



OPEN ACCESS

ORIGINAL ARTICLE

# Mapping of hepatic expression quantitative trait loci (eQTLs) in a Han Chinese population

Xiaoliang Wang,<sup>1</sup> Huamei Tang,<sup>2</sup> Mujian Teng,<sup>3</sup> Zhiqiang Li,<sup>4</sup> Jianguo Li,<sup>5</sup> Junwei Fan,<sup>1</sup> Lin Zhong,<sup>1</sup> Xing Sun,<sup>1</sup> Junming Xu,<sup>1</sup> Guoqing Chen,<sup>1</sup> Dawei Chen,<sup>1</sup> Zhaowen Wang,<sup>1</sup> Tonghai Xing,<sup>1</sup> Jinyan Zhang,<sup>1</sup> Li Huang,<sup>1</sup> Shuyun Wang,<sup>1</sup> Xiao Peng,<sup>1</sup> Shengying Qin,<sup>4</sup> Yongyong Shi,<sup>4</sup> Zhihai Peng<sup>1</sup>

► Additional material is published online. To view please visit the journal (<http://dx.doi.org/10.1136/jmedgenet-2013-102045>).

For numbered affiliations see end of article.

## Correspondence to

Dr Zhihai Peng, Department of General Surgery, Shanghai First People's Hospital, Medical College, Shanghai Jiaotong University, 100 Haining Road, Shanghai, 200080, The People's Republic of China; pengzhihai@sjtu.edu.cn  
Or Dr Yongyong Shi, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Bio-X Institutes, Ministry of Education, Shanghai Jiao Tong University; Shanghai genome Pilot Institutes for Genomics and Human Health, 50 West Guangyuan Road, Shanghai, 200030, The People's Republic of China; shiyongyong@gmail.com

XW, HT, and MT contributed equally to the work.

Received 10 September 2013  
Accepted 27 February 2014  
Published Online First  
24 March 2014



Open Access  
Scan to access more  
free content



CrossMark

To cite: Wang X, Tang H, Teng M, et al. *J Med Genet* 2014;**51**:319–326.

## ABSTRACT

**Background** Elucidating the genetic basis underlying hepatic gene expression variability is of importance to understand the aetiology of the disease and variation in drug metabolism. To date, no genome-wide expression quantitative trait loci (eQTLs) analysis has been conducted in the Han Chinese population, the largest ethnic group in the world.

**Methods** We performed a genome-wide eQTL mapping in a set of Han Chinese liver tissue samples (n=64). The data were then compared with published eQTL data from a Caucasian population. We then performed correlations between these eQTLs with important pharmacogenes, and genome-wide association study (GWAS) identified single nucleotide polymorphisms (SNPs), in particular those identified in the Asian population.

**Results** Our analyses identified 1669 significant eQTLs (false discovery rate (FDR) < 0.05). We found that 41% of Asian eQTLs were also eQTLs in Caucasians at the genome-wide significance level ( $p=10^{-8}$ ). Both *cis*- and *trans*-eQTLs in the Asian population were also more likely to be eQTLs in Caucasians ( $p<10^{-4}$ ). Enrichment analyses revealed that trait-associated GWAS-SNPs were enriched within the eQTLs identified in our data, so were the GWAS-SNPs specifically identified in Asian populations in a separate analysis ( $p<0.001$  for both). We also found that hepatic expression of very important pharmacogenetic (VIP) genes (n=44) and a manually curated list of major genes involved in pharmacokinetics (n=341) were both more likely to be controlled by eQTLs ( $p<0.002$  for both).

**Conclusions** Our study provided, for the first time, a comprehensive hepatic eQTL analysis in a non-European population, further generating valuable data for characterising the genetic basis of human diseases and pharmacogenetic traits.

## INTRODUCTION

The liver is a vital human organ for a variety of physiological processes, and plays a major role in drug metabolism. Elucidating the basis for the genetic variation in hepatic gene expression will significantly further our understanding of human diseases and pharmacogenomics.

Expression quantitative trait loci (eQTLs) is one of the most effective ways to discover gene regulation networks.<sup>1</sup> The eQTLs method measures the variance in gene transcription, followed by

mapping the genetic loci affecting the expression of mRNA.<sup>2</sup> To date, eQTLs mapping has been conducted in many species and in different tissues.<sup>3</sup> Thus far, four genome-wide eQTLs studies in human liver tissue have been performed,<sup>4–7</sup> and numerous eQTLs have been identified. However, no studies have been carried out in an East Asian population, one of the major ethnicities in the world. Detailed eQTLs mapping in different populations is crucial to understand the genetic heterogeneity in gene regulation, and the evolutionary predisposition to diseases. Notably, the hepatic metabolising capacity of Caucasian populations has already been shown to be different from that of Asians, as exemplified in metabolism variability of alcohol, testosterone, bilirubin, etc.<sup>8–10</sup> Hepatic eQTLs studies in East Asians will be also crucial for understanding the genetic basis underlying various diseases and drug response variability, particularly in the East Asian population. For this reason, we have carried out a genome-wide eQTLs mapping in 64 normal livers of Han Chinese. Detailed comparison between the Asian and Caucasian eQTLs was conducted. Experimental validation in an independent sample set (n=54) was also performed. The relationship between Asian eQTLs and trait-associated single nucleotide polymorphisms (SNPs) as well as pharmacogene expression was investigated.

## MATERIALS AND METHODS

### Tissue sample collection

Normal (non-diseased) liver tissues were previously collected from 64 Chinese donors (all male) who provided informed consent. The average age of the subjects was  $34.52\pm 5.98$  years. The independent sample set of liver tissue (n=54) (non-diseased healthy male donors, aged  $35.65\pm 7.34$  years) were newly collected in Shanghai Jiao Tong University Affiliated First People's Hospital (Shanghai, China), Zhangzhou Hospital Affiliated to Fujian Medical University (Zhangzhou, China), and Shandong University Affiliated Qianfoshan Hospital (Jinan, China). This study was approved by the ethics committees of the medical faculties of Shanghai Jiao Tong University Affiliated First People's Hospital.

