Bioinformatics, 34(18), 2018, 3145–3150
doi: 10.1093/bioinformatics/bty318
Advance Access Publication Date: 26 April 2018
Original Paper

Gene expression

A powerful approach reveals numerous expression quantitative trait haplotypes in multiple tissues

Dingge Ying[1], Mulin Jun Li[1,4], Pak Chung Sham[1] and Miaoxin Li[1,2,3,*]

[1]Department of Psychiatry, The Centre for Genomic Sciences, State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China, [2]Zhongshan School of Medicine, Center for Disease Genomics, Sun Yat-Sen University, Guangzhou 510080, China, [3]Key Laboratory of Tropical Disease Control (SYSU), Ministry of Education, Guangzhou 510080, China and [4]Department of Pharmacology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 300070, China

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on November 1, 2017; revised on March 19, 2018; editorial decision on April 18, 2018; accepted on April 25, 2018

Abstract

Motivation: Recently many studies showed single nucleotide polymorphisms (SNPs) affect gene expression and contribute to development of complex traits/diseases in a tissue context-dependent manner. However, little is known about haplotype's influence on gene expression and complex traits, which reflects the interaction effect between SNPs.

Results: In the present study, we firstly proposed a regulatory region guided eQTL haplotype association analysis approach, and then systematically investigate the expression quantitative trait loci (eQTL) haplotypes in 20 different tissues by the approach. The approach has a powerful design of reducing computational burden by the utilization of regulatory predictions for candidate SNP selection and multiple testing corrections on non-independent haplotypes. The application results in multiple tissues showed that haplotype-based eQTLs not only increased the number of eQTL genes in a tissue specific manner, but were also enriched in loci that associated with complex traits in a tissue-matched manner. In addition, we found that tag SNPs of eQTL haplotypes from whole blood were selectively enriched in certain combination of regulatory elements (e.g. promoters and enhancers) according to predicted chromatin states. In summary, this eQTL haplotype detection approach, together with the application results, shed insights into synergistic effect of sequence variants on gene expression and their susceptibility to complex diseases.

Availability and implementation: The executable application 'eHaplo' is implemented in Java and is publicly available at http://grass.cgs.hku.hk/limx/ehaplo/.

Contact: jonsonfox@gmail.com, limiaoxin@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The development of high throughput technologies has stimulated comprehensive surveys on genome-wide gene expression and DNA variation for disentangling the genetic architecture of human diseases. The genetics of transcript abundance has been extensively investigated through genome-wide gene expression studies (Ahuja *et al.*, 2016; Edwards *et al.*, 2013). These studies demonstrated that, for a large fraction of genes, gene expression is influenced by single nucleotide polymorphisms (SNPs) located in the vicinity of the regulated loci, named as expression quantitative trait loci (eQTLs), generally referred as *cis* eSNPs (Garnier *et al.*, 2013). The importance of *cis* eSNPs would be enhanced if they were also associated with a disease, as such data would indicate that the associated gene is a candidate for the disease (Nica and Dermitzakis, 2008). Recent eQTL studies have extended the focus from SNPs to other type of variations, including bi-allelic indels, copy number variations (CNVs) and short tandem repeats as determinants of gene expression (Encode Project Consortium, 2012; Grundberg *et al.*, 2012; Gymrek *et al.*, 2016; Lappalainen *et al.*, 2013; Montgomery *et al.*, 2013; Stranger *et al.*, 2007). Meanwhile, many eQTL studies showed significant contribution of tissue specific eQTLs to common disease heritability (GTex Consortium, 2015; Torres *et al.*, 2014). An eQTL study between blood and brain also found some of the tissue specific eQTLs were associated with related traits (Hernandez *et al.*, 2012). These studies showed the promise of tissue specific eQTLs for the characterizing functional sequencing variation and for interpreting statistic associations of genome-wide association studies (GWAS).

