Background

GEUVADIS (Genetic European Variation in Health and Disease) is a large scale human genomics resource providing sequencing information for individuals from the CEU (Utah residents with European ancestry), FIN (Finns), GBR (British) and TSI (Toscani) populations. For each individual, the genome was sequenced to identify SNP genotypes. In addition, lympoblastoid cell lines were derived from each sample for RNA sequencing. Now 50,000 of the SNP genotypes for 344 samples with the expression levels of five genes are provided for eQTL analysis.

Statistical analysis

All the following analysis was done with R Studio Version 1.0.143 and R 3.4.0

1. Check the data

1.1 Input the data

Five datasheets were provided. QG17_phenotypes, QG17_genotypes are the 5 gene expression and 50,000 SNP information for the 344 samples. QG17_covars includes population and sex as covariates. CEU, FIN, GBR, TSI were assigned with 1, 2, 3, 4. MALE and FEMALE were assigned with 1 and -1 for later analysis. QG17_gene_info includes information for the 5 genes, and they were re-ordered as the same order of columns in the QG_genotypes file. QG17_SNP_info provides SNP id, chromosome and position for the 50,000 SNPs.

1.2 Check the phenotypes

For each of expression information for the five genes across the 344 individuals, a histogram (Figure 1A) and a QQ plot (Figure 1B) were generated. All the phenotypes conformed normal distribution without outliers.

1.3 Check and clean the genotypes

The SNPs are coded as 0, 1, 2 to indicate homozygous (0, 2) and heterozygous (1) in QG17_genotypes. The data was cleaned to make it consistent that 0, 1, 2 refer to the number of minor alleles at each locus. All the individuals have 0, 1, 2 as SNP value, without outlier or missing data. Minor allele frequency was checked to increase power and remove insignificant hits, and no SNPs had MAF < 5%. Hardy Weinberg equilibrium test was applied to each SNP, with HWExact function in "HardyWeinberg" package, to make sure the they fit linkage disequilibrium rule. 277 SNPs were filtered out at this step with a p-value < 0.05 after Bonferroni correction of 50000 tests.

PCA was performed to check the clustering of the individuals for potential covariates, especially population structure. The filtered genotypes were centered with the 'scale' function (center = TRUE, scale = FALSE), and was put into the function 'prcomp' to generate PCA result. 'sdev' of the pca result indicated the variance captured by each PC. Packages "ggplot2" and "ggfortify" were used to plot the PCA result. The function 'autoplot' was used, with pca result as 'object', and covariate information as 'data', population as 'colour' to color different groups based on their population structure.

From the PCA result, population structure affected the clustering. Correspondingly, the correlation between PC1 and population structure was 0.53. FIN and TSI formed distinct clusters while there's overlap between the CEU and GBR cluster (Figure 2). Sex didn't affect the clustering (figure not shown). PC1 captured 0.53% of total variance while PC2 captured 0.38%.

2. Do GWAS analysis:

2.1 Do a linear regression (LRM) test

Maximum likelihood estimation (MLE) were used to estimate the beta values, and likelihood ratio test were used for hypothesis test and p value calculation. To be more specific:

For estimation, $MLE(\hat{\beta}) = (\boldsymbol{x^T x})^{-1}\boldsymbol{x^T y}$
x_i value is 0, 1, 2 for 0, 1 or 2 minor alleles at each SNP locus.
Without covariates, $\boldsymbol{x}$ is a $n * 2$ matrix with 1 as values for the first column, x_i as values for the second column. Covariates are included as extra column for $\boldsymbol{x}$ if consider them. In this study, we used PC components as covariates for linear regression tests.
$\boldsymbol{y} = [y_1, y_2, ..., y_n]$
$Y = \beta X + \epsilon$
H_0 is the simple model with $\beta = 0$.
H_a is the complex model with $\beta != 0$.
y fits normal distribution with $N(\mu, \sigma^2)$
ϵ fits normal distribution with $N(0, \sigma_\epsilon^2)$
To calculate likelihood, $l(\beta, \sigma^2, y, x) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta x)^2$
$\hat{\sigma^2} = \frac{1}{N-1}(y - \hat{y})^2 = \frac{1}{N-1}(y - \hat{\beta}x)^T(y - \hat{\beta}x)$
To calculate LRT for p-values, $LRT = -2\log(\frac{l_s}{l_c}) = -2\log(l_s) + 2\log(l_c)$
LRT fits chi-squared distribution with $df = 1$.

