Acta Biochim Biophys Sin (2008): 426-436 | © 2008 Institute of Biochemistry and Cell Biology, SIBS, CAS | All Rights Reserved 1672-9145
http://www.abbs.info; www.blackwellpublishing.com/abbs | DOI: 10.1111/j.1745-7270.2008.00418.x

How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes

Yanfang Guo[1,2#], Peng Xiao[1,3#], Shufeng Lei[1], Feiyan Deng[1], Gary Guishan Xiao[3], Yaozhong Liu[3], Xiangding Chen[1], Liming Li[1], Shan Wu[1], Yuan Chen[1], Hui Jiang[1], Lijun Tan[1], Jingyun Xie[4], Xuezhen Zhu[2], Songping Liang[4], and Hongwen Deng[1,2,5]*

[1] Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha 410081, China
[2] Key Laboratory of Biomedical Information Engineering, Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
[3] Osteoporosis Research Center and Department of Biomedical Sciences, Creighton University Medical Center, Omaha, Nebraska 68131, USA
[4] Key Laboratory of Protein Chemistry and Developmental Biology of Education Committee, College of Life Sciences, Hunan Normal University, Changsha 410081, China
[5] Departments of Orthopedic Surgery and Basic Medical Sciences, University of Missouri-Kansas City, Kansas City, Missouri 64108, USA

A key assumption in studying mRNA expression is that it is informative in the prediction of protein expression. However, only limited studies have explored the mRNA-protein expression correlation in yeast or human tissues and the results have been relatively inconsistent. We carried out correlation analyses on mRNA-protein expressions in freshly isolated human circulating monocytes from 30 unrelated women. The expressed proteins for 71 genes were quantified and identified by 2-D electrophoresis coupled with mass spectrometry. The corresponding mRNA expressions were quantified by Affymetrix gene chips. Significant correlation (r=0.235, P<0.0001) was observed for the whole dataset including all studied genes and all samples. The correlations varied in different biological categories of gene ontology. For example, the highest correlation was achieved for genes of the extracellular region in terms of cellular component (r=0.643, P<0.0001) and the lowest correlation was obtained for genes of regulation (r=0.099, P=0.213) in terms of biological process. In the genome, half of the samples showed significant positive correlation for the 71 genes and significant correlation was found between the average mRNA and the average protein expression levels in all samples (r=0.296, P<0.01). However, at the study group level, only five studied genes had significant positive correlation across all the samples. Our results showed an overall positive correlation between mRNA and protein expression levels. However, the moderate and varied correlations suggest that mRNA expression might be sometimes useful, but certainly far from perfect, in predicting protein expression levels.

Keywords protein expression; mRNA expression; circulating monocytes; gene function; correlation

Functional genomics, dealing with the functional analysis of genes and their products [1], is an active research area in the post-genomic era. Genetic information stored in DNA is translated into proteins with the intermediate of mRNA, and proteins are directly involved in almost all processes of life by carrying out various biological functions. Extensive studies have been carried out at the mRNA level and protein level separately for illuminating functions of genes. The studies at the mRNA level are largely based on a key assumption that mRNA expression is informative in predicting protein expression level in

Received: February 20, 2008 Accepted: March 27, 2008
The work was partially supported by the grants from the National Natural Science Foundation of China (Nos. 30600364, 30470534, and 30230210) and the Scientific Research Fund of the Hunan Provincial Education Department (No. 05B037). It was also partially supported by the LB595 grant from the State of Nebraska, USA. HWD was partially supported by grants from the US National Institutes of Health (Nos. K01 AR02170-01A2, R01 GM60402, 5R01 AR050496-02, and R21 AG027110-01A1)
Abbreviations: 2-DE, two-dimensional electrophoresis; CMCs, circulating monocytes; GO, gene ontology; MS, mass spectrometry; MS/MS, tandem mass spectrometry
# These authors contributed equally to this work
*Corresponding author: Tel/Fax, 86-731-8872791; Email, hwdeng@hunnu.edu.cn

relationto gene functions. However, the degree of validity of this key assumption has seldom been investigated, and is certainly far from being established.

Powerful high-throughput methods for measuring both mRNA and protein expressions are currently available, such as DNA microarray technologies for mRNA profiling, and 2-D electrophoresis (2-DE) or liquid chromatography coupled with mass spectrometry (MS) approaches for protein profiling. These technologies make it possible to study gene expression at both transcriptome and proteome levels, and provide data for assessing statistical correlations between mRNA and protein expression levels.

So far, only a few studies have investigated the correlation of mRNA and protein expression levels in yeast [2–4] and human tissues [5–7], and the results have been relatively inconsistent. In yeast, Gygi *et al* found that even similar mRNA expression patterns could be accompanied by large variations (more than 20-fold difference) of protein abundances, and vice versa [2]. Conversely, Futcher *et al* found a relatively high genomic mRNA and protein expression correlation (r=0.76) after normalizing the expression data using a logarithmic method [3]. Greenbaum *et al* found significant positive correlations between protein and mRNA expression in both global range and biological subcategories (such as cell cycle and cell rescue) (r=0.20–0.89), by merging many yeast protein-abundance datasets [4]. In humans, Anderson and Seilhamer examined 19 selected mRNA and corresponding protein abundances in a normal human liver expression database and found a positive correlation (r=0.48) [5]. However, a study on the expression correlation of three genes in 17 human prostate cancer tissue samples displayed no significant result for any gene [7]. Among 98 studied genes, 21 showed a significant positive correlation between mRNA and protein expression levels (r>0.2445, P<0.05) for a cohort of 85 lung cancer adenocarcinomas and normal lung samples [6]. The largely inconsistent results in both yeast and human tissues suggested that gene expressions at the transcriptional and translational levels might be different.

Here, to identify the relationship between gene expression at the mRNA and protein levels *in vivo* in normal human cells, we carried out an exhaustive correlation analyses of mRNA and protein expression abundances in human circulating monocytes (CMCs). Relative to using human tissue samples that could be composed of various proportions of different cells, we attempted to study a specific and relatively homogeneous human cell type, trying to avoid potential confounders due to various mixtures of different cells in tissue samples.

