Thus all four labels are isobaric, so that observed differences are not due to the differential mass of the labels. Up to four samples can be labeled so that within a sample every peptide is equally tagged with an additional mass of 114, the peptides of the second sample can be tagged with a mass of 115 and so on. We identified 943 proteins with 95% confidence: approximately 7% were ribosomal and 43% represented novel or uncharacterized gene products.

These data are presented as exploratory; the statistical power is not optimal since only a two-by-two comparison is possible on this proteomics platform. Consequently we are using a
statistical threshold of $P<0.1$, derived from permutation analysis, which we are using to indicate peptides that might be significantly differentially expressed under more reproducible experimental conditions such as Western analysis. Rather than reduce the statistical threshold for this technology and eliminate possible candidates, the higher threshold exposes us to Type I errors, which are more acceptable to us than Type II errors; that is, to eliminate candidate genes that are responsible for the QTL. For example APOA2, which has a p-value of 0.077 in our analysis, has been shown previously to be differentially abundant with no mRNA difference (DOOLITTLE et al. 1990) and is responsible for the HDL QTL linked to it (WANG et al. 2004).

We identified 45 proteins as differentially regulated between the two strains ($P<0.1$) (Supplementary Table S13); 33 of these 45 could be linked to corresponding genes. The remaining 12 peptides included 8 peptides that could not be mapped and 4 problem peptides that did not map to a unique location because they were in highly conserved protein families.

**Candidates based on differences in protein abundance:** Only a few of the 33 differentially expressed proteins mapped to a QTL (Figure 2). For the HDL QTL, apolipoprotein A-II precursor (APOA2) ($P=0.077$) and fumarate hydratase (FH1) ($P=0.073$) are encoded by genes that fall within the Hdlq20 QTL interval on Chr 1. Additionally, hydroxysteroid 11-beta dehydrogenase 1 (HSD11B1) maps close to this QTL and knock-out mice for this gene have increased HDL cholesterol levels (MORTON et al. 2001).

The gene that encodes acyl-coenzyme A dehydrogenase, short chain (ACADS, previously known as SCAD) maps within Hdlq1 on Chr 5 (Table 1), and NADH dehydrogenase (ubiquinone) 1 alpha subcomplex 9 (NDUFA9) is encoded by a gene within the Hdlq24 QTL on Chr 6. The gallstone QTL Lith17 on Chr 5 overlaps with the HDL QTL Hdlq1, and thus ACADS is also a candidate for this gallstone QTL.
In order to validate some of the candidates identified from the iTRAQ ABI 4800 Proteomics Analyzer, antibodies suitable for Western analysis were obtained for several proteins. These include FH1 and HSD11B1 on Chr 1, ACADS on Chr 5, and NDUFA9 on Chr 6. Western analysis confirmed the initial results from MS (Figure 3). Significant differences were observed for all four proteins after normalizing to B-TUBULIN; NDUFA9 (p=0.0033), FH1 (p=0.0002), HSD11B1 (p=0.0091) and ACADS (p=0.016).

DISCUSSION

Often QTL are large, containing hundreds of genes, so different tools and approaches are needed to reduce the number of probable candidate genes that require further testing. Combining three approaches, we generated smaller candidate gene lists for each QTL. This was achieved by searching for differential amounts of either mRNA and protein abundance and by mining publicly available SNPs to exclude regions of the QTL that are inferred to be identical by descent and to identify genes with protein coding differences that may indicate functional differences. Each of these three approaches have their limitations and strengths as discussed below.

If these three approaches have not reduced the possible candidate genes to a manageable list, which we consider 10 genes or less, then other approaches are available, some of which have been discussed previously (DiPetrillo et al. 2005; Flint et al. 2005). For example, comparative genomics is a useful approach both for narrowing the QTL if a homologous QTL has been found in other species or examining the genes found in a human genome wide association study in the mouse. Another approach is to examine the expression databases to determine if the candidate gene is expressed in the relevant tissues. Following this bioinformatics
work or global expression studies, we test the small list of candidates by sequencing and by expression analysis. Our goal is to obtain at least three independent lines of evidence for a candidate gene.