### RNA sample preparation, hybridisation

Total RNA of the human liver tissue samples was extracted and 1.65  $\mu$ g of each RNA sample was

labelled and hybridised to the Agilent 44 K G4112F arrays (GPL4133) at Shanghai genomePilot Technology, Inc (Shanghai, China).

### Gene expression microarray data preprocessing and normalisation

The dye-normalised and post-surrogate processed signal for the green channel, `gProcessedSignal`, obtained from Agilent's Feature Extraction Software was used for downstream analyses. The raw expression data for the 64 samples were evaluated for individual array quality (MA plots), array intensity distributions (box plots and density plots), and between-array differences (heat maps representing the distance between arrays) using the `arrayQualityMetrics` package. One outlier sample was dropped based on the `arrayQualityMetrics` default criteria.<sup>11</sup> The quantile normalisation method was used to normalise the data,<sup>12</sup> and the average values were obtained for the replicate spots. The expression intensity values were log<sub>2</sub> transformed. The processes were implemented using Limma. After data collection, we ran a pipeline to re-annotate the Agilent G4112F microarray for the current reference assembly (NCBI Build 37.3). A total of 29 190 oligonucleotides on the array were validated for subsequent analyses.

### DNA extraction, GWAS genotyping, and quality control

Total DNA of the human liver tissue samples were extracted by using the RNA/DNA Mini Kit (Qiagen, Hilden, Germany). DNA was diluted to working concentrations of 50 ng/μL for SNP chip genotyping. The genome-wide scan was performed using the Affymetrix Genome-Wide Human SNP Array 6.0. Quality control (QC) filtering of the GWAS data was performed by excluding arrays with a contrast QC < 0.4 from further data analysis. The sex of each sample was determined using Genotyping Console, and none of them mismatched established and annotated sexes. Genotype data were generated using the birdseed algorithm.<sup>13</sup> SNPs were further filtered based on annotation, call rate, Hardy-Weinberg equilibrium, and allele frequency information. As a result, among the initially genotyped 909 622 SNPs on the Affymetrix Genome-Wide Human SNP Array 6.0 platform, 4023 duplicated SNPs and 1175 SNPs that did not have chromosomal annotation were first removed from further analysis. We then removed 54 656 SNPs that had a genotyping call rate < 95%, and 737 SNPs that deviated significantly from Hardy-Weinberg equilibrium at the threshold ( $p < 10^{-4}$ ). Considering the small sample size of our dataset, we removed a further 54 656 SNPs that had minor allele frequency (MAF) < 25%. In total, 302 483 SNPs remained for further analysis.

### Quantification of ancestry and sample independence test

All the individuals are of self-reported Han Chinese ancestry. The multidimensional scaling (MDS) analysis was used to calculate the genome-wide IBS (identify-by-state) pairwise distances in PLINK V.1.07. All the samples resided in a single cluster, and no outlier was detected; thus all samples were used for analyses. Sample independence was also examined by using the IBD analysis in PLINK. The results showed that all individuals were independent of each other. No sample contamination, duplication, and significant family relationship were identified.

### eQTLs mapping

We tested all expression traits for their associations with each of the QC passed SNPs using PLINK, which correlates allele dosage with changes in the trait. Only the SNPs within  $\pm 2$  megabase (Mb) of the transcription start or stop of the

corresponding gene were tested for putative *cis*-eQTL. All the rest of the SNPs were tested for association to each expression trait for *trans*-eQTLs. To correct for the number of tests performed, a false discovery rate (FDR) of 0.05 was used as a cut-off for statistical significance.

There were incomplete records for demographic and clinical information for these samples. In order to assess the effect of hidden cofactors on eQTLs mapping, we performed corrections for gene expression variance using PEER (probabilistic estimation of expression residuals).<sup>14</sup> Interestingly, the majority of significant *cis*-eQTLs disappeared when using the PEER corrected residuals for gene expression, suggesting that covariates actually had a minimum effect on the gene expression variance in our data.

To assess the statistical power to detect eQTLs, a power analysis was performed using the GWAPower program.<sup>15</sup> We found that using our 63 samples (one sample was removed after data cleanup), with a validation *p* value of  $10^{-4}$  (a threshold at which we expect to be able to follow-up with at least *in silico* replication studies), we have about 80% power to detect an eQTLs accounting for 25% of the phenotypic variation in expression levels.

### Comparison of eQTLs between Asian and Caucasian populations

To check the effect of ethnicity on eQTLs mapping, we compared the results with that of a Caucasian dataset (GEO accession: GSE26106) by Innocenti *et al.*,<sup>5</sup> which was conducted with the same gene expression microarray platform (GEO#GPL4133). The GSE26105 dataset has 205 Caucasian liver samples which have both genotype and gene expression data. The same preprocessing and eQTLs mapping methods used in the Asian dataset were applied to the raw data of the gene expression and genotype downloaded from GEO. We focused on 64 964 SNPs and 24 340 gene probes which passed the QC in both datasets and have the genomic loci annotations.

To assess the eQTLs replication among the two populations more completely, we also performed a post-hoc genome-wide imputation to infer genotypic information for SNPs that had not been genotyped in our platform using the IMPUTE2 program<sup>16</sup> after prephasing the genotypes with SHAPEIT.<sup>17</sup> The 1000 genome phase I data (NCBI build 37) was used as a reference panel and default parameters were used in prephasing and imputation. After imputation, we obtained imputed data for 37 574 750 and 38 049 377 SNPs in the Han Chinese and Caucasian populations, respectively. The imputed data were further filtered based on the imputation quality (>30%) and MAF (>25% for Chinese and >5% for Caucasians), after which information for 2 624 722 and 6 763 377 SNPs remained for Chinese and Caucasian datasets, respectively, for further analysis.

### Statistical analysis for enrichment tests and population divergence comparison

Trait-associated SNPs were obtained from the NIGHR catalogue (<http://www.genome.gov/26525384>). We downloaded all catalogued SNPs associated with human traits with genome-wide significance ( $10^{-8}$ ) ( $n=6437$ ), among which we defined the Asian GWAS-SNPs ( $n=748$ ) as the subset of SNPs identified in East Asian populations, mainly Chinese and Japanese populations. The very important pharmacogenetic (VIP) genes ( $n=49$ ) were obtained from the Pharmacogenomics Knowledgebase (<http://www.PharmGKB.org>). The list of major pharmacokinetic genes ( $n=409$ ) was obtained from a previous study.<sup>18</sup>

The enrichment significance of one dataset in another was calculated by the  $\chi^2$  test, with all SNPs or genes from the Asian dataset as a background. A value of  $p < 0.05$  was considered significant. The binomial test was used for evaluating the enrichment of correlation with multiple gene probes in eQTLs hotspots.