"MASS" package was used here to generate the inverse of a matrix with function 'ginv'.

Principle components (PCs) from PCA were included as covariates for later tests. For each phenotype, to determine the optimal number of PCs to be included, QQ plots of the quantile distribution of the observed p-values against the quantile distribution of expected p-values were performed to check the quality of the GWAS analysis without covariates or with different number (from 1 to 10) of PCs included. The first 10, 3 and 5 PCs were included as covariates respectively for gene ERAP2, PEX and FAHD1. A QQplot with a big and steep tail of significant hits was preferred, because it narrowed the range of potential causal genotypes. No significant hits were observed for genes GFM1 and MARCH7 (Figure 3A).

In addition, 'manhattan' function from the "qqman" package was used to construct Manhanttan plots for the 5 genes (Figure 3B). Blue line indicated the p-value threshold after Bonferroni correction with total number of tests being the number of genotypes multiplied by the number of phenotypes. Red line was generated with the total number of tests being the number of genotypes tested. The names of the SNPs at the peak that passed the blue line were indicated on the plot.

On the Manhattan plots, we see a single peak for ERAP2. The top SNP that have p-value below the threshold is rs7726445. From the provided SNP information and Genome Brower, we know that the SNP is located within LNPEP gene on chromosome 5, which is next to the ERAP2 gene. This is a good indication that the peak contains a true hit.

There are two peaks for PEX6 gene expression with rs1129187 and rs6854915 on top respectively. Rs1129187 is located within the coding region of PEX6 gene on chromosome 6. Rs6854915 is on chromosome 4, which is in trans with the PEX6 gene, and it is the only SNP that the peak has. It is highly possible to be a false positive. The peak with SNP rs1129187 on top is more likely to have a true causal polymorphism within it.

For FAHD1, there is only one peak with rs11644748 on top. The SNP is located within the coding region of gene FAHD1. It's highly possible that the peak includes the true causal polymorphism.

2.2 Do a linear mixed model (LMM) test
A 344*344 covariance matrix (A matrix) was generated to account for the random effects. The matrix showed the covariance between pairs of vectors with 'cov' function, each vector consisting of the SNP sequencing result for one individual. 'pheatmap' function from the "pheatmap" package was used to generate a heatmap using the covariance matrix (Figure 4). The function aggregated the rows using k-means clustering. From the figure we can see at least 2 major squares and some small squares with higher covariance values, which indicates covariates existing among the individuals.

A linear mixed model with direct EM algorithms was tried but was too slow on the computer. The package "mlmm" (https://github.com/Gregor-Mendel-Institute/mlmm/blob/master/misc/emmax.r) was used instead, together with the "emma" package. The coding necessary for the function 'EMMAX' (https://github.com/Gregor-Mendel-Institute/mlmm/blob/master/misc/emmax.r) was pasted. 'EMMAX' function was used, with Y being the phenotype input, X being the genotype input, and K being the covariance matrix. 'nbchunks' is an integer defining the number of chunks of X to run the analysis, allowing the decrease of memory. Here nbchunks was set to 3.

After finishing the analysis and getting the p-values for each SNP, a QQplot (Figure 5A) and a Manhattan plot (Figure 5B) were generated for each phenotype using the same method as 2.1. Similarly, there were no significant hits for GFM1 and MARCH7. The general locations of the peaks for the other 3 genes were consistent between linear regression model and linear mixed model, with the same SNPs at the peak, which is a good indication of the location contains the true causal polymorphisms.