**mRNA expression data:** Today microarray technology yields reliable, consistent and reproducible data from probe sets that are well annotated. As with the platform used here (Affymetrix GeneChip® Mouse Genome 430 2.0), most technologies report mRNA expression for many thousands of genes, approaching the entire known transcriptome. It is likely, therefore, that if a differentially expressed gene in the liver is causal for any of the nine QTL, we should be able to detect it. Indeed, when we examine the differentially expressed probes in the QTL intervals located in haplotype regions that are not identical by descent between our two strains, we find several genes that are obvious candidates. These include *Scarb1* for *Hdlq1* on Chr 5, *Ctgf* for *Lith21* on Chr 10, *Slc10a2* for *Lith19* on Chr 8 and *Atrnl1* for *Obwq5* on Chr 19 (Figure 3 and Table 3).

Although searching for differential gene expression using microarrays is probably the most advanced technology of the three methods, it has several limitations. Suitable probes for a gene may not be on the microarray and splice variants are difficult to detect. The differential expression could be caused by differential binding of the mRNA due to SNPs in the probe. This can be guarded against by searching the databases for SNPs in the probes and by confirming any expression difference with new probes and RT-PCR. Trans regulatory elements outside of the QTL may cause differential expression of a gene in the QTL region; similarly, gene interaction or epistasis could cause the parental strains to have similar gene expression even though differential expression was causal for the QTL in an inbred cross. Both errors can be guarded against by measuring gene expression in tissues from selected F2 progeny (comparing mice
homozygous for one parent over the QTL region to mice homozygous for the other parent over the same region).

**Protein analysis:** The primary aim of microarrays is to identify differentially expressed genes that are representing differentially abundant proteins. Large scale proteomic techniques have only recently become feasible; however, they are technically and financially restrictive relative to mRNA arrays. Nonetheless, as proteomic technologies improve they may replace microarray gene expression since it will be possible to directly interrogate the presence of differentially abundant proteins. Despite the ability to directly identify 943 proteins, making direct comparisons between mRNA regulation and protein regulation remains difficult, principally because MS/MS proteomics and the accompanying bioinformatics are still at early stages of development. Areas of improvements that will significantly advance the field should include standardized protocols on tissue processing. In this study, we chose a single tissue processing condition (pH 8.5) that would yield the widest variety of protein peptides from liver; this isolation may miss more basic and more acidic proteins. A more thorough investigation of the liver proteome would include a range of tissue processing conditions.

Another desirable improvement is peptide annotation, specifically linked annotation between protein/peptides and gene databases. Although we identified 943 proteins, a significant proportion of these (43%) could not be assigned to a gene, thus direct comparisons with Affymetrix probe IDs could not be made. Furthermore, a significant proportion of the 943 proteins, despite a protein identifier, had little or no annotation in public databases and indeed BLAST searches of the peptides to genome databases (*Ensembl* _mouse_) derived no hits at all, even after decoding the peptide into DNA sequence. Lastly, standardized analytical methods of
MS/MS data are currently lacking and an initiative similar to that seen for mRNA gene analysis is required.

Despite the need for continued technological advancements in the proteomic field, the technology is developed enough to identify changes in protein abundance between samples. ACADS, a protein encoded by a gene in both the HDL QTL Hdlq1 and the gallstone QTL Lith17 on Chr 5, showed significantly increased protein abundance in SM adult female liver relative to the NZB strain (2.6 and 1.6 fold change from MS and Western analysis respectively), but no difference in mRNA expression (Table 2). The difference in protein abundance was confirmed by Western blot analysis. Similar results were obtained for FH1, HSD11B1 and NDUFA9 (Table 3, Figures 2, 3). Finally, our MS data indicates a 4 fold protein increase of APOA2 in NZB relative to SM protein abundance with no mRNA expression difference. A difference in the quantity of APOA2 protein but not in mRNA was reported previously between two different strains (DOOLITTLE et al. 1990). We have previously shown that the gene Apoa2 is responsible for the HDL QTL Hdlq5 (WANG et al. 2004) in mice and associated with HDL levels in human studies (LILJA et al. 2002).