Haplotype, which refers to certain combination of multiple SNP alleles, is often used to explore synergistic or non-additive effects among multiple SNPs. Although methods based on individual SNPs have led to many significant findings in GWAS, haplotype-based methods will be an complementary way to explore extra genetic factors contributing to a disease (Liu *et al.*, 2008). Many GWAS and region-specific association studies have shown the power of haplotype by increasing the amount of explained disease risks and identifying additional disease susceptibility genes (Khankhanian *et al.*, 2015; Solovieff *et al.*, 2014). However, few studies have extended the application of haplotype into eQTL analysis. A gene expression study on HapMap data showed that samples with certain haplotypes tagged by four SNPs located in two enhancers had significantly higher gene expression, while this effect was vanished in single SNP analysis (Corradin *et al.*, 2014).

On the other hand, *cis*-regulatory sequences, such as enhancers and promoters, control development and physiology by regulating gene expression (Wittkopp and Kalay, 2012). Recently, progress has been made on predicting regulatory potential at non-coding sequencing variants in high throughput sequencing studies (Li *et al.*, 2016). Li et al. adopted tissue and cell-type specific epigenomic data to score regulatory variants; and found that the regulatory scores when used as weights substantially improved power of gene-based association analysis (Li *et al.*, 2017). Therefore, prior regulatory predictions may be a valuable resource for effectively selecting functional variants to reduce the computational burden in haplotype analysis.

In this study, we proposed a regulatory region guided approach for detecting eQTL haplotypes on whole genome level. In addition, we utilized this approach to comprehensively explore genome-wide cis-eQTL haplotypes in 20 tissues with the genotype and expression data from Genotype-Tissue Expression Project (GTEx; GTex Consortium, 2015). Furthermore, we also examined their enrichment of regulatory functional elements derived by chromatin states, and investigated their association with complex diseases and traits in a tissue specific manner.

2 Materials and methods

2.1 The proposed regulatory region guided eQTL haplotype analysis method

Here, we propose a regulatory region guided eQTL haplotype analysis method. This approach is made up of three steps. Firstly, the regulatory scores on non-coding variants are utilized to predict regulatory regions. Secondly, common haplotypes formed by SNPs that located in the predicted regulatory regions are explored. Finally, the association of the haplotypes with gene expression is examined and the significance is corrected by an estimation of 'effective number' of tests performed in each gene. The public data used and the key steps of the proposed approach are illustrated in Figure 1.

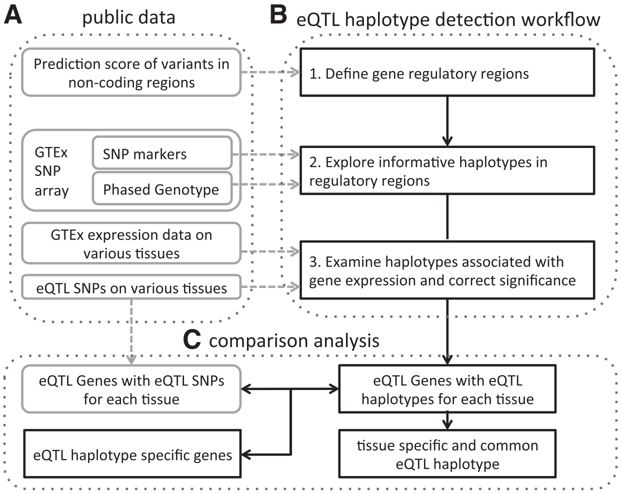
A
public data
Prediction score of variants in non-coding regions
GTEx SNP array
SNP markers
Phased Genotype
GTEx expression data on various tissues
eQTL SNPs on various tissues
B
eQTL haplotype detection workflow
1. Define gene regulatory regions
2. Explore informative haplotypes in regulatory regions
3. Examine haplotypes associated with gene expression and correct significance
C
comparison analysis
eQTL Genes with eQTL SNPs for each tissue
eQTL Genes with eQTL haplotypes for each tissue
eQTL haplotype specific genes
tissue specific and common eQTL haplotype

Fig. 1. Overview of the proposed eQTL haplotype detection approach. (**A**) The public data used for the eQTL haplotype detection workflow. (**B**) The workflow of the proposed approach in Section 2.1. (**C**) The comparison analysis performed of the eQTL haplotypes as described in Section 2.2