Further analysis and interpretations of the results
3.1 Do the LD heatmaps and regional Manhattan plots for each gene
Packages of "LDheatmap", "snpStats" and "snp.plotter" were used. QG17_SNP_info was cleaned to remove the SNPs that didn't pass the initial check of the genotypes in 1.3. The SNPs that were considered significant hits from the previous analysis (linear regression and linear mixed models) were chosen and coerced into a class called 'Snp.Matrix'. 'ld' function was used to calculate measures of linkage disequilibrium between pairs of SNPs using r^2 and to generate a square matrix, which was input for the 'LDheatmap' function to get the LD heatmaps (Figure 6 and Figure 7).

For the local Manhattan plots, each dot was plotted with the position of the SNP as the value on x-axis, with the –log10(p-value) as the value on y-axis. Only SNPs that passed the threshold (blue line in Figure 3B and 5B) were plotted here for a regional magnification. For PEX6, there's one SNP on chromosome 4 instead of chromosome 6. It's highly possible to be a false hit, as discussed earlier, and was removed from the input SNP list for LD heatmap and local Manhattan plot (Figure 6 and Figure 7).

3.2 Comparing the results from linear regression model (LRM) and linear mixed model(LMM).
Comparing the results from LRM and LMM analysis side by side with the LD heatmaps and local Manhattan plots, we can see that the main regions that potentially contained a true causal polymorphism were generally the same with slight different ranges, which was consistent with the Manhattan plots.

For ERAP2 and FAHD1 genes, the LRM located the causal polymorphism within a narrower range that what the LMM did. The ranges of regions determined by LRM and LMM were almost the same for gene PEX6. One possible explanation for this is that in LRM, we used a limited amount PCs (determined by the inspection of QQplots) to calculate the contribution of covariates to the phenotype variance. The PCs included for analysis only captured a small percentage of variance (Even PC1 only captured ~0.5% of total variance). In LMM, we generated a matrix calculating the covariance between pairs of individuals using all their genotype sequencing

results as “random effects” to model the covariate effect. LMM could be more sensitive as it is more objective and has potentially taken more covariate effect into consideration, and this might explain why we got more hits for ERAP2 and FAHD1 with LMM.

3.3 Further interpretation of result with information from Genome Browser, GeneCards and literature searching.
Because the general region of the peak was the same while LRM results had higher resolution, we used the positions of SNPs got from LRM test as input for Genome Browser (Figure 8). Dec.2013(GRCh38/hg38) was used as the reference genome.

ERAP2 stands for Endoplasmic Reticulum Aminopeptidase 2. The gene encodes a zinc metalloaminopeptidase of the M1 protease, which is enriched in the endoplasm reticulum. The coded protein plays a central role in N-terminal trimming of antigenic epitopes, which is essential to customize long precursor peptides to fit them into the correct length, and the step is essential for presentation on MHC class I molecules. Mutations in the gene could be associated with immune diseases including Spondylitis and Birdshot Chorioretinopathy. The position of the genomic sequence is chr5: 96,875,939-96,919,716 (43778bp). The coding region is chr5: 96,879,686-96,917,605.

The target region that potentially includes the causal polymorphism ranges from 96,894,513 to 97,035,174 on chromosome 5, based on the positions of the SNPs that are considered significant. It includes the majority of the genomic region of ERAP2. Besides the coding region of ERAP2, the target region also covers the promoter (ENSR00001286005, Chr 5: 96,875,800-96,877,601) and many predicted enhancers. GH05F096932(chr5:96931318-96937916), GH05F096919(chr5:96919852-96925137), GH05F096960(chr5:96960300-96964068) are some of the top hits among the 43 predicted enhancers and they overlap with the predicted region.