Materials and Methods

Subjects

The samples of this study came from an ongoing genetic project on osteoporosis [8] approved by the Research Administration of Hunan Normal University (Changsha, China). All subjects signed informed consent documents before entering the project. Thirty unrelated women were recruited and each donated 50 ml peripheral blood for this study. They were all randomly selected from a sample composed of 1000 healthy and normal adult Chinese Han females aged 20–45 years, living mainly in Changsha city in southern China. The sampling scheme [8] and exclusion criteria [9] have been detailed elsewhere. Briefly, the exclusion criteria included chronic disorders involving vital organs (heart, lung, liver, kidney, or brain), serious metabolic diseases such as diabetes, hypo- or hyperparathyroidism, hyperthyroidism, other skeletal diseases such as Paget's disease, osteogenesis imperfecta, rheumatoid arthritis, chronic use of drugs affecting bone metabolism such as corticosteroid therapy, anticonvulsant drugs, estrogens, thyroid hormone, and malnutrition conditions such as chronic diarrhea and chronic ulcerative colitis.

Monocyte extraction and purification

Mononuclear cells were isolated from the blood samples using Lymphoprep solution (Axis-Shield, Oslo, Norway), and CMCs were isolated from mononuclear cells using Dynal monocyte negative isolation kit (Dynal Biotech, Oslo, Norway). The purity of the isolated monocyte sample was monitored by BD FACSCalibur flow cytometry (BD Biosciences, San Jose, USA) with fluorescence-labeled antibodies PE-CD14 and FITC-CD45, and determined to be as high as 90%. The obtained CMCs from each sample were divided into two equal portions for RNA and protein extraction.

Protein profiling and identification

The experimental procedures for protein profiling and identification were detailed elsewhere [10]. Briefly, total cell lysates were obtained from monocyte samples and the proteins were quantified. To use as many samples as possible under our budget constraint, we used a sample pooling scheme for 2-DE [11]. Three protein samples were pooled with equal aliquots to form a mixed sample, thus 30 subjects were divided into 10 pooled samples for 2-DE. The composition of pooled samples is given in **Table 1**. Each pooled sample was subject to triplicate 2-DE, running on the IPGphor isoelectric focusing system

Table 1 Characteristics of 30 unrelated healthy adult Chinese Han females aged 20–45 years, living mainly in Changsha, China, who participated in this study

Pooled sample ID (*j*)	Original sample ID	Weight (kg)	Height (cm)	Age (years)
1	A1839*	54.5	156.5	33.69
	A2569*	49.0	156.0	25.23
	A0321*	47.0	151.5	23.33
2	A1203*	61.0	164.0	24.38
	A0171	51.5	166.0	21.85
	A0703	53.0	158.4	44.53
3	A2194*	49.5	159.5	35.49
	A1885*	56.5	165.0	28.01
	A1387	50.5	159.5	26.74
4	A1840*	64.0	168.0	30.06
	A1540*	58.5	161.5	25.07
	A1419	48.0	158.8	24.03
5	A2146*	64.5	161.5	38.50
	A1866*	58.2	152.5	25.05
	A1464*	53.5	152.0	29.20
6	A1098*	53.0	155.0	24.84
	A2593*	56.0	162.5	24.55
	A1002	49.0	151.8	27.70
7	A0989*	45.0	156.8	24.18
	A0953*	49.0	160.2	23.76
	A2384*	67.0	166.0	34.61
8	A0978*	45.5	158.5	24.61
	A0652*	55.0	161.0	24.01
	A2008	57.0	158.0	37.59
9	A0477*	48.5	152.5	22.67
	A1941*	47.5	157.7	25.16
	A1817	40.0	151.0	24.42
10	A0004*	47.0	155.4	25.20
	A0952	48.5	156.0	24.64
	A0690	43.0	150.0	28.77

j, mathematical order number of the pooled sample. *Subjects with available mRNA expression data.

(Amersham Pharmacia Biotech, Pittsburgh, USA) and Protein II xi cell electrophoresis apparatus (Bio-Rad Laboratories, Hercules, USA). PDQuest software (version 7.2; Bio-Rad Laboratories) was used to quantify the expression levels for the target protein spots on triplicate silver-stained 2-DE gels. In order to justify for variability due to silver staining, the intensity of each spot was normalized to the total intensity of all the spots in the gel. The normalized data were exported as expression values for further statistical analyses. The target spots were excised and trypsin digested into peptides. The extracted peptides were analyzed by MS and tandem MS (MS/MS) using a matrix-assisted laser desorption/ionization (MALDI)-time of flight/time of flight (TOF/TOF) Ultraflex instrument (Bruker Daltonik, Bremen, Germany). Finally, the MS and MS/MS data generated from the same protein spot were combined and searched against the NCBInr database of *Homo sapiens* (version 2.23.2005) using Mascot software (version 2.1; Matrix Science, Boston, USA). The proteins were identified using a probability-based Mowse algorithm [12]. A $-10\log(P)$ score, where P is the probability that the observed match is a random event, of >30 (for MS and MS/MS data) was regarded as significant.

mRNA quantification

The total RNA from monocytes was extracted using a Qiagen RNeasy mini kit (Qiagen, Valencia, USA) following the procedures recommended by the manufacturer. Individual samples were used for DNA microarray study. Preparation of cRNA, hybridization, and scanning of the Human Genome U133 Plus 2.0 GeneChip (Affymetrix, Santa Clara, USA) were carried out according to the manufacturer's protocol. The expression values were generated with Microarray Suite 5.0 software (Affymetrix). Expression data at the mRNA level were only available for 21 out of the 30 subject samples that had protein expression data (noted in **Table 1**). To match the corresponding protein expression abundances in each pooled sample in the protein analyses, the average mRNA expression value of the component individual samples in each pooled sample was calculated and used in subsequent analyses.

