HALT-C Ancillary Study PROPOSAL

PART I

Proposal Name: Application of Novel Organ-Specific Serum Protein Biomarker Candidates to Diagnose Hepatic Fibrosis in Hepatitis C Patients in HALT-C Trial

Proposal PI: Leroy Hood

Co-Investigators: David Galas, Institute for Systems Biology, Seattle; Kai Wang, Institute for Systems Biology, Seattle, WA; Gil Omenn and Alex Tsodikov, University of Michigan

HALT-C PI: Anna Lok, University of Michigan

Funding Agency and Review Body (e.g., NIDDK; my university/GAC): Luxembourg Ministries of Health, Research, and Economy and External Trade. Funding announced in July 2008.

I agree to follow HALT-C Policies and Procedures when conducting this study. I acknowledge that the data obtained from this study will belong to the NIH and will be placed in the HALT-C database for use by other investigators. I understand that I cannot begin experiments using HALT-C specimens/data until I receive approval from the HALT-C Ancillary Studies Committee and funding from the Scientific Review Body for my proposal. I also understand that the data analysis for this proposal will be performed by NERI (unless otherwise approved by the HALT-C study) and that Protocols approved by the HALT-C Ancillary Studies Committee will be placed on the HALT-C Restricted Website.

________________________________	____________________
Proposal Principal Investigator	Date
________________________________	____________________
HALT-C Principal Investigator	Date

PART II: PROTOCOL

1. Hypothesis: Organ-specific proteins can be detected in the circulation and can be used to monitor disease-specific changes in the primary organ affected, namely liver for hepatic fibrosis and its progression.

Background: The ISB investigators have utilized findings from highly sensitive transcriptomic analyses with massively parallel signature sequencing and next-generation high throughput sequencing and from proteomics analyses (multiple reaction monitoring (MRM); unique N-glycopeptides) to identify liver-specific protein biomarker candidates detectable in plasma.

Specific Aim 1, Pilot Phase: Use panels of liver-specific protein biomarker candidates in high-throughput format with HALT-C specimens in a preliminary study to determine whether these markers can distinguish patients with cirrhosis from those with fibrosis. Liver biopsy specimens and serum specimens collected at the same time from such patients will be tested to relate protein changes in liver to homologous protein changes in serum.

Specific Aim 2: Validate serum biomarker findings from Aim 1 in a larger number of HALT-C specimens (independent of those used in Aim 1) to evaluate usefulness of this panel in identifying patients with cirrhosis and in predicting which patients will have disease progression.

2. Background/Rationale: Systems Biology with Blood Protein Fingerprints as Disease Diagnostics

2.1 Blood is a window into the status of the body's health. Organs throughout the body are highly dependent upon a continuous supply of energy through the circulatory system and for removal of molecules and cells released in health and disease. Blood is an informative specimen window to assess the functional status of organs, detect biological responses of organs to injuries, and monitor specific markers of developing pathology. In this project, we will explore and validate a panel of blood- and tissue-based liver-specific markers, including both RNA and proteins during the progression of HCV-related diseases.

2.2 **Organ-specific blood molecular fingerprints.** Uncertainty about the tissue of origin of protein biomarker candidates causes diagnostic non-specificity. The solution is to identify and utilize organ-specific biomarkers (majority of the message produced normally by only one organ). To identify organ-specific markers, we have used two different, complementary datasets: (a) a deep (millions of transcripts for each organ) transcript database generated at the Institute for Systems Biology using a technique called massively parallel signature sequencing (MPSS), and (b) data from a microarray platform containing over 42,000 different cDNA clones, from the National Cancer Institute (NCI). The MPSS dataset contains transcripts from 37 distinct human tissues, including replicate data for 7 of these tissues and 13 from perturbations of the normal state of these tissues [Hood et al, 2004]. The NCI microarray dataset contains expression results of 19 different human organs with 158 normal human samples obtained from 30 different individuals [Son et al, 2005]. Since the analyzed data are not available from NCI, we downloaded the entire dataset, reanalyzed with specific selection criteria (e.g., that 50% of the expression signals come from particular tissues), and extracted tissue "specific" genes (see Figure 1).