2.1.1 Define gene regulatory regions according to prediction score of variants in non-coding regions

To select potential variants that may form eQTL haplotypes, the first step is to define regulatory regions that may harbor such variants. Sequence variants are assigned with regulatory prediction scores by a powerful integrative approach (Li *et al.*, 2016). Briefly, the score was calculated using a composite strategy that utilized the predictions from eight different tools on functional annotation of non-coding variants. The tool takes advantage of the complementary attributes of individual tools to achieve a better performance. Supplementary Figure S1 shows the regulatory scores of a 200 kb region in chromosome 1. Variants with relatively higher scores form clusters indicating regulatory regions. Meanwhile the variation of the scores in surrounding variants is quite large, which makes it unsuitable to define clusters based on the score directly. To better identify such clusters, we use a smoothing method to utilize the variants with high scores. By doing so, firstly, the score of each variant is smoothed and replaced by the average scores in surrounding 10 variants (M_{var}). Secondly, the segmental average score (M_{seg}) and the SD ($\mathrm{SD}_{\mathrm{seg}}$) are calculated for each 10 000 consecutive variants, with 1000 variants as segment shifting size (the selection of smoothing parameters were explained in the Supplementary Result S8). Thirdly, variants are considered to be located in critical regulatory regions if its' smoothed regulatory score higher than segmental average score by two segmental $\mathrm{SD}_{\mathrm{seg}}$.

$$M_{\mathrm{var}} > M_{\mathrm{seg}} + \mathrm{SD}_{\mathrm{seg}} * 2 \tag{1}$$

As a result, every critical regulatory region consists of as a series of variants (at least two variants) meeting the above criteria. Regions that overlapping with centromere are spitted into two by removing the centromere region.

2.1.2 Explore informative haplotypes in regulatory regions

The second step is to explore informative haplotypes in the above regulatory regions. All coding genes in RefGene are selected for analysis. For each gene, a variant is considered as potential tag SNPs for eQTL haplotypes of the gene if it is: (i) located in critical regulatory regions defined above, and (ii) within 1 Mb upstream or downstream from the gene coding regions and (iii) with minor allele frequency higher than 0.05.

The initial step of haplotype searching treats the two alleles of the first SNPs (a_1, A_1) in upstream (according to co-ordinates) the gene as the first two haplotypes (h_1, h_2).

$$h_1 = a_1;\ h_2 = a_2$$

1 SNP $$H_w = (h_1; h_2)$$

These two haplotypes are added into a working haplotype group H_w, which contains all the haplotypes that can be extended by more tag SNPs. The extension step of the haplotype searching is carried out by adding one consecutive SNP to all haplotypes in the working group. By doing so, each haplotype in the working group is extended by two alleles of a newly added SNP and the frequencies of the newly extended haplotypes are calculated according to the phased genotype. Those with frequency higher than the preset threshold (0.05 as default) are added to the working group.

$$h_1 = a_1; h_2 = a_2; h_3 = a_1a_2;\ h_4 = a_1A_2;$$

2 SNPs $$h_5 = A_1a_2;\ h_6 = A_1A_2;\ h_7 = A_1;\ h_8 = A_2;$$

$$H_w = (h_i);\ i = 1 \text{ to } 8;\ \text{Freq}\quad (h_i) > 0.05$$

When the number of haplotypes in the working group reached the preset threshold (10 000 as default), the haplotype extension step stops and all obtained haplotypes, except the single allele haplotypes (such as a_1, A_1), are added to an overall haplotype group H_{all} for further analysis.

The above steps are defined as a single window analysis. The working group is cleared for the next single window analysis. If the number of variants in previous window is n and the index of the last variant is l, the index of the first variant in next window analysis would be $l - n/2$.