Statistical analysis

The relationship between the levels of mRNA and protein expression was examined using the Pearson's product-moment correlation coefficient statistical method. P values and interval estimation of correlation coefficients (r) were obtained through Fisher's Z transformation [13]. $P<0.05$ (two-sided) was considered statistically significant. The percentage of protein expression variation explained by corresponding mRNA expression was calculated as r^2. As we used the pooling scheme in the protein expression study, it is important to prove that correlation analyses based on this scheme are feasible. In **Appendix I** (http://www.abbs.info/fulltxt/40-05/supplement08055.htm), we present a detailed analytical rationale justifying our analyses

using averages in the pooled samples. Some of our pooled proteins and the corresponding mRNA samples are not perfectly matched so that, for three subjects composing one pooled protein sample, in a few cases only one or two of them were analyzed for mRNA expression (**Table 1**). We show in **Appendix I** that the correlation coefficients derived from this study are lower bounds approximating true values.

Six different types of correlation analyses were carried out, to reflect various aspects of correlation of mRNA-protein in the genomic and/or study group levels. To aid in comprehension of the computation and thus biological significance of these various types of correlations, the mathematical symbols used to denote the data and the resultant correlation coefficients are outlined below.

i: the *i*th gene, *i*=1, 2, …, 71.
j: the *j*th pooled sample, *j*=1, 2, …, 10.
$\bar{x}_{ij}$: the mRNA expression value for the *i*th gene in the *j*th pooled sample.
$\bar{y}_{ij}$: the protein expression value for the *i*th gene in the *j*th pooled sample.
$\bar{x}_{\cdot j}$: the mRNA expression values for all 71 genes in the *j*th pooled sample.
$\bar{y}_{\cdot j}$: the protein expression values for all 71 genes in the *j*th pooled sample.
$\bar{x}_{i\cdot}$: the mRNA expression values in all 10 pooled samples for the *i*th gene.
$\bar{y}_{i\cdot}$: the protein expression values in all 10 pooled samples for the *i*th gene.
$\bar{X}_i$: the average mRNA expression value of the 10 samples for the *i*th gene.
$\bar{Y}_i$: the average protein expression value of the 10 samples for the *i*th gene.

The analyses include following steps.
(1) An all-against-all correlation for all the studied genes in all the pooled samples between mRNA and protein expression values was conducted to assess the correlation characteristics across both genomic and group levels. In this correlation, the x, y variables, correlation coefficient, and P value are $\bar{x}_{ij}$, $\bar{y}_{ij}$, r_1, and P_1, respectively.
(2) On a genomic scale of individual samples, correlation for all the studied genes was first calculated for each pooled sample then an average correlation coefficient was obtained. The variables and parameters in this correlation analysis are $\bar{x}_{\cdot j}$, $\bar{y}_{\cdot j}$, r_{2j} and P_{2j}, respectively. The average correlation coefficient of r_{2j} is denoted by $\overline{r_2}$.
(3) To investigate the relation between the correlation of average expression values in the study group and the average correlation $\overline{r_2}$ (computed above) among different individuals, we obtained the average mRNA and protein expression values of all samples, then carried out the correlation using these average values. $\bar{X}_i$, $\bar{Y}_i$, r_3, and P_3 represent the variables and parameters in this analysis.
(4) To explore whether different genes behave differently on their mRNA-protein expressions at the group level, for each gene, we calculated the correlation of its mRNA and protein expression values across the pooled samples. The variables and parameters for this analysis are $\bar{x}_{i\cdot}$, $\bar{y}_{i\cdot}$, r_{4i}, and P_{4i}.
(5) To assess the mRNA-protein correlations among different biological categories, we conducted expression correlation analyses for different gene clusters in different biological categories. We first classified the 71 genes according to gene ontology (GO) principles using GFINDer software [14], then carried out the expression correlation for each gene cluster across all the pooled samples. $\bar{x}_{ij_m}$, $\bar{y}_{ij_m}$, r_{5m}, and P_{5m} are the variables and parameters involved (*m*, the mth biological category, *m*=1, 2, …, 12; j_m, the *j*th gene in the *m*th biological category, j_m ranges from 1 to the total number of genes in the mth biological category).
(6) To find whether the expression correlation is dependent on expression values per se, correlation analysis was carried out between the mRNA-protein correlation coefficients and average mRNA or protein expression values of all samples. In this correlation, *x*, *y*, correlation coefficient, and *P* value are r_{2j}, $\bar{X}_i$ / $\bar{Y}_i$, r_{6x}/r_{6y}, and P_{6x}/P_{6y}, respectively.

Results

The characteristics of the studied samples are shown in **Table 1**. We selected 105 protein spots on 2-DE gels for quantification and identification. Some proteins migrated to more than one spot (commonly seen in proteomics, possibly due to differential protein processing or modification) and abundances for such proteins were calculated by integrating the intensities of the different spots. In total, protein quantities of 71 genes were determined (**Table 2**). The range of average protein expression values after normalization was from 12580.07 (*ENO1*) to 161743.13 (*PARK7*), whereas that of average mRNA expression values was from 3.45 (*PKLR*) to 5578.65 (*ACTB*). All of the mRNA expression data was submitted to the National Center for Biotechnology Information's Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/, accession No. GSE7158).

Correlation for all genes in all samples

mRNA expression abundance was highly significantly positively correlated with the protein expression level for

Table 2 Characteristics of 71 studied genes and results of the correlation for individual genes across all the samples