Peroxisomal Biogenesis Factor 6 (PEX6) codes a member of the ATPase family, and is involved in peroxisome biosynthesis. To be more specific, it is directly involved in peroxisomal protein import, and is also required for the stability of the PTS1(peroxisomal targeting signal 1) receptor, which is a receptor that recognizes PTS tag, and responsible for the localization of proteins tagged with PTS to peroxisome. The position of the gene is hg38 chr6:42,963,870-42,979,220 (15,351 bp), and the coding region locates on hg38 chr6: 42,964,335-42,979,150. Diseases associated with PEX6 include Heimler Syndrome 2 and Peroxisome Biogenesis Disorder 4A.

In the Manhattan plot, two peaks were discovered, with one of them considered as a false positive, as we discussed earlier. The region specified by the tagged SNP hits is chr6:42,873,885-43,108,015. On the local Manhattan plot, we saw one SNP located on the right end of the figure while the rest of the hits gathered between 42,870,000 and 43,000,000. The LD heatmap also showed two half-squares with darker color and higher LD value. Combined information from two figures indicates that the true causal polymorphism is more likely to locate within the first half (42,870,000-43,000,000). The region specified by the SNPs covers the genomic region of PEX6, including its coding region, promoter region(ENSR00001216951, Chr 6: 42,977,800-42,979,801), and several enhancer regions (GH06F042959, chr6:42959666-42961460; GH06F042977, chr6:42977872-42980808; GH06F042927, chr6:42927970-42933607; GH06F042987, chr6:42987306-42990687; GH06F042910, chr6:42910703-42912472). In addition, the region also covers the genomic regions for several genes, including PTCRA, GNMT, CNPY3, etc.

Fumarylacetoacetate Hydrolase Domain Containing 1 (FAHD1) is part of FAH (fumarylacetoacetate hydrolase) protein superfamily. It was initially discovered as a novel mitochondrial enzyme with acylpyruvate hydrolase

activity (PMID: 21878618). More recently, it was also identified as a type of oxaloacetate decarboxylase in eukaryotes (PMID: 25575590). In addition, it is required for mitochondrial function in C. elegans and human endothelial cells. Depletion of the gene leads to mitochondrial dysfunction and cell senescence (PMID: 28286170, PMID: 26266933). The genomic location of the gene is chr16:1,826,967-1,840,207 (13,267 bases). The coding region is chr16:1,827,230-1,838,126.

The tag SNPs with a significant p-value were located within chr16:1,828,065 – 1,868,123. It overlaps with the genomic location of FAHD1, and the coding region of the gene. Similar to the previous cases, the FAHD1 gene promoter (ENSR00000502344, Chr 16: 1,825,800-1,829,001) and predicated enhancer regions (GH16F001825, chr16:1825160-1832070; GH16F001870, chr16:1870199-1871347; GH16F001812, chr16:1812320-1813237; GH16F001832, chr16:1832612-1834274) either overlaps with the predicted region, or are in *cis* with it.