The proteomics analysis reveals two important aspects of this study. First, differentially abundant proteins can be identified and may underlie QTL, and second, as others have shown in yeast (GYGI et al. 1999), these differentially expressed proteins are not necessarily accompanied by differences at the mRNA level. Of the 33 named differentially expressed proteins identified with corresponding Affymetrix probe sets, 12 had differentially expressed mRNA in the same direction with an FDR of q<0.05 (39%) (Supplementary Table S13). We expect that differentially abundant proteins whose mRNA levels are similar will have changes at the gene
(e.g. 5’ and 3’ UTRs) or protein level (e.g. phosphorylation) that will affect protein synthesis and degradation. APOA2 is known to have such an amino acid change (WANG et al. 2004).

**SNPs and haplotype data:** We used public SNP information to first eliminate those regions that were in the 95% confidence interval of the QTL but were identical by descent between NZB and SM. We then examined the remaining genes by identifying codon changing SNPs. Once genes with non-synonymous SNPs were identified between well characterized strains, they were compared to the haplotypes of SM and NZB, which allowed us to infer if the same non-synonymous polymorphism might be present between SM and NZB as illustrated in Table 2. This generated a list of genes that may have non-synonymous changes between the two strains, which may affect the function of the proteins thus causing the QTL. However the number of genes identified using this method for some QTL was still prohibitively large for testing (e.g. 88 genes in the *Obwq4* QTL). In addition, it has recently been shown that SNPs causing synonymous codon changes, for which the alternate codon is rare, can lead to a difference in protein efficiency or function (KIMCHI-SARFATY et al. 2007).

The method of using SNPs does have several drawbacks that will become less important in the future as SNP databases improve (REUVENI et al. 2007). Using the SNP database to eliminate regions that are identical by descent could lead to error by incorrectly eliminating regions; this type of error will be reduced as the density of the SNPs improves. The SNP database, even for those strains sequenced by Celera and Perlegen, is incomplete because both groups used rigorous criteria to guard against false positives. Based on the Perlegen data, about 50% of the SNPs will be identified if the minor allele is shared by three strains (58% if the minor allele frequency is shared by seven strains) (YANG and CHURCHILL 2007) or somewhat better if Celera data is included or only classic inbreds (not wild-derived strains) are considered.
Summary: In spite of the limitations for each method, we were reasonably effective in reducing the number of candidate genes in each QTL region. Some genes had multiple lines of evidence such as *Acads* for *Hdlq1* on Chr 5 and *Adipor2* for *Obwq3* on Chr 6 (Table 3 and S2) but multiple lines of evidence do not prove that it is the causal gene. In one case we were able to identify a known gene for a QTL (*Apoa2*); and each method revealed likely candidates that the other methods did not. We conclude that QTL candidate genes should be assessed for changes in coding and both protein and message levels.
Acknowledgements

This work was supported by The American Heart Association grant 0525816T (to IMS) National Institute of Health grants HL77796 and HL66611 (to B.P), and GM070683 (to G.C). Special thanks to Stephen Grubb, Allison Cox and Sikander Hayat for assistance with bioinformatics; Douglas Hinerfeld and Gene Expression and Protein Chemistry Services at The Jackson Laboratory; to Dr. Vockley’s Lab from the Children’s Hospital of Pittsburg, for the ACADS antibody; to Applied Biosystems for the kind donation of the TaqMan assays; and to Sarah Langley for assistance in proofreading.
LITERATURE CITED


IRIZARRY, R. A., B. HOBBS, F. COLLIN, Y. D. BEAZER-BARCLAY, K. J. ANTONELLIS et al., 2003b
Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.