Gene	Gene ID	Description of identified protein	$\overline{X}_i$	$\overline{Y}_i$	r_{4i}	P_{4i}	*Var* (%)	*CI*	
ATIC	471	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/ IMP cyclohydrolase	100.295	30994.847	0.832	**<0.001**	69.22	0.624	0.93
SLA2	84174	Src-like adapter protein-2	95.498	45300.323	0.823	**<0.001**	67.73	0.607	0.926
TPI1	7167	Triosephosphate isomerase complexed with 2-phosphoglycolic acid	756.845	54846.313	0.542	**0.01**	29.38	0.144	0.789
PRDX3	10935	Peroxiredoxin 3	226.114	47483.085	0.528	**0.013**	27.88	0.124	0.781
SOD2	6648	Human manganese superoxide dismutase mutant Q143n	390.917	109078.721	–0.497	**0.021**	24.70	–0.765	–0.08
RSU1	6251	Ras suppressor protein 1	277.876	78456.198	0.476	**0.028**	22.66	0.056	0.753
GPX1	2876	Glutathione peroxidase	3153.781	78238.175	–0.466	**0.032**	21.72	–0.747	–0.04
GSTP1	2950	Glutathione S-transferase	551.405	51642.377	0.416	0.06	17.31	–0.019	0.719
ENO1	2023	Enolase 1	571.306	12580.068	–0.405	0.069	16.40	–0.712	0.033
PHB	5245	Prohibitin	90.926	22537.456	–0.389	0.082	15.13	–0.703	0.051
TPM4	7171	Tropomyosin 4	131.421	46213.315	0.383	0.087	14.67	–0.059	0.699
THBS1	7057	Thrombospondin	176.578	51256.133	–0.382	0.088	14.59	–0.698	0.06
GPD2	2820	Glycerol-3-phosphate dehydrogenase	50.327	29877.945	–0.335	0.139	11.22	–0.67	0.113
MMP28	79148	Eilysin	7.546	26447.573	–0.333	0.142	11.09	–0.669	0.115
ACTG1	71	γ-Actin	3622.884	24256.309	–0.332	0.143	11.02	–0.668	0.116
ATP5H	10476	ATP synthase, H^+ transporting, mitochondrial F0 complex, subunit d	380.398	91213.998	–0.308	0.176	9.49	–0.653	0.142
CA2	760	Carbonic anhydrase Ii	416.300	42594.415	0.298	0.193	8.88	–0.154	0.646
PSMA2	5683	Proteasome (prosome, macropain) subunit, α type, 2	344.745	66401.627	0.286	0.212	8.18	–0.166	0.639
VCL	7414	Vinculin	692.257	37294.101	0.272	0.236	7.40	–0.181	0.63
PGLS	25796	6-phosphogluconolactonase	220.833	51723.569	–0.269	0.242	7.24	–0.628	0.184
OXCT1	5019	3-oxoacid CoA transferase 1 precursor	39.552	21043.245	0.268	0.243	7.18	–0.185	0.627
K-ALPHA-1	10376	α Tubulin; K-α-1 protein	1643.770	38099.539	0.265	0.249	7.02	–0.188	0.625
CLIC1	1192	Chloride intracellular channel 1	1231.814	43734.630	0.263	0.254	6.92	–0.191	0.624
PLEK	5341	Pleckstrin	875.107	31149.048	0.261	0.256	6.81	–0.192	0.623
NP	4860	Purine nucleoside phosphorylase	188.248	60802.262	0.256	0.267	6.55	–0.198	0.619
CCT2	10576	Chaperonin containing TCP1	192.919	38678.090	0.252	0.274	6.35	–0.201	0.617
MAPK9	5601	JNK2 α1 protein kinase	93.731	19380.322	0.233	0.313	5.43	–0.22	0.604
ECH1	1891	Enoyl coenzyme A hydratase 1, peroxisomal	152.867	21340.807	–0.231	0.319	5.34	–0.602	0.223
ETFA	2108	3-D structure of human electron transfer flavoprotein to 2.1 Å resolution	392.705	57922.725	–0.229	0.322	5.24	–0.601	0.225
NRN1	51299	Neuritin	10.933	26612.918	–0.228	0.324	5.20	–0.601	0.225
NME1	4830	Nucleoside-diphosphate kinase 1	112.800	53339.370	–0.211	0.364	4.45	–0.589	0.243
C21orf33	8209	Homolog of zebrafish ES1	102.776	57340.718	0.201	0.386	4.04	–0.252	0.582
MGLL	11343	Monoglyceride lipase isoform 2	185.324	63934.777	–0.184	0.43	3.39	–0.57	0.269
ESRRB	2103	Estrogen receptor-related receptor β short form	6.921	15515.017	–0.182	0.436	3.31	–0.569	0.271
PBP	5037	Prostatic binding protein (phosphatidylethanolamine binding protein)	127.290	57571.265	0.179	0.443	3.20	–0.274	0.567