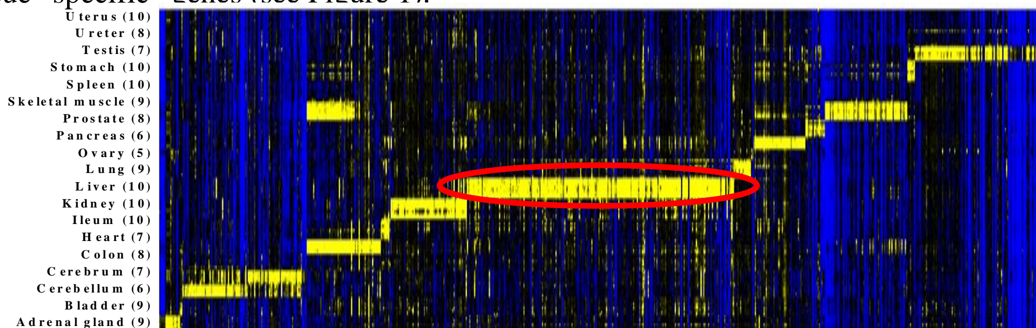

Figure 1: Clustering of 730 organ specific genes identified from NCI normal human tissue profiling data based on gene expression levels. The individual genes are arranged on the X-axis, while 158 different samples grouped into tissues are on the Y-axis. Individual tissue groups are listed on the left followed by number of samples in each tissue group. The expression levels from highest to lowest are represented from yellow to blue colors. A yellow "line" across the image from left to right indicates a group of tissue specific genes; the "line" circled in red are liver specific genes.

2.3 **Proof-of-principle: The ISB systems approach can identify early stage biomarkers for complex neurological diseases.** Hood et al [Mol Syst Biol 2009 in press] studied in brain mRNA the dynamic onset of infectious prion disease in several strains of mice, and showed that interlocking protein networks surrounding the prion protein are significantly perturbed across the 150-day span from disease initiation to death. Figure 2 shows a small subset of the subtractive networks that were derived from comparisons of mRNA-expression patterns in normal and diseased brains at each of three time points after infection. The use of time-dependent whole transcriptome data from multiple inbred mouse strains enabled signal to be extracted from a very noisy environment. In addition, the temporal analysis reveals the alteration of network transcription level changes prior to symptom onset. Thus, biomarkers that are specific to the changing nodes, or to the biological processes they regulate, can be *predictive and useful as diagnostic biomarkers, even before symptoms arise*. If some of these altered nodes encode secreted proteins, they could become readily accessible in the blood as markers for early disease detection.

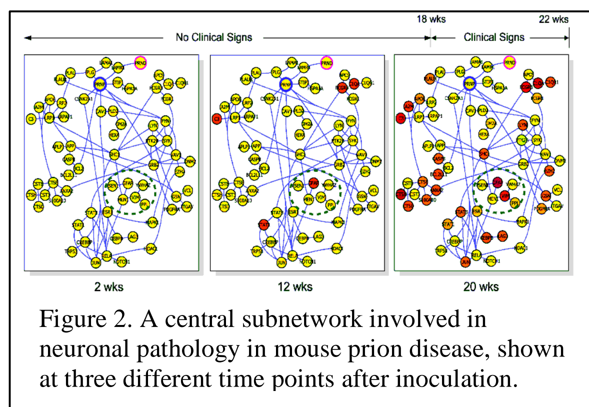

Figure 2. A central subnetwork involved in neuronal pathology in mouse prion disease, shown at three different time points after inoculation.

2.4 **The development of MRM peptides for liver-specific proteins**. From liver-specific transcripts, we selected 36 proteins as highly specific for liver and known to be detectable in plasma, since they were reported already in the Human Proteome Organization (HUPO) international Plasma Proteome Project led by Gil Omenn, Young-Ki Paik, and Ruedi Aebersold and in the PeptideAtlas database (www.peptideatlas.org, Eric Deutsch, ISB). Based on this list of proteins (Table 1), we have designed peptide standards for MRM assays, at least two per protein for most of the proteins. These peptides are unique or highly associated with the particular protein, and are designed to be detectable in the mass spectrometer and quantifiable [Anderson & Hunter, 2006].These prepared peptides have been extensively tested on Thermo Finnigan LCQ deca instruments with full LC-MS/MS and SRM (selected reaction monitoring) runs. The current list of MRM peptides for liver-specific proteins is included below (Table 1, at end).