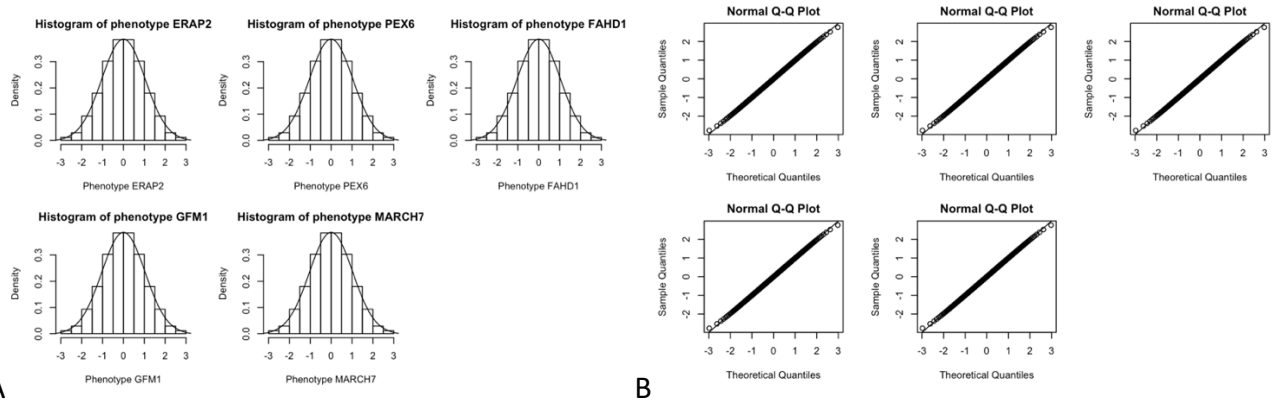
Summary

In summary, linear regression model and linear mixed model were used to do eQTL with the sequencing data of 50,000 SNPs for 344 individuals, and RNAseq data of the expression level for 5 genes. PCA shows that population structure is a main covariate. Results implied from the Manhattan plots, LD heatmaps and regional Manhattan plots with the two methods are generally consistent. For each of the genes ERAP2, PEX6 and FAHD1, there's a single peak that indicates the region which potentially contains the true causal polymorphism affecting the gene expression. It overlaps with the genomic region of the gene itself, including coding region, promoter and enhancers. No SNPs with significant p-values were found for genes GFM1 and MARCH7. One limitation of the analysis is only one variable was used here for the genotype, instead of using both Xa and Xd. Dominant effects were not captured here.

The result is consistent with previous studies, in which people have found that eQTLs tend to cluster near the transcription start time or may be enriched within the transcript regions of the target genes. Moreover, in one study using the microarray data and SNP sequencing data for lymphoblastoid cell line (Gaffney DJ et al. Genome Biology, 2012), they found that 40% of the eQTLs occurred in open chromatin, and particularly transcription factor binding site. Further confirmation of the causal polymorphism requires functional validation with biological experiments.

References

1. https://genome.ucsc.edu
2. http://www.genecards.org
3. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed model association methods. *Nat Genet* **46**, 100-106, doi:10.1038/ng.2876 (2014).
4. Gaffney, D. J. *et al.* Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* **13**, doi:10.1186/gb-2012-13-1-r7 (2012).
5. Shin, J.-H., Blay, S., McNeney, B. & Graham, J. LDheatmap: An R Function for Graphical Display of Pairwise Linkage Disequilibria Between Single Nucleotide Polymorphisms. *16*, doi:https://www.jstatsoft.org/index.php/jss/article/view/v016c03 (2006).



A B

Figure 1. Check the phenotypes. A) shows the histogram of the gene expression values. In B) A QQplot is shown for each gene expression to assess normality of the phenotypes.