To assess the influence of allele frequency in eQTLs replication between populations, we compared the distribution of  $F_{ST}$  (fixation index) values for the significant eQTLs between overlapped and non-overlapped groups, using the  $\chi^2$  test. An  $F_{ST}$  value of 0.5 was used as a cut-off to compare the difference in the number of SNPs with or without overlap between the two populations.  $F_{ST}$  values were obtained from the FstSNP-HapMap3 database.<sup>19</sup>

### Experimental confirmation of the gene expression and SNP gene expression association

The mRNA levels of seven genes (*DDT*, *ERAP2*, *MRPL43*, *FADS1*, *BRCA1*, *CCND2*, and *PTPRE*) were quantified in the livers using quantitative PCR (qPCR). Primers sequences are provided in online supplementary table S5. SNPs that were significantly associated with these gene expression profiles were also genotyped using PCR sequencing. Primer sequences are also listed in online supplementary table S5. The relationships between qPCR data and microarray data or SNP genotypes were determined using linear regression, with the significance cut-off set as 0.05.

## RESULTS

### Gene expression profiling and SNP genotyping

We conducted a genome-wide eQTLs mapping in a set of Han Chinese liver tissue samples ( $n=63$ ). SNPs were genotyped using Affymetrix SNP 6.0 chip, and genome-wide gene expression levels were profiled using Agilent G4112F array. After QC, 302 483 SNPs remained for analysis. Microarray expression probes were re-annotated using a previous reported pipeline.<sup>9</sup> A total of 29 190 probes were considered to be valid for subsequent analyses.

### eQTLs mapping

At the 5% FDR level ( $p < 9.45 \times 10^{-9}$ ), we identified a total of 1669 eQTLs with 1322 SNPs significantly associated with the expression of 282 genes. Among these eQTLs, 1465 were classified as *cis*-eQTLs including 1198 SNPs for 217 genes, and 204 *trans*-eQTLs, with 178 SNPs significantly associated with 68 genes. The full list of association results at a study-wide significance level are provided in online supplementary table S1. Figure 1 shows the distribution of *cis*- and *trans*-eQTLs at the  $p < 10^{-5}$  level, eQTLs hotspots, as well as the location of genes in the entire genome.

We also studied the physical distance distribution (base pairs) of the most significant eQTLs for each gene to the gene's transcription start site (TSS). We found that the most significant eQTLs were enriched around TSS (figure 2), which is also consistent with previous reports.<sup>5 20</sup>

### Comparison between Caucasian and Asian eQTLs

We set out to compare the eQTLs in the Han Chinese population with the previously published data in an American population.<sup>5</sup> Given the different platforms used in the two studies, we focused our analyses only on the common SNPs ( $n=64\ 964$ ) and gene probes (27 340) between the two datasets. We found that at the general genome-wide significance level ( $p < 10^{-8}$ ), 41% (113 out of 277) of Asian SNP-gene association pairs were also significant pairs in the Caucasian population (table 1). To confirm whether

there was a significant overlap in eQTLs between the two populations, we used a liberal  $p$  value cut-off of  $10^{-4}$  for both populations, and found that the eQTLs in the Asian population was also significantly more likely to be an eQTL in Caucasians ( $p < 2.2 \times 10^{-16}$ ). This enrichment remained to be significant for both *cis*- and *trans*-eQTLs ( $p < 2.8 \times 10^{-5}$  for both) (see online supplementary table S2).

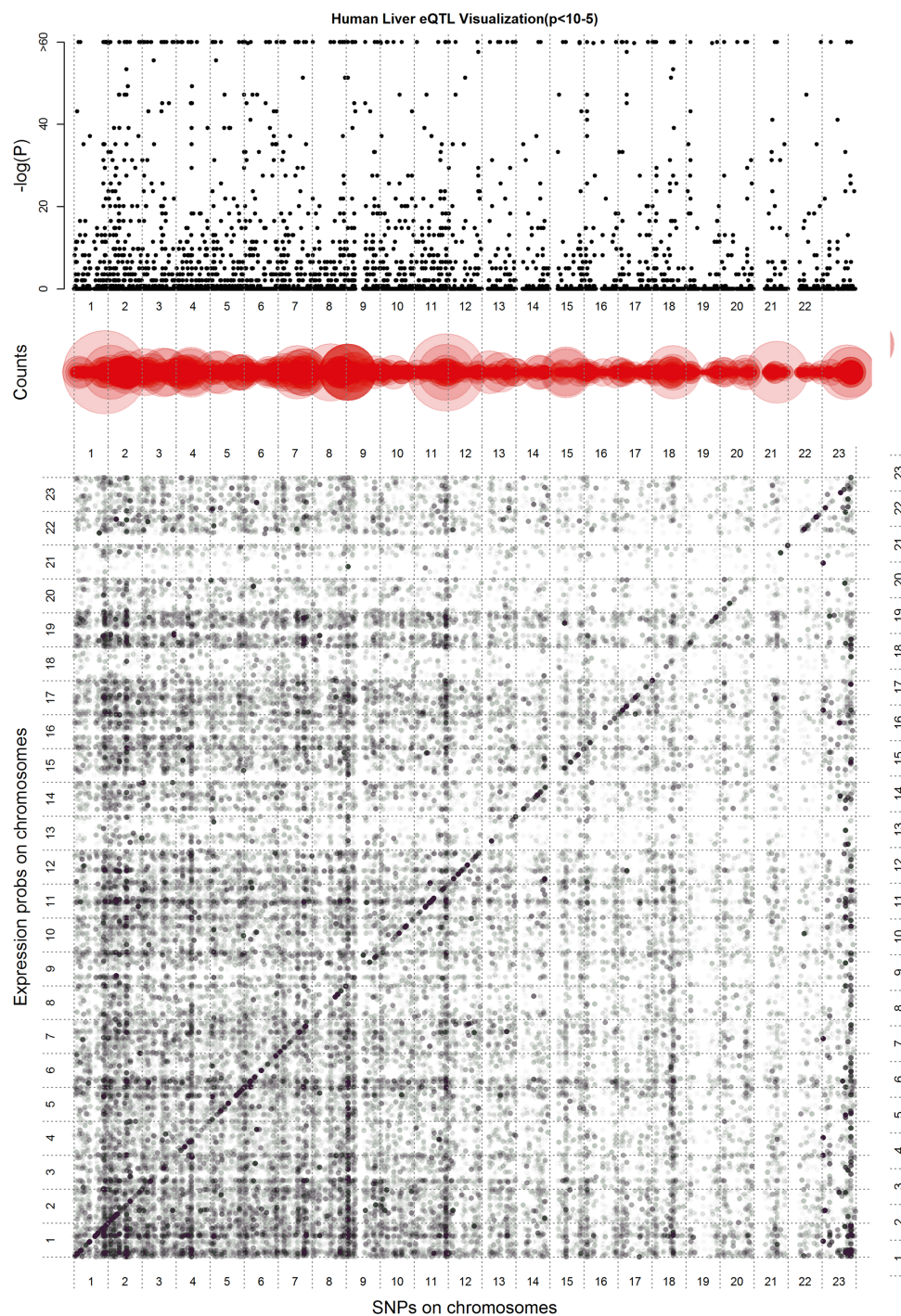
To further evaluate the eQTL overlapping between the two populations, we performed a genome-wide post-hoc imputation analysis to infer genotypes for the SNPs that had not been genotyped in both sample sets. After imputation, 19 703 SNPs were found to be significantly ( $p < 9.45 \times 10^{-9}$ ) associated with gene expression in the Han Chinese dataset, which included 18 186 *cis*-eQTLs and 1517 *trans*-eQTLs. After comparison with the Caucasian dataset, we found that 11 841 (60%) SNPs were also significantly associated with expression of the same genes at  $10^{-4}$  level in the Caucasian dataset, which included 11 284 (62%) *cis*-eQTLs and 557 (36.7%) *trans*-eQTLs.