Based on the results, MRM transition candidates from each parent peptide standard ion were selected and tested for MRM analysis. We have developed MRM methods to scout several precursor-fragment ion transition pairs from each peptide over a large time period in order to observe elution for each peptide on a Waters nanoAcquity UPLC system coupled to a Waters Quattro Premier XE triple quadrupole mass spectrometer or an Eksigent Tempo nano-LC system coupled to an ABI Q-Trap 4000 mass spectrometer. These observed retention times will be mapped to a time-segmented MRM method capable of analyzing hundreds of peptide precursor-fragment ion transition pairs in a single chromatographic run. Results for peptide standards were obtained that show capturing of transition daughter ions with reasonably strong signals. We create a liver-specific fingerprint by comparing the isotopically-labeled (heavy) peptide standards to endogenous (light) blood protein in mouse or human.

We will interrogate protein level changes with disease phenotypes in samples spiked with isotopically labeled peptide standards. We will also search the transcript, proteome, and peptide databases for known liver markers: AFP (alpha-fetoprotein), N-fucosylated AFP-L3, DCP (des-gamma carboxyprothrombin), Glypican 3, GP73 and N-fucosylated GP73, GEP (granulin-epithelin precursor), and mortalin [Block et al, 2008]. We will compare with standard liver function assays for SGOT, SGPT, and GGT transaminase activity. In all of these cases, we will explore the feasibility of MRM peptide standards.

3. Relation to Aims of HALT-C Study

This ancillary study is closely related to the aims of HALT-C as the goal is to develop and evaluate non-invasive tests for diagnosis of fibrosis and progression of liver disease. If the organ-specific markers, defined

with both mRNA and protein studies of the liver, can be linked with organ-specific proteins in the circulation, we may be able to diagnose cirrhosis and progressive liver disease non-invasively.

4. Study design, experimental groups

4.1 **Specific Aim 1: Use panels of liver-specific protein biomarker candidates in high-throughput format with HALT-C specimens in a preliminary study to determine whether these markers can distinguish hepatitis C patients with cirrhosis versus fibrosis**. Liver biopsy specimens and serum specimens collected as close as feasible to the same time from such patients will be tested to relate protein changes in liver to homologous protein changes in serum.

We request baseline liver biopsy specimens and serum specimens within 90 days from 20 patients with Ishak 3 or 4 (bridging fibrosis) who did not have histological or clinical outcome during the course of the study and 20 patients with cirrhosis who developed clinical decompensation during the study. We will first document what protein changes in liver can be related to and account for homologous protein changes in serum among the 36 proteins in Table 1 and any others added with the same criteria of liver-specific transcripts and detectability and quantifiability in serum.

We will compare expression levels of the organ-specific proteins in the two groups of patients, using a cross-sectional design. Longitudinal progression within individual patients will be left for Specific Aim 2. Given the complications of sampling error for each phenotype, the preferred design is cross-sectional, informed by knowledge that the patients with fibrosis did not progress to cirrhosis on follow-up biopsies, allowing clear separation of the groups.

Sample Size. For this pilot study, we plan the sample size to detect a relative difference of 40% in protein levels between any two groups of patients in the pairwise comparisons between fibrosis and cirrhosis. Assuming a coefficient of variation of <0.5, we would have at least 80% power to detect the difference with 20 patients per group.

Statistical Analysis Plan. Multiple linear regression will be used to assess the effects of candidate liver-specific proteins associated with the list of MRM peptides. Transformations will be considered should model diagnostics plots reveal any evidence of departure from linearity. In a secondary analysis, we will reverse the problem formulation with the disease category (fibrosis, cirrhosis, HCC) being the response while protein expression levels are the explanatory variables. Due to small sample size, principal component analysis on protein expression will be used to reduce the dimension of the feature vector with the first two/three principal components entering a multinomial logistic regression model. Since data are sampled based on the disease status, no prediction of absolute probabilities of the type of disease based on the protein level will be attempted. The focus will be on identifying significant relative effects potentially allowing us to discriminate between the pairs of disease categories.

4.2 **Specific Aim 2, Validation Study:** Validate serum biomarker findings from Aim 1 in a larger number of HALT-C specimens (independent of those used in Aim 1) to evaluate usefulness of this panel in the diagnosis of hepatitis C-related cirrhosis and in the prediction of progression of liver disease.

We will proceed with aim 2 only if results of aim 1 are positive. Based on the results of Aim 1, a panel of serum biomarkers that differentiates patients with Ishak 3 or 4 and no histological or clinical outcome and patients with Ishak 5 or 6 who developed clinical outcome will be assembled. Results will be considered positive if this panel offers an accuracy $\geq$70% in differentiating these two groups of patients.