Lyons, M. A., R. Korstanje, R. Li, S. M. Sheehan, K. A. Walsh et al., 2005 Single and interacting QTLs for cholesterol gallstones revealed in an intercross between mouse strains NZB and SM. Mamm Genome 16: 152-163.


<table>
<thead>
<tr>
<th>Trait</th>
<th>QTL</th>
<th>Chr</th>
<th>LOD</th>
<th>Nearest marker</th>
<th>High allele</th>
<th>Mode of inheritance</th>
<th>Peak (CI start-end)</th>
<th>Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL</td>
<td>Hdlq20</td>
<td>1</td>
<td>10.6</td>
<td>D1Mit291</td>
<td>NZB</td>
<td>Dom</td>
<td>174 (168-184)</td>
<td></td>
</tr>
<tr>
<td>HDL</td>
<td>Hdlq1</td>
<td>5</td>
<td>10.6</td>
<td>D5Mit161</td>
<td>NZB</td>
<td>Add</td>
<td>122 (116-132)</td>
<td></td>
</tr>
<tr>
<td>HDL</td>
<td>Hdlq24</td>
<td>6</td>
<td>6.4</td>
<td>D6Mit259</td>
<td>NZB</td>
<td>Add</td>
<td>128 (115-144)</td>
<td></td>
</tr>
<tr>
<td>Gallstone presence</td>
<td>Lith17</td>
<td>5</td>
<td>7.1</td>
<td>D5Mit24</td>
<td>SM</td>
<td>Add</td>
<td>102 (88-122)</td>
<td></td>
</tr>
<tr>
<td>Gallstone presence</td>
<td>Lith19</td>
<td>8</td>
<td>5.6</td>
<td>D8Mit155</td>
<td>SM</td>
<td>Rec</td>
<td>6 (0-16)</td>
<td></td>
</tr>
<tr>
<td>Gallstone presence</td>
<td>Lith21</td>
<td>10</td>
<td>5.7</td>
<td>D10Mit214</td>
<td>NZB</td>
<td>Rec</td>
<td>48 (18-58)</td>
<td></td>
</tr>
<tr>
<td>QTL for obesity and body weight</td>
<td>Obwq4</td>
<td>17</td>
<td>6.8</td>
<td>D17Mit20</td>
<td>NZB</td>
<td>Add</td>
<td>54 (34-62)</td>
<td></td>
</tr>
<tr>
<td>QTL for obesity and body weight</td>
<td>Obwq5</td>
<td>19</td>
<td>9.3</td>
<td>D19Mit71</td>
<td>NZB</td>
<td>Rec</td>
<td>56 (48-60)</td>
<td></td>
</tr>
<tr>
<td>QTL for obesity and body weight</td>
<td>Obwq3</td>
<td>6</td>
<td>5.8</td>
<td>D6Mit105</td>
<td>SM</td>
<td>Rec</td>
<td>100 (88-122)</td>
<td></td>
</tr>
</tbody>
</table>

LOD scores are given with sex as an additive covariate except for the body weight and obesity phenotypes. Mb indicates millions of base pairs from the centromere (Build 36). Dom – dominant; Add – additive; Rec – recessive; CI – 95% confidence interval.
### Table 2: Inferring codon changing SNPs between SM and NZB by examining SNPs between highly sequenced strains