Given the allele frequency difference for the SNPs between the two populations, it is possible that population divergence may have affected the identification and confirmation of eQTLs in different populations. To test this hypothesis, we calculated the  $F_{ST}$  (fixation index) value for the significant eQTLs overlapped and non-overlapped between the two populations. Interestingly, using an  $F_{ST}$  value of 0.5 as a cut-off, we found that about 4% of the overlapped eQTLs had an  $F_{ST} > 0.5$  compared to about 7% in non-overlapped eQTLs. Despite a small number, this comparison was statistically significant ( $p = 7.6 \times 10^{-4}$ ), suggesting that the SNPs with less population divergence were more likely to be eQTLs shared by the two populations (data not shown).

### Enrichment of trait-associated GWAS SNPs in Asian eQTLs

To test the hypothesis that the GWAS-identified SNPs significantly associated with human traits are also more likely to be eQTLs,<sup>21</sup> we checked the enrichment of trait-associated GWAS SNPs in Asian eQTLs. We first focused our analysis on the genome-wide significant ( $p \leq 10^{-8}$ ) SNPs associated with any trait in all populations deposited in the National Human Genome Research Institute (NHGRI) database ( $n=6437$ ). Among 1322 SNPs significantly ( $FDR < 0.05$ ) associated with gene expression in Han Chinese livers, 17 were also trait-associated SNPs (see online supplementary table S1), which was significantly enriched compared to the SNPs that were not significant eQTLs ( $FDR > 0.05$ ) ( $p < 1.4 \times 10^{-8}$ ) (see online supplementary table S3). We further divided the GWAS-SNPs into Asian (Chinese and Japanese population only) ( $n=748$ ) and European ( $n=5689$ ) populations. Out of these 17 SNPs, four SNPs were also trait-associated SNPs in the Asian population, which was significantly enriched in the Han Chinese eQTLs ( $p=0.0009$ ) (see online supplementary table S3). Examples here included rs12506899 located in the  $\alpha$ -fetoprotein gene (*AFP*), that was significantly associated with the cancer antigen 19-9 (*CA19-9*) in a GWAS conducted in a Han Chinese population.<sup>22</sup> In our data this SNP was significantly associated with *AFP* gene expression ( $p=1.42 \times 10^{-10}$ ). More interestingly, rs3077 located in the *HLA-DPA1* gene—which exhibited a relatively low allele frequency ( $MAF=0.11$ ) in Caucasians but was reported to be more common in Asians with an allele frequency of 0.62—was significantly associated with increased risk for hepatitis B virus (HBV) infection in an Asian population in a recent GWAS.<sup>23</sup> Similarly, we also noticed rs9277378—a significant *HLA-DPB1* eQTL which is in complete linkage disequilibrium (LD) with rs9277535 from





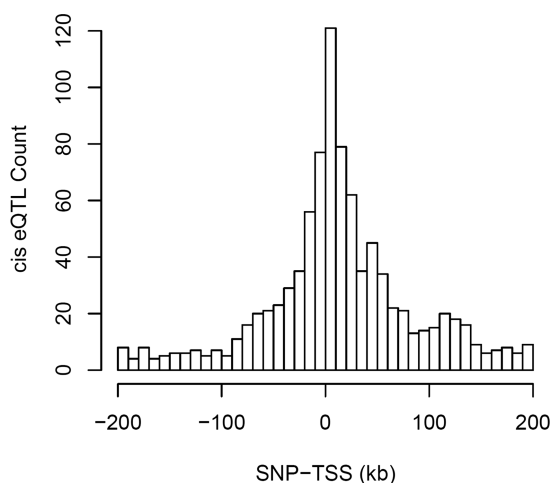
**Figure 1** The combined plot depicting all eQTLs for hepatic gene expression traits with  $p < 10^{-5}$  in our study. The bottom plot shows the genome-wide distribution of eQTLs results, with the SNP distribution on the x axis and expression probes on the y axis. Each dot represents a significant SNP-expression pair. Cis-eQTLs associations are shown in a diagonal direction and trans-eQTLs are shown in a vertical direction. Darker colour indicates more significant association. The middle plot shows hotspot eQTLs enrichment in correlation with expression probes, indicating the eQTLs as possible key regulators affecting expression of multiple genes. The size of the red circles indicate the number of gene expression traits correlated with the particular SNP. The top plot is the enrichment scale ( $-\log p$  value), which shows the enrichment of correlation with multiple genes ( $-\log p$  value from binomial test after Bonferroni correction). eQTL, expression quantitative trait loci; SNP, single nucleotide polymorphism.