To be continued

Continued

Gene	Gene ID	Description of identified protein	$\overline{X}_i$	$\overline{Y}_i$	r_{4i}	P_{4i}	*Var* (%)	*CI*	
ANXA2	302	Annexin A2, isoform 2	1370.193	45047.525	0.177	0.447	3.13	–0.276	0.566
ELN	2006	Tropoelastin	17.921	14789.108	0.173	0.458	2.99	–0.279	0.563
SULT1A3	6818	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3	303.467	46022.128	–0.162	0.487	2.62	–0.555	0.29
FGA	2243	Fibrinogen α chain preproprotein	7.712	67180.750	0.157	0.502	2.46	–0.295	0.551
U2AF1	7307	U2 small nuclear RNA auxiliary factor 1	139.643	107271.835	0.147	0.529	2.16	–0.304	0.544
PSMA5	5686	Proteasome (prosome, macropain) subunit, α type, 5	136.810	79263.113	0.129	0.582	1.66	–0.321	0.531
GSTO1	9446	Glutathione-S-transferase ω 1	838.886	84369.845	–0.123	0.6	1.51	–0.527	0.326
FGB	2244	Fibrin β (fibrinogen β chain)	12.890	43585.840	0.118	0.614	1.39	–0.33	0.523
TLN1	7094	Talin 1	176.583	84783.605	–0.117	0.619	1.37	–0.522	0.332
RAB7B	338382	Member RAS oncogene family	19.974	68940.302	0.114	0.628	1.30	–0.334	0.52
ALDOA	226	Human muscle fructose 1,6-bisphosphate aldolase complexed with fructose 1,6-bisphosphate	1190.295	63151.838	–0.109	0.642	1.19	–0.517	0.338
ARHGDIB	397	GDP dissociation inhibitor β	1750.240	41000.739	0.108	0.646	1.17	–0.34	0.516
GAPDH	8302	Glyceraldehyde-3-phosphate dehydrogenase	166.786	25979.082	–0.108	0.646	1.17	–0.515	0.34
RWDD2	112611	Hypothetical protein MGC13523	12.662	26890.752	0.103	0.66	1.06	–0.344	0.512
TTC17	55761	Tetratricopeptide repeat domain 17	91.760	36430.382	0.099	0.675	0.98	–0.348	0.509
PKM2	5315	Pyruvate kinase 3	701.676	42357.755	–0.095	0.685	0.90	–0.506	0.351
POLB	5423	DNA polymerase β	109.155	71119.917	0.085	0.717	0.72	–0.36	0.499
TUBA6	84790	α-Tubulin; tubulin α 6	1433.833	44170.669	0.084	0.722	0.71	–0.361	0.497
UGP2	7360	UDP-glucose pyrophosphorylase 2	257.860	36345.692	0.077	0.743	0.59	–0.367	0.492
AK2	204	Adenylate kinase 2 isoform b; adenylate kinase, mitochondrial	114.939	54253.569	0.077	0.744	0.59	–0.367	0.492
P4HB	5034	Prolyl 4-hydroxylase, β subunit	412.310	88347.105	0.076	0.745	0.58	–0.367	0.492
CORO1A	11151	Coronin-like protein	1195.014	30591.578	0.072	0.761	0.52	–0.372	0.488
PARK7	11315	DJ-1 protein	1528.052	161743.128	0.067	0.774	0.45	–0.375	0.485
IDH2	3418	Isocitrate dehydrogenase 2 ($NADP^+$), mitochondrial precursor	93.071	49580.305	–0.063	0.788	0.40	–0.482	0.379
WDR1	9948	WD repeat-containing protein 1, isoform 1	374.337	40004.797	–0.053	0.822	0.28	–0.474	0.388
PON3	5446	Paraoxonase-3	3.629	15305.943	0.051	0.83	0.26	–0.39	0.472
YWHAG	56010	Tyrosine 3-monooxgenase/tryptophan 5-monooxgenase activation protein, γ polypeptide	888.286	54409.031	0.046	0.847	0.21	–0.394	0.468
CSF1	1435	Macrophage colony-stimulating factor	17.423	56586.350	0.042	0.858	0.18	–0.397	0.465
HRH4	59340	Histamine receptor subunit peptide 4	20.007	83392.122	0.042	0.86	0.18	–0.397	0.465
MPST	4357	Mercaptopyruvate sulfurtransferase	111.448	16354.185	–0.032	0.892	0.10	–0.457	0.405
PKLR	5313	Pyruvate kinase	3.450	33920.012	–0.026	0.913	0.07	–0.452	0.41
CASP8	841	Procaspase-8	175.638	17170.903	–0.019	0.936	0.04	–0.447	0.416
ACTB	60	Actin, β	5578.650	117215.305	–0.016	0.947	0.03	–0.444	0.419
CAP1	10487	Adenylyl cyclase-associated protein	1519.945	58209.193	–0.001	0.996	0.00	–0.433	0.431
GSN	2934	Gelsolin isoform b	95.329	26928.388	–0.001	0.997	0.00	–0.432	0.431
ARID1B	57492	RP11-419L10.1	52.962	27253.323	0.000	0.999	0.00	–0.431	0.432

CI, 95% confidence interval estimated for the correlation coefficient; r_{4i} and P_{4i}, correlation coefficient and *P* value for the *i*th gene, respectively; Var, the percentage of protein expression variation explained by mRNA expression (determined as r^2); $\overline{X}_i$, the average mRNA expression values for the *i*th gene; $\overline{Y}_i$, the average protein expression value for the *i*th gene. *P* values less than 0.05 are in bold.

71 genes in all the samples (r_1=0.235, P_1<0.0001). **Fig. 1 (A)** shows that increasing mRNA expression is generally accompanied by increasing protein expression, but with relatively large noise. Gene expression variation at the mRNA level accounts for approximately 5.52% of the protein expression variation.

Correlation for all genes in each pooled sample

Positive mRNA-protein correlation was observed for all genes in each pooled sample, with r_{2j} (j=1, 2, …, 10) ranging from 0.109 to 0.443 and the percentage of protein variation explained by mRNA variation ranging from 1.19 to 19.62 (**Table 3**). Half of the correlations displayed statistical significance (r>0.234, P<0.05), and another two (j=5, r_{25}=0.217; j=6, r_{26}=0.230) showed marginal significance (P_{25}=0.069 and P_{26}=0.070). The $\overline{r_2}$ was 0.270 with a standard deviation of 0.111.

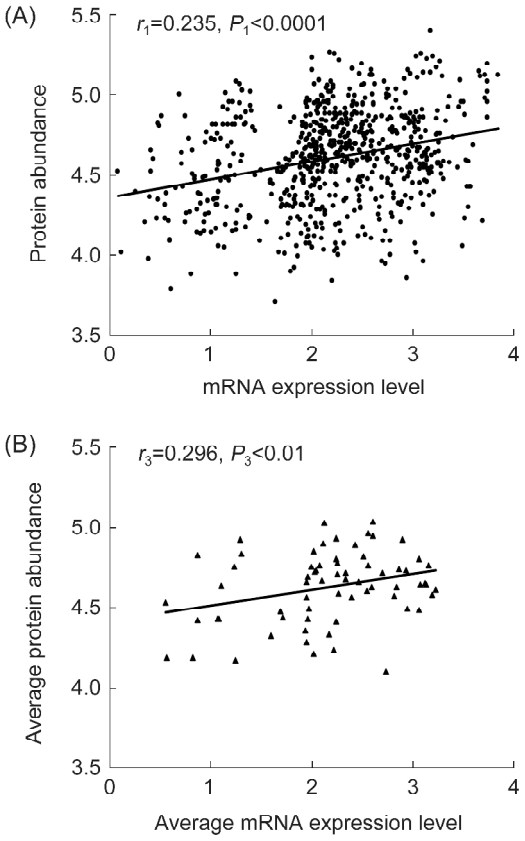

Fig. 1 Correlation between protein and mRNA expression levels of 71 studied genes Expression values are plotted on a log scale, whereas original data were used to calculate the correlation coefficients (for both data forms, the correlation coefficients were very similar to each other). The solid line indicates the correlation trend. (A) mRNA-protein correlation for expression values of all the genes across all the samples. (B) mRNA-protein correlation with the expression values of each gene averaged over all the samples.