We request serum specimens from a total of approximately 150 patients: 30-40 patients with baseline biopsy showing Ishak 3 or 4 and no progression to histological cirrhosis, clinical decompensation, or HCC; 30-40 patients with baseline biopsy showing Ishak 3 or 4 and progression to histologic cirrhosis but no clinical decompensation or HCC; 30-40 patients with baseline biopsy showing Ishak 5 or 6 and no

progression to clinical decompensation or HCC, and 30-40 with baseline Ishak 5 or 6 and progression to clinical decompensation but not HCC. For those who progressed, we would like to obtain 2 earlier samples at baseline and approximate midpoint to time of cirrhosis or clinical decompensation. It would be desirable to match groups for gender, ethnicity, and HCV genotype.

Statistical Analysis Plan: As for Aim 1, we will use multiple linear regression with the disease status as a categorical covariate. Adjusted R-Squared, Mallow's Cp, and AIC criteria will be used to ensure reproducibility of the predictive properties of the protein panel in future observations. In order to handle multiple observations made on a subset of study subjects, a random intercept term will be introduced in the linear mixed model. As a secondary analysis, survival analysis methods (proportional and non-proportional hazards regression models) will be used on a subset of subjects with follow-up information on their progression from fibrosis to cirrhosis and/or HCC during the follow-up. There is insufficient information at this point to assess the power and precision of the validation analysis. Therefore, updated sample size assessment will be performed after Aim 1 has been completed, yielding quantitative data on coefficient of variation and on group differences by protein. Tentatively, using the same assumptions as in Aim 1, we expect to have improved sensitivity with 30 patients per group, instead of 20, resulting in an ability to detect a 27% difference in protein levels. We will analyze both individual proteins and all proteins, as well as potentially related, clustered proteins as secondary analyses.

5. Methods & Data Usage

a) Novel technology platforms: Next generation deep sequencing for transcripts (such as Solexa); N-glycopeptide proteomics and multiple reaction monitoring for proteins by mass spectrometry.

b) Protein biomarker panel: See Table 1 for proteins and MRM peptides to be assayed.

c) Assays of liver and serum specimens: We will use multiple reaction monitoring of the 2 or 3 specific peptides associated with each of the 36 proteins in the Table below in an Orbitrap or LTQ-FT mass spectrometer; the peptides have already been designed (Table 1).

d) Statistical design and analysis: See specific sections under Aim 1 and Aim 2.

e) Data usage: We will fully comply with the HALT-C Data Use Agreement. Samples will be labeled with HALT-C subject no. and sample ID. We have no interest in specific-patient identifiers. We will use newly-generated identifiers held by the HALT-C PI to permit longitudinal analyses of serum specimens from the patients studied. We will report all findings on a group basis.

To address Aim 1, a defined data set on patients whose samples are studied will be created by NERI and sent to University of Michigan (or Institute of Systems Biology) for analysis. The dataset will include:

- Demographics, duration of HCV infection, alcohol and smoking history, BMI, HCV genotype
- Baseline information: histology (date of biopsy, inflammation, fibrosis and steatosis score, length of biopsy and fragmentation), baseline labs (date of labs, CBC, liver panel, Cr, INR, AFP, HCV RNA), baseline ultrasound (date of ultrasound or other imaging, spleen size), baseline EGD (date of EGD, presence or absence of esophageal varices / portal gastropathy)
- Month 24 information: histology (date of biopsy, inflammation, fibrosis and steatosis score, length of biopsy and fragmentation), labs (date of labs, CBC, liver panel, Cr, INR, AFP, HCV RNA), ultrasound (date of ultrasound or other imaging, spleen size)
- Month 48 information: histology (date of biopsy, inflammation, fibrosis and steatosis score, length of biopsy and fragmentation), labs (date of labs, CBC, liver panel, Cr, INR, AFP, HCV RNA), ultrasound (date of ultrasound or other imaging, spleen size), EGD (date of EGD, presence or absence of esophageal varices / portal gastropathy)
- Information at last follow-up: labs (date of labs, CBC, liver panel, Cr, INR, AFP, HCV RNA), ultrasound (date of ultrasound or other imaging, spleen size)
- Treatment assignment: peginterferon vs. control, for peginterferon group – date of last dose
- Occurrence of clinical outcomes: date when each outcome is reached

6. Anticipated Results

a) We are confident that the new technology platforms and powerful transcriptomic and proteomic assays will be shown to identify organ-specific proteins for liver detectable in serum.

b) We expect to generate novel protein biomarkers that differentiate progressors from non-progressors among hepatitis C patients in HALT-C trials.

c) We will test in Aim 2 whether serum levels of these protein biomarkers can differentiate and predict progression to cirrhosis and clinical decompensation. We hope to provide new biomarker assays for the non-invasive assessment of liver fibrosis and disease progression to HALT-C investigators.