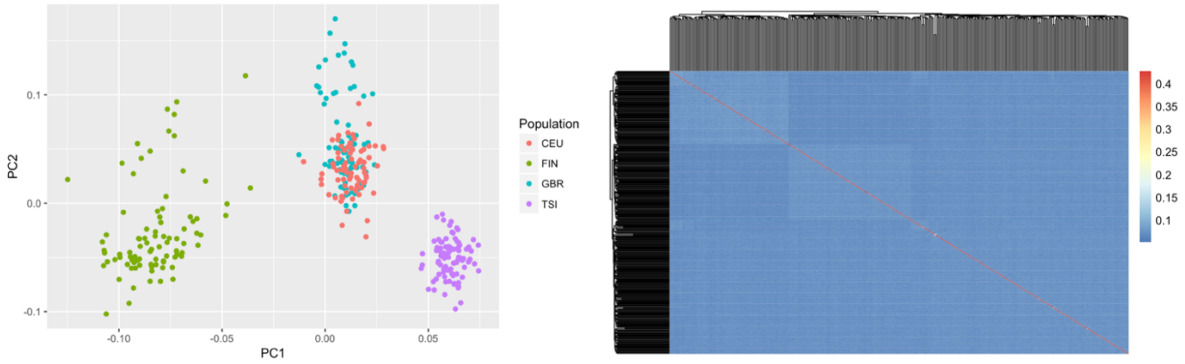
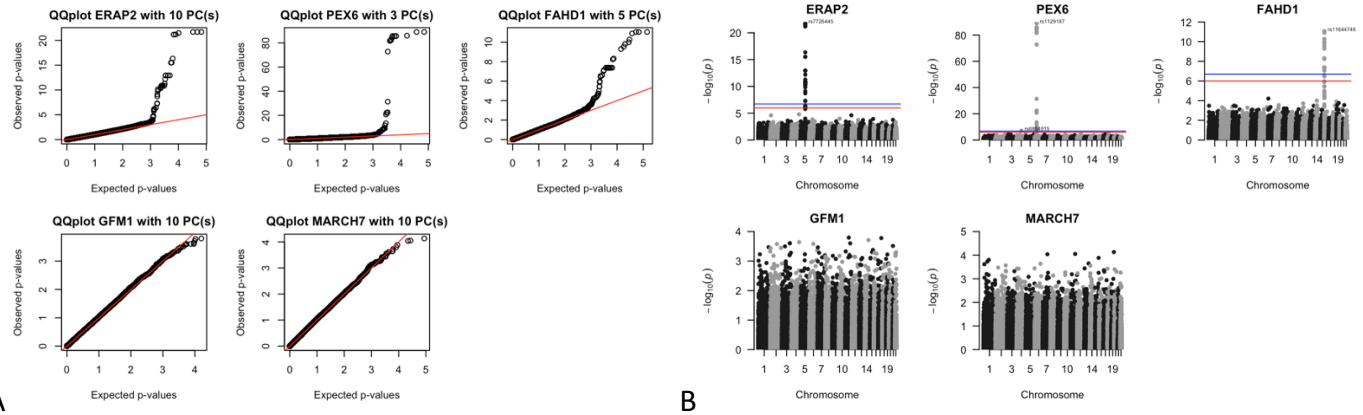
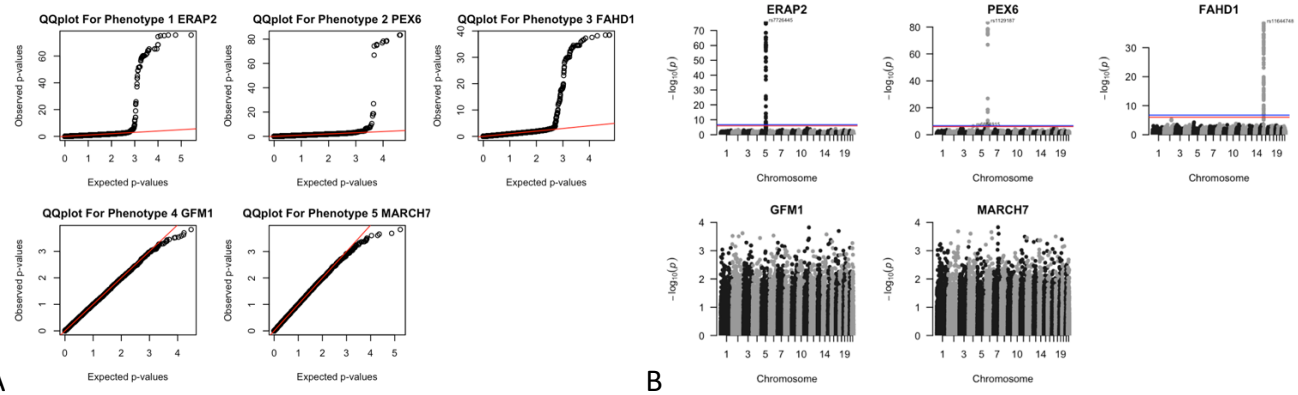


Figure 2. The PCA plot shows the clustering based on genotypes. **Figure 4.** The heatmap of covariance matrix.