HapMap Asian data—was significantly associated with HBV infection in Asian populations in two previously published GWAS.<sup>24 25</sup> On the other hand, 15 out of the aforementioned 19 SNPs were GWAS SNPs in European populations, and were also significantly enriched in the Han Chinese eQTLs ( $p < 0.0001$ ) (see online supplementary table S1).

#### Association between Asian eQTLs and expression variability of pharmacogenes

The liver is the most important organ for drug metabolism. We hypothesised that genetic polymorphisms can explain inter-individual variability in pharmacogene expression, which would be of importance to pharmacogenetics. To confirm this, we tested





**Figure 2** Histogram showing distances from each gene's best associated SNP to its TSS. Negative and positive values denote SNPs 5' and 3' of TSS, respectively. The unit interval was set as 10 KB. eQTL, expression quantitative trait loci; SNP, single nucleotide polymorphism; TSS, transcription start site.

the overlap between Asian eQTLs and the VIP genes ( $n=49$ ), a list of genes drawing most attention in the pharmacogenetic research area and recently identified by the Pharmacogenomics Knowledgebase (<http://www.PharmGKB.org>). As a result, 44 genes were profiled in our platform. Four genes, *BRCA1*, *CYP2D6*, *CYP3A5*, and *GSTT1*, were significantly ( $FDR < 0.05$ ) associated with at least one eQTLs in our dataset (see online supplementary table S1). Compared to the number of genes that were not significant eQTLs, the expression of VIP genes were significantly more likely to be controlled by eQTLs ( $p=0.002$ ) (see online supplementary table S4). To expand this analysis to other important pharmacogenes, we tested the association between eQTLs and 409 genes encoding major phase I/phase II drug metabolism enzymes, transporters, as well as nuclear factors regulating pharmacogene expression.<sup>17</sup> Among these 409 genes, 341 were profiled in our study, and 17 genes (see online supplementary table S1) were found to be significantly associated with at least one eQTL. Similar to the VIP genes, this was also a significant enrichment ( $p=2.8 \times 10^{-8}$ ) (see online supplementary table S4). Notably, a number of glutathione-s-transferase genes (*GST*) including *GSTA4*, *GSTM1*, *GSTM2*, *GSTM2P1*, *GSTM4*, *GSTM5*, *GSTT1*, and *GSTT2* were found to be significantly associated with at least one eQTLs (see online supplementary table S1).

#### Experimental validation of the gene expression and SNP gene expression associations

In order to confirm the findings using independent techniques, we quantified mRNA levels of seven randomly selected genes that were significantly associated eQTLs, using qPCR. Among

these seven genes, five (*DDT*, *ERAP2*, *MRPL43*, *FADS1*, and *BRCA1*) were significantly associated with *cis*-SNPs, and two (*CCND2* and *PTPRE*) were associated with *trans*-SNPs. The qPCR measurements were then correlated with gene expression profiles in the microarray. We found that qPCR measurements of all seven genes were significantly correlated with their microarray profiles ( $p < 5 \times 10^{-4}$  for all). To confirm the SNP genotypes determined by the DNA chip, we also performed Sanger sequencing for the seven SNPs significantly associated with these genes' expression. We found that the genotype of one SNP (rs1006771) had 100% concordance with the DNA chip data, while the remaining six SNPs all had 98% concordance, with discrepancy in the genotype for one sample for each SNP between sequencing and DNA chip results. Note that this discrepancy randomly occurred in different samples. Meanwhile, the qPCR measurements of six genes (*DDT*, *ERAP2*, *MRPL43*, *FADS1*, *BRCA1*, and *PTPRE*) were also significantly associated with the originally identified SNPs ( $p < 0.03$  for all), while no association between the qPCR level of *CCND2* and the originally identified *trans*-SNPs was observed ( $p > 0.18$  for all) (table 2).

To further validate the reliability of the eQTLs mapping in our study, we also experimentally validated the aforementioned seven SNP-gene pairs in an independent liver tissue set ( $n=54$ ). Again, gene expression and SNP genotype were determined using qPCR and sequencing. We found that all five *cis*-eQTLs identified in the original sample set were also significantly associated with gene expression profiles in the new sample set ( $p < 0.008$  for all), while the two *trans*-eQTLs were not ( $p > 0.4$  for both).

#### DISCUSSION

Previous studies on eQTLs mapping in human liver have demonstrated its power for understanding inter-individual variability in disease aetiology and response to therapeutic treatments.<sup>5 21</sup> We have performed, for the first time, an eQTLs study in human livers of a Han Chinese population. This unique dataset may extend our understanding of the genetic basis underlying the variability in gene expression, aetiology of human diseases, and various traits in pharmacotherapy.

Although we have a relatively small sample size ( $n=64$ ), our study was able to replicate a large proportion of hepatic eQTLs identified in previous studies, suggesting the high quality of our dataset. This is also confirmed by our experimental validation. Without considering the SNPs in linkage disequilibrium, 41% of significant eQTLs in Asians were consistent with those found in Caucasians. However, by including imputed genotypic data, this number increased to 62%, further indicating that the regulatory mechanism underlying many SNP-gene expression correlations is actually common across the entire human population. By dividing the eQTLs into *cis* ( $\pm 2$  Mb) and *trans*, at  $10^{-4}$  level, about 30% of Asian *cis*-eQTLs were also Caucasian *cis*-eQTLs, while only 0.023% of Asian *trans*-eQTLs were consistent with Caucasian findings. This further highlighted the reliability of *cis*-eQTLs mapping as indicated in previous studies.<sup>5</sup> This inference was also confirmed by our experimental validation studies conducted in an independent sample cohort. However, in spite of the small proportion of overlapping *trans*-eQTLs between the two populations, this overlap still represented a statistically significant enrichment, suggesting that these *trans*-eQTLs are likely to be true signals.

eQTLs information was deemed to be important to establish the causal role of genetic variants and genes involved in human disease.<sup>4 5 26-28</sup> Our data further confirmed the previous findings that trait-associated GWAS-SNPs are more likely eQTLs.<sup>21</sup>

**Table 1** Overlap of eQTLs ( $p < 10^{-8}$ ) between Asian and Caucasian populations

Dataset	No. of significant SNP-gene pairs	No. of SNPs	No. of genes
Asian all	277	215	138
Caucasian all	684	486	435
Overlap	113	100	61

eQTLs, expression quantitative trait loci; SNPs, single nucleotide polymorphisms.