Highly sequenced strains eg C57BL/6, DBA/2 etc…

<table>
<thead>
<tr>
<th>Chr6</th>
<th>Gene</th>
<th>i</th>
<th>iv</th>
<th>v</th>
<th>ii</th>
<th>iii</th>
<th>SM</th>
<th>NZB</th>
<th>SM and NZB are:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C57BL/6</strong></td>
<td>Syn2</td>
<td></td>
<td></td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>IBD</td>
</tr>
<tr>
<td><strong>DBA/2</strong></td>
<td>Syn2</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Syn2</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>non-IBD</td>
</tr>
<tr>
<td></td>
<td>Syn2</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td><strong>C57BL/6</strong></td>
<td>Pparg</td>
<td></td>
<td></td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>non-IBD</td>
</tr>
<tr>
<td><strong>DBA/2</strong></td>
<td>Pparg</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pparg</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pparg</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>non-IBD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td><strong>C57BL/6</strong></td>
<td>Tsen2</td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>non-IBD</td>
</tr>
<tr>
<td><strong>DBA/2</strong></td>
<td>Tsen2</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tsen2</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tsen2</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tsen2</td>
<td>C</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>IBD</td>
</tr>
</tbody>
</table>
Cs – coding synonymous; Cn – coding non-synonymous; IBD – identical by descent. Although SM and NZB are not as well sequenced as other strains (missing SNP data indicated by the white space), Cns between SM and NZB can be inferred by examining the haplotypes of the other strains. Here, for the Cn in Ppar indicated in bold, SM is inferred to be ‘C’ and NZB is inferred to be ‘T’.
Table 3: Candidate genes for each QTL based from our data and published evidence – the genes to test first.

<table>
<thead>
<tr>
<th>Trait</th>
<th>QTL</th>
<th>Gene</th>
<th>Chr</th>
<th>Mb</th>
<th>mRNA</th>
<th>Protein</th>
<th>Sequence</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDL</td>
<td>Hdlq20</td>
<td>Apoa2</td>
<td>1</td>
<td>173.1</td>
<td>4.1</td>
<td>X</td>
<td></td>
<td>(LILJA et al. 2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fh1</td>
<td></td>
<td>177.5</td>
<td>-2.8</td>
<td></td>
<td>X</td>
<td>(MORTON et al. 2001)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hsd11b1*</td>
<td></td>
<td>194.9</td>
<td>-3.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hdlq1</td>
<td>Acads</td>
<td>5</td>
<td>115.3</td>
<td>-2.6</td>
<td></td>
<td>X</td>
<td>(ACTON et al. 1996)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scarb1</td>
<td></td>
<td>125.6</td>
<td>-2.1</td>
<td></td>
<td></td>
<td>(CHINETTI et al. 2001)</td>
</tr>
<tr>
<td></td>
<td>Hdlq24</td>
<td>Pparg</td>
<td>6</td>
<td>115.4</td>
<td></td>
<td>X</td>
<td></td>
<td>(YAMAUCHI et al. 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rassf4</td>
<td></td>
<td>116.6</td>
<td>-8.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lith17</td>
<td>Scarb2</td>
<td>5</td>
<td>93.6</td>
<td></td>
<td>X</td>
<td></td>
<td>(ACTON et al. 1996)</td>
</tr>
<tr>
<td>Gallstone</td>
<td></td>
<td>Acads</td>
<td></td>
<td>115.3</td>
<td>-2.6</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>presence</td>
<td>Lith19</td>
<td>Slc10a2</td>
<td>8</td>
<td>5.1</td>
<td>5.2</td>
<td></td>
<td></td>
<td>(SHIH et al. 2001)</td>
</tr>
<tr>
<td></td>
<td>Lith21</td>
<td>Ctgf</td>
<td>10</td>
<td>24.3</td>
<td>-2.6</td>
<td></td>
<td>X</td>
<td>(KONINGER et al. 2005)</td>
</tr>
<tr>
<td>Obesity</td>
<td>Obwq3</td>
<td>Pparg</td>
<td>6</td>
<td>115.4</td>
<td></td>
<td>X</td>
<td></td>
<td>(COCK et al. 2004)</td>
</tr>
<tr>
<td>Genes</td>
<td>fold change</td>
<td>wt fold</td>
<td>protein fold</td>
<td>ref.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>---------</td>
<td>--------------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rassf4</td>
<td>116.6</td>
<td>-8.2</td>
<td></td>
<td>(YAMAUCHI et al. 2003)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adipor2</td>
<td>119.3</td>
<td>1.9</td>
<td>X</td>
<td>(PUIGSERVER et al. 1998)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obwq4</td>
<td>Pgc</td>
<td>17</td>
<td>47.2</td>
<td>X</td>
<td>(GUNN et al. 1999)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obwq5</td>
<td>Atrnl1</td>
<td>19</td>
<td>57.7</td>
<td>3.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Genes listed are in QTL regions that are not identical by descent between SM and NZB (*Hsd11b1* maps close to *Hldq20*). Protein and mRNA fold changes are given where a positive value indicates up-regulation and a negative indicates down-regulation in NZB relative SM.