A B

Figure 3. QQplot (A) and Manhattan plot (B) with linear regression model.



A B

Figure 5. QQplots (A) and Manhattan plots (B) with linear mixed models.

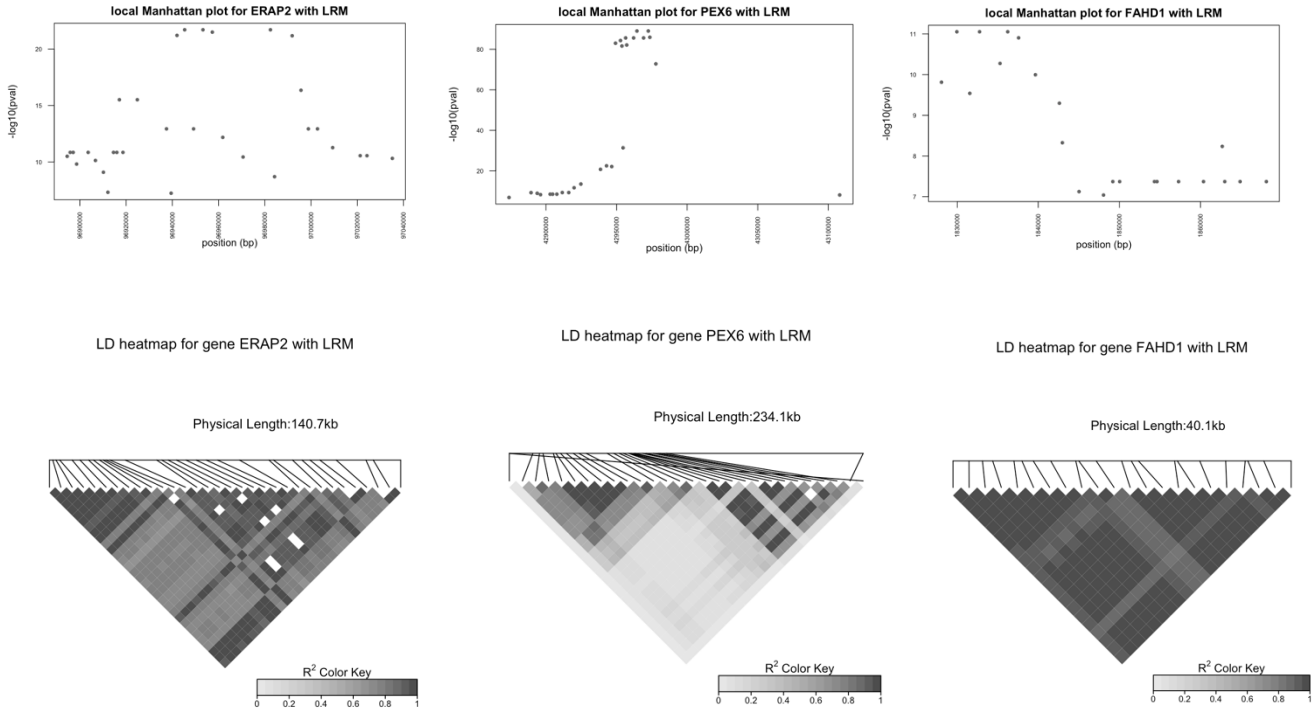


Figure 6. The local Manhattan plots and LD heatmaps for ERAP2, PEX6 and FAHD1 with linear regression.