After all variants were analyzed in window analysis, duplicated haplotypes in H_{all} will then be removed. Haplotypes are considered as redundant haplotypes if any combination of its tag SNP alleles (defined as sub-haplotype) has r^2 (Coefficient of determination) higher than 0.8 between the two haplotypes. To exclude redundant haplotypes, each haplotype in H_{all} would be checked if any of the sub-haplotype was also in H_{all} and with r^2 higher than 0.8. All the remaining haplotypes with frequency higher than threshold are considered as informative haplotypes and are stored in the overall haplotype group H_{all} for further analysis.

2.1.3 Examine haplotypes associated with gene expression and correct significance by an estimation of effective number

The last step is to perform association analysis at the above informative haplotype. The expression level of each coding genes for further analysis is the normalized RPKM value from RNA sequencing or relative value from expression array. For each haplotypes (h_i) in H_{all} of each gene, the genotype for each sample j is coded as 0, 1 or 2 if it contains 0, 1 or 2 copies of the haplotype according to the phased genotypes of the tag SNPs of the haplotype. Linear regression analysis is applied to each haplotype using the additive model:

$$\text{RPKM}_j = \mu + \beta G_{hi,j} + \varepsilon$$

where $G_{hi,\,j}$ is the number of haplotype h_i in sample j, μ is a constant term while ε is the error term. β is the regression coefficient.

The testing of a large number of haplotypes needs to be taken into account in the interpretation of statistical significance for each gene, but this is complicated for the non-independence of haplotypes because of linkage disequilibrium. A correction of multiple testing by the number of haplotypes through the Bonferroni approach will produce conservative P-values. Therefore, we adopt an effective number of independent tests estimator (Li *et al.*, 2012b) we proposed previously to adjust the multiple testing issue. The estimator uses genotype correlation to approximate the effective number of independent tests. The effective number of independent test is smaller than the actual number of tests when genotypes are correlated. For each haplotype, genotypes are encoded by the counts of the haplotype at every subject. The genotype correlation matrix of all haplotypes is calculated by Pearson correlation method. The effective number of independent tests (m_e) is approximated by Li *et al.* (2012b) with the usage of genotype correlation matrix.

The corrected P-value of a haplotype i is:

$$\widehat{p}_i = m_e * p_i$$

where m_e is the estimated effective number of tests among all extracted haplotypes of a gene and p_i is the original P-value of the eQTL association analysis.

As the haplotype analysis is designed to find synergistic effect, haplotypes with association P-value higher than 0.05*P-value of any of its tag SNPs are removed. The approach has been implemented in the programming language Java. The executable application and example data can be accessed at http://grass.cgs.hku.hk/limx/ehaplo/.

2.2 eQTL haplotype analysis for GTEx data

Fully processed, normalized and filtered gene expression data from 8555 samples across 55 tissues and phased genotype data of the corresponding 450 individuals were obtained from GTEx v6 through dbGap authorized access. Individual genotype and expression data from top 10 sample size tissues, ranging from 241 to 385, together with all 10 brain tissues with sample size ranging from 83 to 113, were used for eQTL haplotype analysis. GTEx eQTL SNPs were collected from GTEx public release (http://www.gtexportal.org/home/datasets) for comparison, which used the FastQTL (http://fastqtl.sourceforge.net/) to handle the multiple testing issues. After the eQTL haplotype analysis, number of genes having eQTL SNPs, eQTL haplotypes and both were counted for all 20 tissues.

For each pair of tissues, their eHap genes were compared and the number of the eHap genes that only showed in one of the tissue pairs was counted as its specific eHap genes. The average number of specific eHap genes for the 19 comparisons of each single tissue was calculated and compared with its number of eHap genes as the average percent of tissue specific genes of each tissue. For instance, all genes with eQTL haplotypes from whole blood were compared with skeletal muscle and the fraction of them that are not in skeletal muscle is therefore calculated. By comparing to other 19 tissues individually, the 19 fractions were calculated and the average number of is named as ‘percentage of tissue specific eQTL genes’ for whole blood.