1 Fold changes for mRNA are significant (Fs<1x10^-4); 2 Proteins are significantly different as determined by Western analysis (Figure 3), except for APOA2 which is suggestive (P<0.1) based from MS/MS data; 3 Genes that contain known or extrapolated codon changing SNPs leading to potential functional differences are indicated by ‘x’; 4 Gene names from table S2 were systematically interrogated against the PubMed database using relevant keywords (STYLIANOU et al. 2006).
Table 4: Summary of candidate genes identified by the various methods used

<table>
<thead>
<tr>
<th>Method</th>
<th>Significance level</th>
<th>Condition</th>
<th>No. of genes/proteins</th>
<th>Total No. of genes/proteins*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray</td>
<td>FDR 0.05</td>
<td>over all chromosomes</td>
<td>11223</td>
<td>24530 (^a)</td>
</tr>
<tr>
<td>Microarray</td>
<td>FDR 0.0001</td>
<td>over all chromosomes</td>
<td>831</td>
<td>24530 (^a)</td>
</tr>
<tr>
<td>Microarray</td>
<td>FDR 0.0001</td>
<td>in QTL</td>
<td>87</td>
<td>3763 (^b)</td>
</tr>
<tr>
<td>Proteomics</td>
<td>P&lt;0.1</td>
<td>over all chromosomes</td>
<td>45</td>
<td>943</td>
</tr>
<tr>
<td>Proteomics</td>
<td>P&lt;0.1</td>
<td>in QTL region</td>
<td>5</td>
<td>unknown (^c)</td>
</tr>
<tr>
<td>SNP analysis</td>
<td>N/A</td>
<td>in QTL and non-IBD</td>
<td>336</td>
<td>2752</td>
</tr>
</tbody>
</table>

* Gene here also refers to an Affymetrix probe for the microarrays, and a peptide linked to a protein entry in the Swiss-Prot or TrEMBL databases for the proteomics. \(^a\) There are 45,037 non-control probe sets on the arrays; however 20,507 are not expressed in any of the six samples indicating genes for those probes are not expressed in adult female mouse liver. \(^b\) The total number of probes within the total space of the 9 QTL including IBD regions between SM and NZB. \(^c\) Many protein hits could not be mapped to genomic locations so the genes located only in QTL regions cannot be determined.
<table>
<thead>
<tr>
<th>Gene</th>
<th>SM</th>
<th>NZB</th>
<th>p-Value*</th>
<th>FC'</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-TUBULIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDUFA9</td>
<td></td>
<td></td>
<td>0.0002</td>
<td>-1.9</td>
</tr>
<tr>
<td>FH1</td>
<td></td>
<td></td>
<td>0.0033</td>
<td>-1.2</td>
</tr>
<tr>
<td>HSD11B1γ</td>
<td></td>
<td></td>
<td>0.0091</td>
<td>-1.4</td>
</tr>
<tr>
<td>ACADS‡</td>
<td></td>
<td></td>
<td>0.0162</td>
<td>-1.6</td>
</tr>
</tbody>
</table>