Table 3 mRNA-protein expression correlation results for all genes in each pooled sample

Pooled sample ID (j)	r_{2j}	P_{2j}	Var (%)	CI	
1	0.162	0.189	2.62	–0.074	0.381
2	0.288	**0.015**	8.29	0.059	0.488
3	0.109	0.369	1.19	–0.127	0.334
4	0.350	**0.003**	12.25	0.127	0.539
5	0.217	0.070	4.71	–0.018	0.428
6	0.230	0.069	5.29	–0.004	0.439
7	0.299	**0.012**	8.94	0.072	0.498
8	0.183	0.126	3.35	–0.052	0.399
9	0.423	**0.0002**	17.89	0.210	0.597
10	0.443	**0.0002**	19.62	0.234	0.613

CI, 95% confidence interval estimated for the correlation coefficient; j, mathematical order number of the pooled sample; r_{2j} and P_{2j}, correlation coefficient and P value of the genomic correlation for the jth pooled sample, respectively; *Var*, the percentage of protein expression variation explained by mRNA expression (determined as r^2). P values less than 0.05 are in bold.

Correlation between average mRNA and average protein expression values

Significant correlation was detected using the average mRNA and protein expression values of each of the 71 genes across the human genome (r_3=0.296, P_3<0.01) [**Fig. 1(B)**]. The r_3 was almost equal to the $\overline{r_2}$.

Correlation across all the samples for each gene

The r_{4i} (i=1, 2, …, 71) ranged from –0.497 to 0.892. Although relatively large correlation coefficients were achieved, only five genes (*ATIC*, *SLA2*, *TPI1*, *PRDX3*, and *RSU1*) showed significant positive correlations (**Table 2**), and two of the negative correlations were significant.

Correlation for each gene category in all samples

For each GO principle (biological process, cellular component, and molecular function), the corresponding gene categories at level one were studied. **Table 4** shows the correlation results for each GO category with the percentage of genes over 5% (the proportion of the number of genes belonging to the GO category to that of the parent ontology). There are overlaps among different categories due to some genes with multiple functions.

The correlation strength for most of the categories turned out to be similar to the whole gene set. However,

Table 4 mRNA-protein expression correlation for all samples in each gene ontology (GO) category

GO term	*Num* (%)	*m*	r_{5m}	P_{5m}	*Var* (%)	*CI*	
Biological process	64						
Cellular (GO:0009987)	56 (88)	1	0.212	**<0.0001**	4.49	0.131	0.290
Physiological (GO:0007582)	40 (63)	2	0.202	**<0.0001**	4.08	0.106	0.294
Regulation (GO:0050789)	16 (25)	3	0.099	0.2130	0.98	–0.057	0.250
Development (GO:0050896)	10 (16)	4	0.163	0.1050	2.66	–0.035	0.348
Cellular component	62						
Cell (GO:0005623)	54 (87)	5	0.221	**<0.0001**	4.88	0.139	0.300
Organelle (GO:0043226)	45 (73)	6	0.246	**<0.0001**	6.05	0.157	0.331
Protein complex (GO:0043234)	16 (26)	7	0.388	**<0.0001**	15.05	0.248	0.512
Extracellular region (GO:0005576)	15 (24)	8	0.643	**<0.0001**	41.35	0.538	0.728
Molecular function	69						
Binding (GO:0005488)	41 (59)	9	0.274	**<0.0001**	7.51	0.182	0.361
Catalytic activity (GO:0003824)	41 (59)	10	0.302	**<0.0001**	9.12	0.211	0.388
Structural molecule activity (GO:0005198)	10 (14)	11	0.250	**0.0120**	6.25	0.056	0.426
Signal transducer activity (GO:0004871)	9 (13)	12	0.476	**<0.0001**	22.66	0.298	0.622

CI, 95% confidence interval estimated for the correlation coefficient; *m*, the *m*th biological category; *Num*, number of genes; r_{5m} and P_{5m}, correlation coefficient and *P* value for all genes in the *m*th biological category for all samples respectively; *Var*, the percentage of protein expression variation explained by mRNA expression (determined as r^2). *P* values less than 0.05 are in bold.

genes belonging to regulation (n=16, GO: 0050789, r_{53}=0.099, P_{53}=0.213) and development (n=10, GO: 0050896, r_{54}=0.163, P_{54}=0.105) of biological processes had non-significant and lowest mRNA-protein expression correlations. However, genes belonging to the extracellular region (GO: 0005576, n=15, r_{58}=0.643, P_{58}<0.0001) of cellular component and signal transducer activity (GO: 0004871, n=9, r_{512}=0.476, P_{512}<0.0001) of molecular function showed significant and highest correlation (**Table 4**). Interestingly, the percentage of protein expression variation explained by mRNA expression variation varied from 0.98% to 41.35% for different categories studied here.

mRNA-protein correlation strength is independent of mRNA or protein expression abundance *per se*

No significant relationship was observed for mRNA, r_{6x}=0.034, P_{6x}>0.05, or for protein, r_{6y}=0.082, P_{6y}>0.05 (**Fig. 2**).