3 Results

3.1 Systematic eQTL haplotype identification in GTEx data significantly expanded tissue specific eQTL genes in different tissues

We first systematically investigated the existence of the eQTL haplotypes in a series of tissues with a proposed regulatory region guided

Table 1. Summary of (A) GTEx top 10 sample size tissues and (B) Ten brain tissues

	Sample size	eSNP gene	eHap gene	All eGene	eHap gene / all eGene	eHap only gene	Avg.% TS eHap gene* (%)
(A) GTEx top 10 sample size tissues							
Muscle skeletal	361	7079	1628	7670	19	591	80
Whole blood	338	6782	2372	7517	26	735	85
Skin sun exposed	302	8558	2023	9288	19	730	82
Adipose subcutaneous	298	8493	1648	9013	16	520	79
Transformed fibroblasts	292	8751	1547	9206	15	455	81
Artery tibial	285	8050	1881	8744	19	694	81
Lung	278	7224	1553	7869	18	645	79
Thyroid	278	9916	1887	10 435	16	519	80
Nerve tibial	256	9849	1859	10 370	16	521	80
Esophagus mucosa	241	7411	1588	8006	18	595	81
Top 10 tissue average	293	8211	1799	8812	18	601	81
(B) Ten brain tissues							
Brain cerebellum	103	4162	1667	4728	29	566	81
Brain caudate basal ganglia	100	2446	1226	3107	33	661	83
Brain cortex	96	2566	1360	3279	35	713	78
Brain nucleus accumbens basal	93	2017	1046	2600	34	583	80
Brain_Frontal_Cortex_BA9	92	2008	971	2548	33	540	81
Brain cerebellar hemisphere	89	3249	1342	3823	29	574	79
Brain Putamen basal ganglia	82	1588	1033	2234	39	646	82
Brain hippocampus	81	1134	893	1765	44	631	80
Brain hypothalamus	81	1157	871	1731	43	574	83
Brain anterior cingulate cortex	72	1211	837	1762	41	551	79
Ten brain tissue average	89	2154	1125	2758	36	604	81

Note: *Average fraction of the tissue specific (TS) genes with eQTL haplotypes. The number is calculated by comparing the eGenes in the tissue with other 19 tissues individually.

eQTL haplotype approach (See details in Section 2). GTEx project (GTex Consortium, 2015) provided excellent resources for this purpose, in which there are genotypes and gene expression data of 8555 samples of 450 individuals, crossing 55 tissues. The top 10 sample size tissues and 10 brain tissues were used for tissue specific analysis, which took around 2 h for each tissue on an ordinary workstation (3.2 GHz CPU, 16 GB RAM, Supplementary Result S7). Genes with eQTL SNPs in all these tissues were collected from GTEx public release (v6). The critical regulatory regions were obtained according to integrative regulatory scores compiled from eight different tools in non-coding regulatory variants by a composite model (Li *et al.*, 2016), which consist of 3.8% of the whole genome (see details in Section 2). As a result, 540 455 SNPs located in the critical regulatory regions were selected out of 14 354, 092 SNPs from the GTEx project. After eQTL haplotype detection on the selected SNPs and gene expression of the GTEx samples, the number of genes with eQTL markers (SNPs and haplotypes) increased remarkably. The average number of genes with eQTL haplotypes (eHap genes) in the top 10 tissues was 1799, ranging from 1547 to 2372. Among these genes, on average, 601 genes have no eQTL SNPs at all. The number of eHap genes accounted for 18% of all the genes with eQTL markers (eGenes) on average, including SNPs and haplotypes. In the other hand, the average number of eQTL haplotypes in all 10 brain tissues was 1125, and explained 36% of the eGenes. By pairwise tissue comparison (Section 2), the average fraction of tissue specific eHap genes account for 81% of all eHap genes in each tissue, suggesting the high influence of sequence variation on genes varies from tissue to tissue (Table 1, Fig. 1C).