**Table 2** qPCR validation of gene expression and SNP-gene correlation

Gene	eQTLs association	Probe	SNPs	Correlation with microarray data		Microarray data with SNPs		qPCR data with SNPs		Independent sample set (n=54)	
				$\beta$	p Value	$\beta$	p Value	$\beta$	p Value	$\beta$	p Value
<i>DDT</i>	<i>cis</i>	A_23_P17769	rs1006771	0.98	1.90E-18	0.67	3.11E-25	0.45	6.97E-14	0.58	7.4E-03
<i>MRPL43</i>	<i>cis</i>	A_23_P382154	rs3824783	0.81	1.69E-07	-1.16	3.59E-24	-0.64	2.61E-02	-1.21	9.01E-09
<i>ERAP2</i>	<i>cis</i>	A_23_P30243	rs10434709	1.12	2.91E-31	2.35	8.41E-26	1.80	3.02E-20	-2.06	1.76E-09
<i>FADS1</i>	<i>cis</i>	A_24_P192994	rs174547	0.81	4.36E-13	-1.27	7.87E-09	-1.13	5.83E-07	-1.04	5.02E-06
<i>BRCA1</i>	<i>cis</i>	A_23_P207400	rs9911630	1.08	6.28E-09	-0.84	6.61E-11	-0.79	0.00135	-0.52	6.19E-03
<i>CCND2</i>	<i>trans</i>	A_24_P270235	rs1404608	0.71	4.77E-04	0.38	8.97E-11	0.12	3.80E-01	-0.13	0.53
<i>PTPRE</i>	<i>trans</i>	A_23_P138495	rs1463389	2.17	3.04E-06	0.34	3.03E-10	0.50	3.07E-02	-0.53	0.41

eQTLs, expression quantitative trait loci; SNPs, single nucleotide polymorphisms; qPCR, quantitative PCR.

More importantly, we hypothesised that the Asian eQTLs may be particularly useful for understanding the mechanism underlying genotype-phenotype correlations in Asian populations. Indeed, we found that GWAS-SNPs identified in Asian populations were significantly enriched in our eQTLs. One interesting example is the SNPs associated with HBV infection. HBV is endemic in East Asian population, and over 75% of the world's estimated 350 million carriers are located in Western Pacific and South East Asian countries.<sup>29</sup> The prevalence rate of HBV infection in Han Chinese is extremely high (up to 12%).<sup>29</sup> Previous GWAS identified two *HLA-DP* loci, rs3077 in *HLA-DPA1* and rs9277378 in *HLA-DPB1*, that were significantly associated with increased risk for HBV infection in Asian populations.<sup>23-25</sup> Our data revealed that these two SNPs were actually significant eQTLs for *HLA-DPA1* and *HLA-DPB1* gene expression. More interestingly, the allele frequencies of rs3077 and rs9277378 are significantly different between Caucasian, Asian, and African populations, with the rare alleles among Caucasians being common alleles in Asians and Africans according to the HapMap data (MAF for rs3077 is 0.11, 0.61, and 0.76, and for rs9277378 is 0.29, 0.54, 0.82 among the three populations, respectively). This may indicate that altered hepatic gene expression of these two genes among these populations confers differential risks for HBV infection. This is further supported by the epidemiological observation that both Asian and African populations have much higher HBV infection rates than Caucasians.<sup>29</sup>

We also observed a few SNPs (rs174547, rs174548, and rs174549) consistently identified from GWAS in both European and Asian populations were associated with multiple metabolic perturbations and lipid metabolism traits.<sup>30-32</sup> We found that these SNPs were significantly associated with gene expression of *FADS1*, one of the fatty acid desaturase genes. Our finding is further supported by studies in Caucasian populations.<sup>26</sup> Both our original and validation studies in additional samples confirmed the significant association between rs174547 and *FADS1* gene expression. This further indicates that mechanisms underlying expression regulation of many genes are actually shared by different populations, and consequently the associated diseases and traits may have a common natural history.

Besides diseases and trait-associated SNPs and genes, we also tested the association between Asian eQTLs and important pharmacogenes. As expected, four VIP genes and 17 major genes involved in pharmacokinetics were significantly associated with at least one eQTLs. This underscored the importance of our dataset in understanding the inter-individual difference in drug response and toxicities. We observed that hepatic expression levels of a group of *GST* genes were significantly affected by eQTLs. Expression of several important P450 genes including *CYP2D6*, *CYP3A5*, *CYP3A7*, and

*CYP4V2* were also found to be controlled by eQTLs. Although high impact polymorphisms from these genes have been identified in previous studies,<sup>33</sup> our study provided new candidate polymorphisms that may be important to pharmacogenetics in Asian populations. *CYP3A7*, the most important P450 gene in fetal liver,<sup>34</sup> was significantly associated with multiple SNPs; further investigations are therefore warranted to address the question of whether these polymorphisms confer susceptibility to the inter-patient differences in drug efficacy or toxicity in paediatric populations.

Replication of eQTLs results between populations were often observed to be highly variable in different studies, which could be attributed to many reasons.<sup>4</sup> We found in our study that differences in allele frequency can significantly affect eQTL replication in different populations. An SNP with higher allele frequency in one population may also have greater power in association with gene expression compared to a relatively lower allele frequency in another population. In addition, as we calculated in the power analysis, the small sample size in our study might also limit the power for detecting eQTLs with moderate effect, which further led to non-replicable eQTLs. Our sample set was also limited by the incomplete covariate information (demographic, clinical, etc) collected during the sample procurement process. Nevertheless, our future goal is to collect a larger sample set, as well as to perform multi-population meta-analyses to address these questions.