7. Statistical Support

Dr. Omenn has obtained consulting statistical guidance from Dr. Alex Tsodikov, Professor of Biostatistics at the University of Michigan and Affiliate Faculty in the Center for Computational Medicine and Biology.

The statistical analysis plans are provided in Aim 1 and Aim 2.

All findings will be shared with the HALT-C DCC, and all presentations and manuscripts will be submitted to the DCC for critical comment and clearance.

8. HALT-C Samples (see Part III)

9. Financial Issues

Funding for this ancillary study will come from non-U.S. government sources of support to the Institute for Systems Biology, including a large new 10-year project that includes development of organ-specific molecular diagnostic assays supported by the Ministries of Health, Research, and Economy, External Trade, and Economic Development of Luxembourg.

We will budget funds to cover the costs of obtaining the samples from the HALT-C Repository and of NERI for data analysis.

At the present, there is no industry involvement in this project. In the future, a molecular diagnostics company may be founded or become a partner for certain aspects of this work, especially if successful. The Institute for Systems Biology and the University of Michigan have standard policies for intellectual property rights and licensing.

10. References

Anderson NL, Hunter CL. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 2006;5:573-588.

Block TM, Marrero J, Gish RG, Sherman M, London WT, Srivastava S, Wagner PD. The degree of readiness of selected biomarkers for the early detection of hepatocellular carcinoma: notes from a recent workshop. Cancer Biomarkers 2008; 4:19-33.

Hood, L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science 2004;306:640-643.

Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N, Khan J. Database of mRNA gene expression profiles of multiple human organs. Genome Research 2005;15;443-450.

PART III: Sample Requirements

Patients - Lead-in and Express patients who are randomized are eligible, as long as we have a blood sample within 90 days of the baseline biopsy

For blood samples, serum 300 uL for all patients at all time points.

Aim 1

Visit	Liver # patients, mm	Blood # patients, ml	Other
Baseline	Flash frozen liver tissue, each ≥3 mm, from up to 40 randomized patients a) 20 with Ishak 3 or 4 and no histological or clinical outcome during the study b) 20 with Ishak 5 or 6 who developed clinical decompensation during the study	Serum 300 uL on all patients Blood samples from same patients as liver samples, collected within 90 days of biopsy	

Aim 2 – We will proceed with aim 2 only if results of aim 1 are positive and after HALT-C ancillary studies committee have reviewed those results. Based on the results of Aim 1, a panel of serum biomarkers that differentiates patients with Ishak 3 or 4 and no histological or clinical outcome and patients with Ishak 5 or 6 who developed clinical outcome will be assembled. Results will be considered positive if this panel offers an accuracy ≥70% in differentiating these two groups of patients.

Visit	Liver # patients, mm*	Blood # patients, ml	Other (describe) # pts, amount
Baseline		Serum 300 uL from 4 groups of randomized patients (30-40 patients in each group) a) 30-40 patients baseline Ishak 3 or 4 – no progression to cirrhosis on M48 biopsy b) 30-40 patients baseline Ishak 3 or 4 with progression to cirrhosis on M48 biopsy c) 30-40 patients with baseline Ishak 5 or 6 with no decompensation d) 30-40 patients with baseline Ishak 5 or 6 with decompensation	See below table for visual presentation of the 4 patient groups
Month 24 or mid-point between baseline and Outcome		Serum 300 uL from 4 groups of 30-40 patients listed above – at M24 for groups (a-c) and at the half way point between baseline and outcome for patients in group (d)	
Outcome or Month 48		Serum 300 uL from 4 groups of 30-40 patients listed above – at M48 for groups (a-c) and at the time point closest to meeting outcome for patients in group (d)	

Aim 2	Initial Stage	Disease Progression Up to Month 48

Number of Samples	Amount	Sample Type	Initial Ishak Score	**Cirrhosis**	**Decompensation**	**HCC**
30-40	300 µl	Serum	3 to 4	No	No	No
30-40 (with serial samples if available)	300 µl	Serum	3 to 4	Yes	No	No
30-40	300 µl	Serum	5 to 6		No	No
30-40 (with serial samples if available)	300 µl	Serum	5 to 6		Yes	No