To investigate the basic properties of eQTL haplotypes, all eQTL haplotypes identified in whole blood from GTEx data were collected for further detailed analysis. In total 12 954 eQTL haplotypes of 2372 genes were identified. The average number of tag SNPs in these eQTL haplotypes was 4.5, ranging from 2 to 9, while the expanding range of the tag SNPs within the haplotypes (length of the haplotype) was 207 kb by average, ranging from 1.58 Mb to 36 bp. Given so large distance, there were originally many SNPs and many more potential haplotypes tagged by these SNPs. The regulatory region guided approach substantially reduced the number of candidate haplotypes and made the identification of eQTL haplotype with such long length possible. The frequency threshold of the eQTL haplotype identification was set as >0.05 in this analysis, while the median frequency was 0.233. The fraction of eQTL haplotypes with frequency higher than 0.1, 0.2 and 0.3 were 76%, 46% and 25%. The median *P*-value of the eQTL haplotypes was 3.17×10^{-7}, ranging from 1.73×10^{-5} to 2.22×10^{-16}.

3.2 Exploration on eQTL haplotypes located chromatin states identified interaction enrichment in particular chromatin states pairs

The next interesting question is what types of functional elements support the interaction of variants on the eQTL haplotypes. To answer this question, we utilized a set of fully processed finely-mapped chromatin states obtained from ENCODE, which were learned by integrating ChIP-seq data from nine cell lines using a Hidden Markov Model (ChromHMM; Ernst and Kellis, 2012). All eQTL haplotypes identified in whole blood from GTEx data and eight major chromatin states from human blood cell line (GM12878) were selected for the analysis, including active, weak and inactive promoters, strong and weak enhancers in upstream or downstream of genes and insulators (Supplementary Method S1). Therefore, there are 36 pairwise combinations. For each eQTL haplotype, all SNP located in the eight chromatin-state regions were collected and each SNP pairs were allocated in one of the 36 state combinations.

Table 2. The enrichment of the chromatin states between the tag SNPs in eQTL haplotypes

Chromatin state	Active promoter	Weak promoter	Inactive promoter	Strong enhancer 5’	Strong enhancer 3’	Weak enhancer 5’	Weak enhancer 3’	Insulator
(A) Chrome state enrichment ratio for eQTL haplotypes derived from whole blood.								
	ratio(−log(P))							
Active promoter	1.02							
Weak promoter	**1.91 (12.47)**	**1.21 (12.47)**						
Inactive promoter	0.55 (6.09)	0.61	0.68					
Strong enhancer 5’	**1.24 (12.47)**	1.07	**1.43 (8.72)**	**1.22 (12.47)**				
Strong enhancer 3’	0.84	0.85	1.15	**1.54 (12.47)**	1.27			
Weak enhancer 5’	0.77 (12.47)	0.79 (11.22)	1.18	0.96	**1.54 (12.47)**	0.74 (7.19)		
Weak enhancer 3’	0.79 (11.65)	0.58 (12.47)	1.31	1.06	0.86	1.15	1.19	
Insulator	**1.41 (12.47)**	1.18	0.71	1.09	0.77	0.98	0.64 (9.86)	**1.57 (12.47)**
(B) Chrome state enrichment ratio for eQTL haplotypes derived from all brain tissues.								
Active promoter	1.07							
Weak promoter	**1.45 (12.47)**	1.05						
Inactive promoter	0.77	1.18	1.32					
Strong enhancer 5’	**1.28 (12.47)**	1.05	1.04	**1.25 (12.47)**				
Strong enhancer 3’	1.01	1.17 (6.61)	0.74	**1.36 (12.47)**	1.13			
Weak enhancer 5’	0.93	0.94	1.05	0.99	1.06	1		
Weak enhancer 3’	1.04	0.83 (10)	1.19	1.02	0.85	1.14	1.12 (10.28)	
Insulator	1.12	**1.35 (12.47)**	0.94	0.98	0.61 (11.73)	0.96	0.94	**1.38 (5.86)**

Note: The chromatin states were annotated by Chrom-HMM based on ChIP-seq data of human blood cell line (GM12878).
Significance of bold: ratio >1.2 and *P*<10E-5.