Discussion

In this study, we investigated the mRNA and protein expression levels of 71 genes in human CMCs and carried out comprehensive correlation analyses on different levels and aspects of mRNA-protein expression. This is the first expression correlation study in a specific type of normal human cell. We found that, overall, gene expression at the mRNA level is generally correlated with that at the protein level for all the genes studied, reflected by the findings in the whole studied samples (the all-against-all correlation and correlations for GO categories) and in individual subjects (at least 50% of the subjects showed significant positive correlations). Interestingly, we also found that

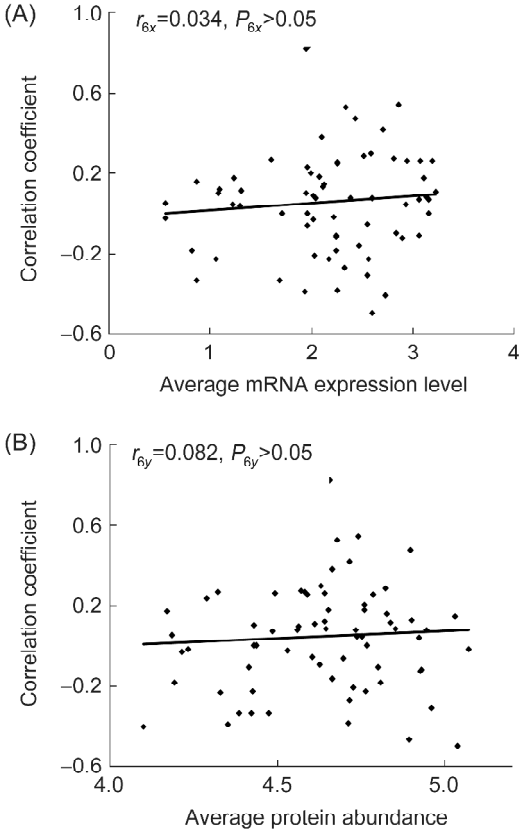

Fig. 2 Relationship between mRNA or protein expression values *per se* versus correlation strength of mRNA-protein expression of 71 studied genes Expression values are plotted in a log scale. (A) The relationship in terms of mRNA expression levels. (B) The relationship in terms of protein expression levels. r_{6x}/r_{6y} and P_{6x}/P_{6y}, the correlation coefficients and *P* value between the mRNA-protein correlation coefficients and average mRNA or protein expression values of all samples.

mRNA-protein expression correlation largely varies for different genes and different gene biological categories. These findings confirmed previous results in human normal liver cells [5], lung adenocarcinomas [6], and in yeast [4] on a genomic scale. All of these studies found an overall significant positive correlation between mRNA and protein expression. However, this study and others also suggested that the correlation strength varies among genes, and only a subset of the studied mRNAs had a significant positive correlation with the protein expression.

At the genomic level, interestingly, we found positive correlations in the whole study group (r_1=0.235, P_1 <0.0001) [**Fig. 1(A)**], individual samples (r_{2j}=0.109–0.443) (**Table 3**), and the average expression level (r_3=0.296, P_3 <0.01) [**Fig. 1(B)**]. Although r_{2j} displayed variation among different samples, the average value ($\overline{r_2}$ =0.270) showed consistency with the other two coefficients (r_1=0.235 and r_3=0.296). Those overall similar positive correlations suggest that there is a general approximate concordance between the mRNA and the protein expression quantities on the genomic scale, both for the group and the individuals. However, alarmingly, the mRNA variation could only account for less than 9% of the protein variation in all the three types of correlation analyses, which might indicate that, at the genomic level, mRNA expression poorly predicts the corresponding protein expression level. Our results also suggest that mRNA-protein expression correlation might also vary in different individuals. Whether this initial finding holds, especially in relation to the health status of individual subjects, might deserve further investigation.

For individual genes across all the samples, however, only five genes (**Table 2**) showed significant positive mRNA-protein correlations. For these five genes, expression at the mRNA level could explain as much as 69.22% of the expression variation at the protein level. These five genes were *ATIC*, *SLA2*, *TPI1*, *PRDX3*, and *RSU1*. They are involved in important molecular functions such as protein binding, enzyme activities, and signal transduction. This finding suggests that some biologically critical genes possibly have highly correlated mRNA-protein expression and those key high correlations might be important for the whole expression system for critical organism function or health.

However, lack of significant correlation findings for other genes could be due to technological and biological reasons, as well as the limitation of the experimental design of this study. First, currently available technological methods are not perfectly accurate in either mRNA or protein quantifications. These factors potentially influence the detection of biologically truly significant correlations. Second, complicated biological processes, such as transcriptional splicing, post-transcriptional splicing, translational modifications, translational regulation, and protein complex formation, might affect the relative quantities of mRNA and protein of various genes to various degrees. Third, different biological or experimental mRNA and protein degradation rates might affect the mRNA and protein correlations. Fourth, different mRNA secondary structures might result in different protein translation efficiencies. Finally, the pooling scheme resulting in a relatively small sample size (n=10) might limit the power to detect significant mRNA-protein correlations. As pointed out in "Materials and Methods", and shown analytically in **Appendix I**, missing information regarding mRNA expression for some subjects, as indicated in **Table 1**, renders the correlations computed in this study the lower

bounds (approximately two-thirds) of true correlations. These factors considered, it suffices to say that the correlations found in this study are lower bounds of true biological correlations.

We found two genes with significant negative expression correlation. This might be due to artifacts due to the technical noise, and/or statistical multiple testing, or a novel mRNA to protein regulation mechanism (e.g., negative feedback) that might need to be explored.

Genes belonging to the same functional category might share similar stringency or regulatory mechanisms for the expressions of mRNA and corresponding proteins. This could underlie the findings here that the correlations vary by different gene category. Functionally similar categories of genes/proteins might contribute importantly to the degree of correlation between mRNA and protein expression [4]. Motivated by this hypothesis, we grouped the genes into 12 categories according to three GO principles (biological process, cellular component, and physical biology) and carried out expression correlations for genes in each category across all the samples. All the 12 categories showed positive correlations, and categories of biological process showed relatively low correlations compared to other categories. The non-significant correlation for "regulation" and "development" categories of biological process (**Table 4**) might suggest that genes involved in regulation and development of a biological process undergo relatively less stringent translational regulation with more discordant expression patterns at both the mRNA and protein levels, in the specific cell type under study, CMCs. This finding might be surprising given that, intuitively, genes for "regulation" and "development" categories of biological process should be under stringent regulation for translation. This result could be related to the roles of these genes in CMCs and they might be under different translation regulation mechanisms in other cell types that could be more relevant to growth and development.