In conclusion, our first eQTLs analysis in the East Asian Han Chinese population revealed both homogeneity and heterogeneity in genetic variations in gene expression among different human populations. Many of our findings provided further supportive evidence for recent genomic discoveries in human diseases and pharmacogenetic traits, and more importantly fostered new rationales for continued investigation. Our data thus provide an additional valuable resource to the existing data in other populations.

#### Author affiliations

- <sup>1</sup>Department of General Surgery, Shanghai First People's Hospital, Medical College, Shanghai Jiaotong University, Shanghai, China
- <sup>2</sup>Department of Pathology, Shanghai First People's Hospital, Medical College, Shanghai Jiaotong University, Shanghai, China
- <sup>3</sup>Department of General Surgery, Shandong University Affiliated Qianfoshan Hospital, Jinan, China
- <sup>4</sup>Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Bio-X Institutes, Ministry of Education, Shanghai Jiao Tong University; Shanghai genome Pilot Institutes for Genomics and Human Health, Shanghai, China
- <sup>5</sup>Department of Hepatobiliary Surgery, Zhangzhou Hospital Affiliated to Fujian Medical University, Zhangzhou, China

**Acknowledgements** We are deeply grateful to all the participants as well as to the doctors working on this project. This work was supported by the National Natural Science Foundation of China (No. 81000188, 81270557).

**Contributors** ZHP supervised sample recruitment. XLW, HMT, MJT and ZQL conducted data analyses and drafted the manuscript. JWF, LZ, XS, JMX, GQC, DWC and ZWW recruited samples. THX, JYZ, LH, SYW, XP and SYQ performed or contributed to the main experiments. All authors critically reviewed the manuscript and approved the final version.

**Funding** The Natural Science Foundation of China.

**Competing interests** None.

**Patient consent** Obtained.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Additional data file** The full eQTLs dataset is accessible in an online data resource (<http://analysis2.bio-x.cn/SHEsisMain.htm>). The data were also deposited in the NCBI GEO database (accession numbers: GSE53792). The following additional data are available. Supplemental file 1 contains table S1, which provides a full list for association results at a study-wise significance level (FDR <0.05) for gene expression and genotyping data. Supplemental tables S2–S5 include enrichment analyses and primer sequences.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/3.0/>