Within all the eQTL haplotypes, 67 270 SNP pairs were identified in all 12 954 whole blood eQTL haplotypes, while majority states of the involved SNPs were annotated as strong (20%), weak promoters (20%) and upstream strong enhancers (26%). By calculating the expected number of all combinations under random distribution, and the actual number of SNP pairs for each state combinations, the enrichments of all pairwise combinations were calculated as the ratio of the observed value and expected value, and the corresponding *P*-values were calculated by chi-square test, respectively.

The result of this analysis on whole blood eQTL haplotypes and blood cell chromatin states showed that they were not equally distributed. Some state combinations are enriched in haplotype eQTLs more often than expected by random. Specifically, nine combinations of regulatory elements were significantly positively enriched in eQTL haplotypes (ratio > 1.2 and $P < 10^{-5}$) and two combinations were negatively enriched (ratio < 0.8, and $P < 10^{-5}$), out of all 36 combinations. Individually, the most significant positively enriched combinations were strong promoter and weak promoter pair, with ratio of 1.91 and *P*-value of 3.1×10^{-13}, followed by insulator-insulator pair ($1.57, 3.1 \times 10^{-13}$), upstream with downstream strong enhancer pair ($1.54, 3.1 \times 10^{-13}$) and upstream weak with downstream strong enhancer pair ($1.54, 3.1 \times 10^{-13}$; Table 2A). Collectively, 3’ and 5’ strong enhancer participated in five significant combinations, suggesting the strong enhancer may be a dominant player of synergistic effect on eQTL haplotypes. The active promoter participated in three of the nine combinations.

To further explore the conservative of these enrichments, we applied the same analysis on eQTL haplotypes from all brain tissues, which are **unmatched** tissues for the chromatin states from blood cell lines. Only six combinations showed significantly positively enrichment and two were negatively enriched (Table 2B). The numbers were half (8/16) of the ones from eQTL haplotypes from matched tissue (whole blood), while only five of them were occurred in the two analyses. Again the 3’ and 5’ strong enhancers are the most frequent participants in the significant combinations.

By comparing the results from the matched tissue analysis and unmatched, it strongly indicated that segments that harbor the eQTL haplotypes were strongly enriched in certain combination of the regulatory elements annotated by chromatin states, and furthermore, only half of the enrichment were conserved between tissues and the rest half were tissue specific.

4 Discussion

In this study, we proposed a regulatory region guided eQTL haplotype association analysis approach. The essentialness of the three haplotype-number-deduction steps of the approach was confirmed by comprehensive evaluation (Supplementary Results S1). By applying it to the whole-genome gene expression across multiple tissues from the GTEx project, we successfully found a non-trivial fraction of genes having significant eQTL haplotypes. The result suggests that haplotype eQTL, which can be detected by the proposed method, is an important complementary of SNP based eQTL in which the former considers synergistic effect of the later.

Moreover, we also showed that most of the genes with eQTL haplotypes (eHap gene) were tissue specific. After mapped onto regulatory functional elements, the eQTL haplotypes are overwhelmingly covered by a combination of strong enhancer and another element. Further analysis in GWAS data showed that the eQTL haplotypes also tend to have higher significant association with human complex diseases only when the eQTL's tissues are related to the diseases (Supplementary Method S2 and Result S3).This work highlights a need for conducting haplotype-based *cis* eQTL analysis for various tissues and the potential of the tissue-matched eQTL haplotypes for prioritizing disease-associated loci.

To further excavate the underlying principle of the eQTL haplotypes detected by the proposed approach on GTEx dataset, we analyzed the combination of different chromatin states of the tag SNPs of the eQTL haplotypes. The expected numbers of all 36 combinations were compared with observed number in all eQTL haplotypes in GTEx whole blood, which was matching the cell type that generating the chromatin states. Interestingly, nine combinations showed significant enrichment in the observed number with large effect size.

Interactions between two insulators which locates upstream and downstream of the under regulated enhancer-promoter pair have been mapped in previous studies (Cavalli and Misteli, 2013; Mora *et al.*, 2016). In the other hand, the enrichment of the interaction between weak and strong promoters would also support the observations in previous report, showing that weak promoters conveyed significant enhancer function to their stronger interacting partners (active promoters) to control the gene expression (Li *et al.*, 2012a; Supplementary Result S2).