Some studied genes were classified into more than one functional category, which might mask the correlation difference of different categories. The difference would be more apparent if more genes were involved in the study. However, the percentage of protein variation explained by mRNA still showed a relatively large range (from 0.98% to 41.35%) for different categories. Most of the categories showed significant mRNA-protein correlations, substantiating that gene expression at transcription and translation levels were correlated globally. However, the large difference among the percentage of protein expression accounted for by mRNA could suggest that there is relatively large expression divergence between the transcription and translation levels for genes in different biological categories.

In normal human CMCs, our correlation results in both individual genes and different categories showed big variations. We speculate that in different developmental stages or disease conditions, the expression correlations for specific genes or biological categories might be more discordant, for example, some genes or gene categories have significant and high correlations and other genes or gene categories showed non-significant or low correlations, all perhaps related to the specific functions of the genes in specific cell types. It is conceivable that the same gene or gene category has very different mRNA-protein correlations in different cell types. Therefore, in gene expression studies for development or diseases, the developmental or disease gene expression measurement at the mRNA level using methods such as real-time reverse transcription-polymerase chain reaction might not be sufficient and protein level studies using, for example, Western blot analysis, are necessary in cells that are related to the development or disease etiology under study.

Studies suggested that mRNA or protein expression abundance *per se* is probably a factor influencing the mRNA-protein expression correlation [4,6]. However, we found that the mRNA-protein expression correlation coefficients are independent of the mRNA or protein abundances, results that are not in agreement with the findings in yeast [4] but are consistent with those from human lung adenocarcinomas [6]. Whether this discrepancy represents the species or cell type differences under study is not clear.

Proteins are the major direct executors of life processes and might reflect gene function more directly than mRNA. In addition, some messengers are transcribed but not translated, thus the number of mRNA copies does not necessarily reflect the number of functional protein molecules. However, mRNA expression is still informative to protein expression, given the study here and those reported earlier. Therefore, both mRNA and protein expression studies are important and complementary to each other. In combination, they should provide us with a more detailed and comprehensive understanding of gene function. The positive correlation between mRNA and protein expression as observed in human CMCs (in the current study) and in human cancer tissues [6,7], as well as in yeast [3,4], are biologically significant and worth further investigation. Currently, a challenge facing scientists is to develop and implement methodologies to integrate and analyze large-scale datasets at different levels of gene/protein expression to unravel the underlying regulatory processes or complex metabolic networks *in*

vivo. Ideker *et al* reported on an integrated approach to build, test, and refine a model of a cellular pathway, in which perturbations to critical pathway components are analyzed using DNA microarray, quantitative proteomics, and databases of known physical interactions [15]. In addition, the models jointly analyzing transcriptomic and proteomic data of *Desulfovibrio vulgaris* [16] correctly predicted an abundance of undetected proteins in some cases. With better understanding of the relationship between mRNA and protein expression, we can optimize combining analyses and provide additional insights into the metabolic mechanisms underlying complex biological systems [17,18].

In summary, this is the first multilevel expression correlation study in normal human cells, CMCs. CMCs are an important cell type in the human immune system. They play an important role in many physiological and pathological conditions, such as cancer, inflammation, infection, and bone metabolism. Our findings indicated that gene expression at the mRNA level is generally informative but not predictive for that at the protein level. Thus, functional genomics approaches at both mRNA and protein levels are necessary and complementary for human complex disease studies.

References

1 Celis JE, Kruhoffer M, Gromova I, Frederiksen C, Ostergaard M, Thykjaer T, Gromov P *et al*. Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. FEBS Lett 2000, 480: 2–16

2 Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 1999, 19: 1720–1730

3 Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrels JI. A sampling of the yeast proteome. Mol Cell Biol 1999, 19: 7357–7368

4 Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. Genome Biol 2003, 4: 117

5 Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 1997, 18: 533–537

6 Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ *et al*. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics 2002, 1: 304–313

7 Lichtinghagen R, Musholt PB, Lein M, Romer A, Rudolph B, Kristiansen G, Hauptmann S *et al*. Different mRNA and protein expression of matrix metalloproteinases 2 and 9 and tissue inhibitor of metalloproteinases 1 in benign and malignant prostate tissue. Eur Urol 2002, 42: 398–406

8 Tan LJ, Lei SF, Chen XD, Liu MY, Guo YF, Xu H, Sun X *et al*. Establishment of peak bone mineral density in Southern Chinese males and its comparisons with other males from different regions of China. J Bone Miner Res 2007, 25: 114–121

9 Deng HW, Xu FH, Huang QY, Shen H, Deng H, Conway T, Liu YJ *et al*. A whole-genome linkage scan suggests several genomic regions potentially containing quantitative trait loci for osteoporosis. J Clin Endocrinol Metab 2002, 87: 5151–5159

10 Deng FY, Liu YZ, Xiao P, Jiang C, Li LM, Wu S, Chen Y *et al*. *In vivo* proteomics analysis of circulating monocytes in Chinese females underlying osteoporosis. Proteomics 2008, in Press

11 Peng X, Wood CL, Blalock EM, Chen KC, Landfield PW, Stromberg AJ. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics 2003, 4: 26

12 Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20: 3551–3567

13 Rosner B. Fundamentals of Biostatistics. 5th edn. Pacific Grove: Duxbury Press 2005

14 Masseroli M, Galati O, Pinciroli F. GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. Nucleic Acids Res 2005, 33: W717–W723

15 Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 2001, 292: 929–934

16 Nie L, Wu G, Brockman FJ, Zhang W. Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: Zero-inflated Poisson regression models to predict abundance of undetected proteins. Bioinformatics 2006, 22: 1641–1647

17 Lee PS, Shaw LB, Choe LH, Mehra A, Hatzimanikatis V, Lee KH. Insights into the relation between mRNA and protein expression patterns: II. Experimental observations in *Escherichia coli*. Biotechnol Bioeng 2003, 84: 834–841

18 Mehra A, Lee KH, Hatzimanikatis V. Insights into the relation between mRNA and protein expression patterns: I. Theoretical consideration. Biotechnol Bioeng 2003, 84: 822–833