## REFERENCES

- Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* 2005;102:1572–7.
- Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Gene* 2001;17:388–91.
- Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet* 2011;27:72–9.
- Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008;6. <http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.0060107> (accessed Jun 2013).
- Innocenti F, Cooper GM, Stanaway IB, Gamazon ER, Smith JD, Mirkov S, Ramirez J, Liu W, Lin YS, Moloney C. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet* 2011;7:e1002078. <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002078> (accessed Jun 2013).
- Schröder A, Klein K, Winter S, Schwab M, Bonin M, Zell A, Zanger UM. Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *Pharmacogenomics J* 2011;13:12–20.
- Greenawald DM, Dobrin R, Chudin E, Hatoum JJ, Suver C, Beaulaurier J, Zhang B, Castro V, Zhu J, Sieberts SK, Wang S, Molony C, Heymsfield SB, Kemp DM, Reitman ML, Lum PY, Schadt EE, Kaplan LM. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res* 2011;21:1008–16.
- Brooks PJ, Enoch MA, Goldman D, Li TK, Yokoyama A. The alcohol flushing response: an unrecognized risk factor for esophageal cancer from alcohol consumption. *PLoS Med* 2009;6:e50. <http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000050> (accessed June 2013).
- Jakobsson J, Ekström L, Inotsume N, Garle M, Lorentzon M, Ohlsson C, Roh HK, Carlström K, Rane A. Large differences in testosterone excretion in Korean and Swedish men are strongly associated with a UDP-glucuronosyl transferase 2B17 polymorphism. *J Clin Endocrinol Metab* 2006;91:687–93.
- Lampe JW, Bigler J, Horner NK, Potter JD. UDP-glucuronosyltransferase (UGT1A1\*28 and UGT1A6\*2) polymorphisms in Caucasians and Asians: relationships to serum bilirubin concentrations. *Pharmacogenetics* 1999;9:341–9.
- Kauffmann A, Gentleman R, Huber W. ArrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* 2009;25:415–16.
- Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003;19:185–93.
- Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 2008;40:1253–60.
- Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 2012;7:500–7.
- Feng S, Wang S, Chen CC, Lan L. GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits. *BMC Genet* 2011;12:12.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5:e1000529.
- Delaneau O, Zagury JF, Marchini J. Improved whole chromosome phasing for disease and population genetic studies. *Nat Methods* 2013;10:5–6.
- Wei R, Yang F, Urban TJ, Li L, Chalasani N, Flockhart DA, Liu W. Impact of the interaction between 3'-UTR SNPs and microRNA on the expression of human xenobiotic metabolism enzyme and transporter genes. *Front Genet* 2012;3:248.
- Duan S, Zhang W, Cox NJ, Dolan ME. FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformatics* 2008;3:139–41.
- Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008;4:e1000214. <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000214> (accessed June 2013).
- Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010;6:e1000888. <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000888> (accessed June 2013).
- He M, Wu C, Xu J, Guo H, Yang H, Zhang X, Sun J, Yu D, Zhou L, Peng T, He Y, Gao Y, Yuan J, Deng Q, Dai X, Tan A, Feng Y, Zhang H, Min X, Yang X, Zhu J, Zhai K, Chang J, Qin X, Tan W, Hu Y, Lang M, Tao S, Li Y, Li Y, Feng J, Li D, Kim ST, Zhang S, Zhang H, Zheng SL, Gui L, Wang Y, Wei S, Wang F, Fang W, Liang Y, Zhai Y, Chen W, Miao X, Zhou G, Hu FB, Lin D, Mo Z, Wu T. A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19–9, carcinoembryonic antigen and  $\alpha$  fetoprotein and their associations with cancer risk. *Gut* 2014;63:143–51. Published online first:7 January 2013.
- Nishida N, Sawai H, Matsuura K, Sugiyama M, Ahn SH, Park JY, Hige S, Kang JH, Suzuki K, Kurosaki M, Asahina Y, Mochida S, Watanabe M, Tanaka E, Honda M, Kaneko S, Orito E, Itoh Y, Mita E, Tamori A, Murawaki Y, Hiasa Y, Sakaida I, Korenaga M, Hino K, Ide T, Kawashima M, Mawatari Y, Sageshima M, Ogasawara Y, Koike A, Izumi N, Han KH, Tanaka Y, Tokunaga K, Mizokami M. Genome-wide association study confirming association of HLA-DP with protection against chronic hepatitis B and viral clearance in Japanese and Korean. *PLoS One* 2012;7:e39175. <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039175> (accessed Jun 2013)
- Kamatani Y, Watanapokayakit S, Ochi H, Kawaguchi T, Takahashi A, Hosono N, Kubo M, Tsunoda T, Kamatani N, Kumada H. A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* 2009;41:591–5.
- Mbarek H, Ochi H, Urabe Y, Kumar V, Kubo M, Hosono N, Takahashi A, Kamatani Y, Miki D, Abe H. A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. *Hum Mol Genet* 2011;20:3884–92.
- Mirkov S, Myers JL, Ramirez J, Liu W. SNPs affecting serum metabolic traits may regulate gene transcription and lipid accumulation in the liver. *Metabolism* 2012;61:1523–7.
- Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, Palmer CD, Gudnason V, Eiriksdottir G, Garcia ME, Launer LJ, Nalls MA, Clark JM, Mitchell BD, Shuldiner AR, Butler JL, Tomas M, Hoffmann U, Hwang SJ, Massaro JM, O'Donnell CJ, Sahani DV, Salomaa V, Schadt EE, Schwartz SM, Siscovick DS, NASH CRN, GIANT Consortium, MAGIC Investigators, Voight BF, Carr JJ, Feitosa MF, Harris TB, Fox CS, Smith AV, Kao WH, Hirschhorn JN, Borecki IB, GOLD Consortium. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet* 2011;7. <http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1001324> (accessed June 2013).
- Chambers JC, Zhang W, Sehmi J, Li X, Wass MN, Van der Harst P, Holm H, Sanna S, Kavousi M, Baumeister SE, Coin LJ, Deng G, Gieger C, Heard-Costa NL, Hottenga JJ, Kühnel B, Kumar V, Lagou V, Liang L, Luan J, Vidal PM, Mateo Leach I, O'Reilly PF, Peden JF, Rahmioglu N, Soininen P, Speliotes EK, Yuan X, Thorleifsson G, Alizadeh BZ, Atwood LD, Borecki IB, Brown MJ, Charoen P, Cucca F, Das D, de Geus EJ, Dixon AL, Döring A, Ehret G, Eyjolfsson GI, Farrall M, Frouhi NG, Friedrich N, Goessling W, Gudbjartsson DF, Harris TB, Hartikainen AL, Heath S, Hirschfeld GM, Hofman A, Homuth G, Hyppönen E, Janssen HL, Johnson T, Kangas AJ, Kema IP, Kühn JP, Lai S, Lathrop M, Lerch MM, Li Y, Liang TJ, Lin JP, Loos RJ, Martin NG, Moffatt MF, Montgomery GW, Munroe PB, Musunuru K, Nakamura J, O'Donnell CJ, Olafsson I, Penninx BW, Pouta A, Prins BP, Prokopenko I, Puls R, Ruokonen A, Savolainen MJ, Schlessinger D, Schouten JN, Seedorf U, Sen-Chowdhry S, Siminovich KA, Smit JH, Spector TD, Tan W, Teslovich TM, Tukiainen T, Uitterlinden AG, Van der Klauw MM, Vasari RS, Wallace C, Wallaschofski H, Wichmann HE, Willemsen G, Würtz P, Xu C, Yerges-Armstrong LM, Abecasis GR, Ahmadi KR, Boomsma DI, Caulfield M, Cookson WO, van Duijn CM, Froguel P, Matsuda K, McCarthy MI, Meisinger C, Mooser V, Pietiläinen KH, Schumann G, Snieder N, Sternberg MJ, Stolk RP, Thomas HC, Thorsteinsdottir U, Uda M, Waeber G, Wareham NJ, Waterworth DM, Watkins H, Whitfield JB, Witteman JC, Wolfenbuttel BH, Fox CS, Ala-Korpela M, Stefansson K, Vollenweider P, Völzke H, Schadt EE, Scott J, Järvelin MR, Elliott P, Kooper JS. Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* 2011;43:1131–8.
- Feng ID. Epidemiology of hepatitis B infection in the Western Pacific and South East Asia. *Gut* 1996;38(Suppl 2):18–23.



- 30 Han Y, Pei Y, Liu Y, Zhang L, Wu S, Tian Q, Chen X, Shen H, Zhu X, Papasian CJ, Deng H. Bivariate genome-wide association study suggests fatty acid desaturase genes and cadherin DCHS2 for variation of both compressive strength index and appendicular lean mass in males. *Bone* 2012;51:1000–7.
- 31 Hong KW, Jin HS, Song D, Kwak HK, Soo Kim S, Kim Y. Genome-wide association study of serum albumin:globulin ratio in Korean populations. *J Hum Genet* 2013; 58:174–7.
- 32 Hong MG, Karlsson R, Magnusson PK, Lewis MR, Isaacs W, Zheng LS, Xu J, Grönberg H, Ingelsson E, Pawitan Y, Broeckling C, Prenti JE, Wiklund F, Prince JA. A genome-wide assessment of variability in human serum metabolism. *Hum Mutat* 2013;34:515–24.
- 33 Hu L, Zhuo W, He YJ, Zhou HH, Fan L. Pharmacogenetics of P450 oxidoreductase: implications in drug metabolism and therapy. *Pharmacogenet Genom* 2012;22:812–19.
- 34 Hines RN. Ontogeny of human hepatic cytochromes P450. *J Biochem Mol Toxicol* 2007;21:169–75.