To further investigate whether these enrichments were tissue specific, we further analyzed the eQTL haplotypes from GTEx brain tissues, with the chromatin states generated from blood cells, as an unmatched tissue analysis. Five out of the nine state pairs in previous matched tissue analysis still showed significant enrichment while the signals in the other four faded out. Probably, the interactions in the former five chromatin-state pairs were relatively conserved in different tissues (Table 2B). A matched analysis of the brain eQTL haplotypes on chromatin state from brain cells would better address this phenomenon but it was not available in public ChIP-seq dataset.

In summary, we proposed a regulatory region guided eQTL haplotype detection approach and successfully identified many eQTL haplotypes and genes in multiple tissues, based on the application on GTEx data. In the proof-of-principle examples, we showed variant synergistic effect on haplotypes may also play an important role in regulation of gene expression. The haplotype eQTLs can substantially extend the number of eQTLs genes. The synergistic effect may be based on a combination of certain functional elements in which the strong enhancers are heavily involved. These explorations improved our standing of the mechanism of the interaction of variants that influence gene expression and then the risk of complex diseases.

Acknowledgement

The authors thank The GTEx Consortium to provide access to both genotype and expression data on thousands of samples from various tissues.

Funding

This work was supported by Natural Science Foundation of China (31771401), Hong Kong Health and Medical Research Fund (02132236), Hong Kong Research Grants Council GRF 17128515, 17124017, 17121414 HKU 776412M and N_HKU736/14; Hong Kong Research Grants Council Theme-Based Research Scheme T12-705/11; and Hong Kong Theme-based Research Scheme T12C-714/14-R.

Conflict of Interest: none declared.

References

Ahuja,V. *et al.* (2016) Genome-wide gene expression analysis for target genes to differentiate patients with intestinal tuberculosis and Crohn's disease and discriminative value of FOXP3 mRNA expression. *Gastroenterol. Rep. (Oxf)*, **4**, 59–67.

Cavalli,G. and Misteli,T. (2013) Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, **20**, 290–299.

Corradin,O. *et al.* (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.

Edwards,S.L. *et al.* (2013) Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, **93**, 779–797.

Encode Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Garnier,S. *et al.* (2013) Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes. *PLoS Genet.*, **9**, e1003240.

Grundberg,E. *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084–1089.

GTex Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Gymrek,M. *et al.* (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.*, **48**, 22–29.

Hernandez,D.G. *et al.* (2012) Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.*, **47**, 20–28.

Khankhanian,P. *et al.* (2015) Haplotype-based approach to known MS-associated regions increases the amount of explained risk. *J. Med. Genet.*, **52**, 587–594.

Lappalainen,T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

Li,G. *et al.* (2012a) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.

Li,M.J. *et al.* (2016) Predicting regulatory variants with composite statistic. *Bioinformatics*, **32**, 2729–2736.

Li,M.J. *et al.* (2017) cepip: context-dependent epigenomic weighting for prioritization of regulatory variants and disease-associated genes. *Genome Biol.*, **18**, 52.

Li,M.X. *et al.* (2012b) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.*, **131**, 747–756.

Liu,N. *et al.* (2008) Haplotype-association analysis. *Adv. Genet.*, **60**, 335–405.

Montgomery,S.B. *et al.* (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes, *Genome. Res.*, **23**, 749–761.

Mora,A. *et al.* (2016) In the loop: promoter-enhancer interactions and bioinformatics. *Brief. Bioinform.*, **17**, 980–995.

Nica,A.C. and Dermitzakis,E.T. (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.*, **17**, R129–R134.

Solovieff,N. *et al.* (2014) Genetic association analysis of 300 genes identifies a risk haplotype in SLC18A2 for post-traumatic stress disorder in two independent samples. *Neuropsychopharmacology*, **39**, 1872–1879.

Stranger,B.E. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848–853.

Torres,J.M. *et al.* (2014) Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *Am. J. Hum. Genet.*, **95**, 521–534.

Wittkopp,P.J. and Kalay,G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.*, **13**, 59–69.