Table 1

List of MRM peptides for liver-specific protines

Gene name	Sequence ID	Number of MRM peptide synthesized	Peptide Sequence
Fumarylacetoacetate hydrolase	NP_000128.1	2	AHEHIFGMVLMNDWSAR
			LLDMELEMAFFVGPGNR
dihydropyrimidinase	NP_001376	2	IPNGVNGVEDR
			MFMAYK
glycine N-methyltransferase	NP_061833.1	3	AWLLGLLR
			VWQLYIGDTR
			SLGVAAEGLPDQYADGEAAR
argininosuccinate synthetase 1	NP_446464.1	2	EQGYDVIAYLANIGQK
			GQVYILGR
4-hydroxyphenylpyruvate dioxygenase	NP_002141.1	3	FLHFHSVTFWVGNAK
			GYLLQIFTK
			TEDIITAIR
argininosuccinate lyase	NP_000039.2	3	WSHWILSHAVALTR
			FVGAVDPIMEK
			VAEEWAQGTF
aldolase B, fructose-bisphosphate	NM_000035.2	3	GIVVGIK
			ALNDHHVYLEGTLLK
			DGVDFGK
betaine-homocysteine methyltransferase	NP_001704.2	2	VNEAACDIAR
			AIAEELAPER
homogentisate 1,2-dioxygenase	NP_000178.2	2	LLIYTEFGK
			CFYNSDGDFLIVPQK
beta-carotene 15, 15'-monooxygenase 1	NP_059125.2	2	FAVPLHVDK
			YVFATGVQWSPIPTK
glutamate dehydrogenase 1	NP_005262.1	2	YSTDVSVDEVK
			MVEGFFDR
catalase	NP_001743.1	2	HMNGYGSHTFK
			FYTEDGNWDLVGNNTPIFFIR
retinol-binding protein 4	NP_006735.2	2	YWGVASFLQK
			FSGTWYAMAK
fibrinogen-like 1	NP_004458.3	3	SDGSENFNR
			FSTWDR
			AQVRLLETR
fructose-1,6-bisphosphatase 1	NP_000498.2	2	TLVYGGIFLYPANK
			DFDPAVTEYIQR
methionine adenosyltransferase I, alpha	NP_000420.1	3	FVIGGPQGDAGVTGR
			YLDEDTVYHLQPSGR
			FVIGGPQGDAGLTGR
hydroxyacid oxidase 1	NP_060015.1	2	AIFVTVDTPYLGNR
			LAMALSGCQNVK
aldehyde dehydrogenase 1A1	NP_000680.2	2	IFINNEWHDSVSGK
			VAFTGSTEVGK
isocitrate dehydrogenase 1	NP_005887.2	2	TVEAEAAHGTVTR
			LIDDMVAQAMK
proteasome alpha 2 subunit	NP_002778.1	2	YNEDLELEDAIHTAILTLK
			HIGLVYSGMGPDYR
DNA directed RNA polymerase II polypeptide B	NP_000929.1	2	MTIGHLIECLQGK
			DCQIAHGAAQFLR
glutamate-ammonia Ligase	NP_002056.2	1	DIVEAHYR
asialoglycoprotein receptor 1	NP_001662	1	LLHVK
Asialoglycoprotein receptor 2	NP_550434	2	WVDGTDYR
			ADHDALLFHLK
Carbamoyl-phosphate synthetase 1, mitochondrial	NP_001866	1	GNDVLVIECNLR
3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2	NP_005509	3	LSIQCYLR
			YTVGLGQTR
			VSPEEFTEIMNQR
Transferrin	NP_001054	3	SCHTGLGR
			DQYELLCLDNTR
			TAGWNIPMGLLYNK
Plasminogen	NP_000292	3	TPENFPCK
			EAQLPVIENK
			HSIFTPETNPR
complement factor B preproprotein	NP_001701	2	DFHINLFQVLPWLK
			GDSGGPLIVHK
paraoxonase 1	NP_000437	2	IHVYEK
			HANWTLTPLK
ceruloplasmin precursor	NP_000087	2	IYHSHIDAPK
			TYCSEPEK
Interleukin 6 signal transducer	NP_002175	2	SHIAQWSPHTPPR
			HQVPSVQVFSR
Peroxisome biogenesis factor 1	NP_000457.1	2	DFTVLVDR
			GFLPASLR
Importin 4	NP_078934.3	2	ACYALENFVENLGPK
			LLPPLLQIVCK
Purine nucleotide phosphorylase	NP_000261.2	2	FEVGDIMLIR
			VFGFSLITNK
Malate dehydrogenase	NP_005908.1	2	ENFSCLTR
			LSSAMSAAK