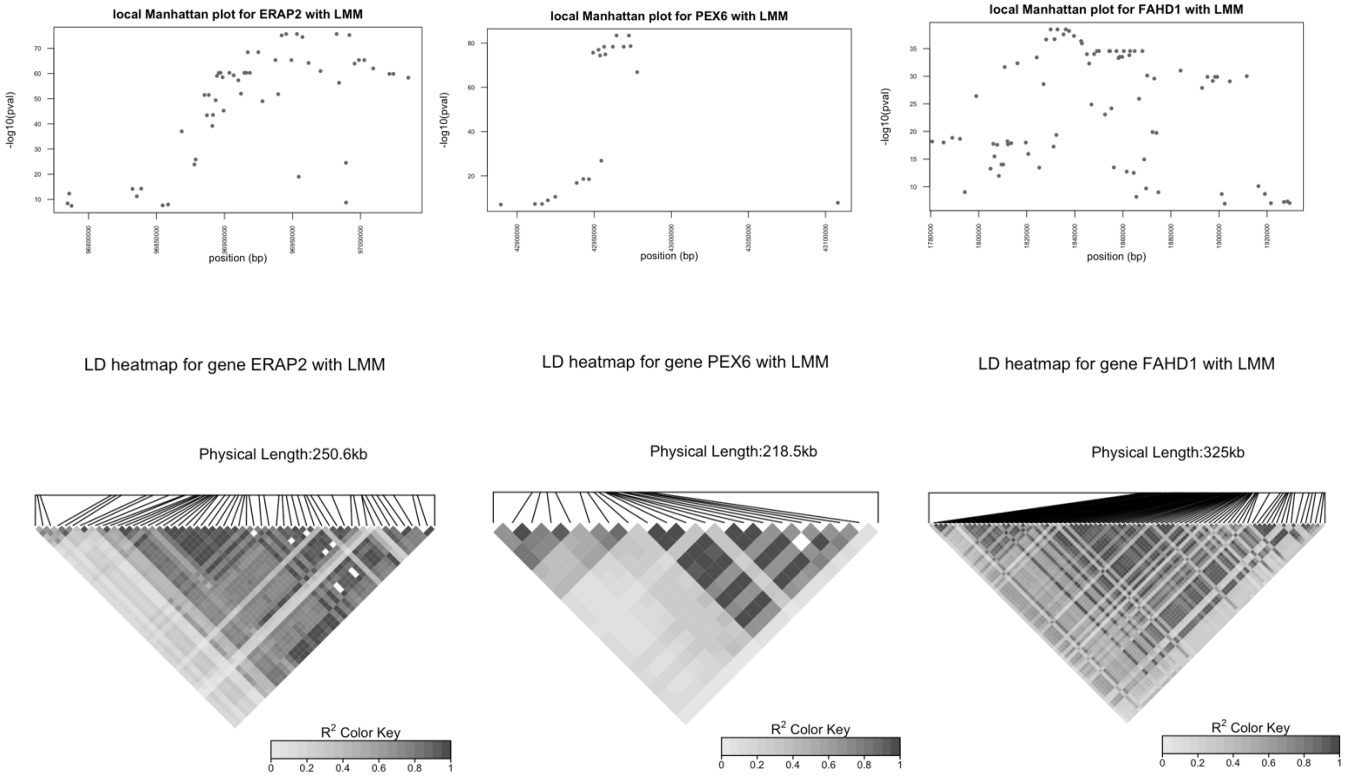
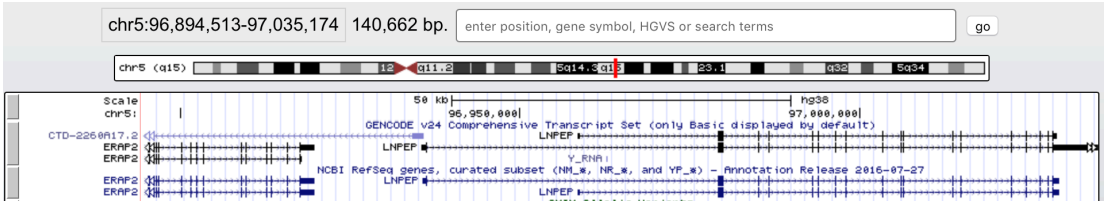
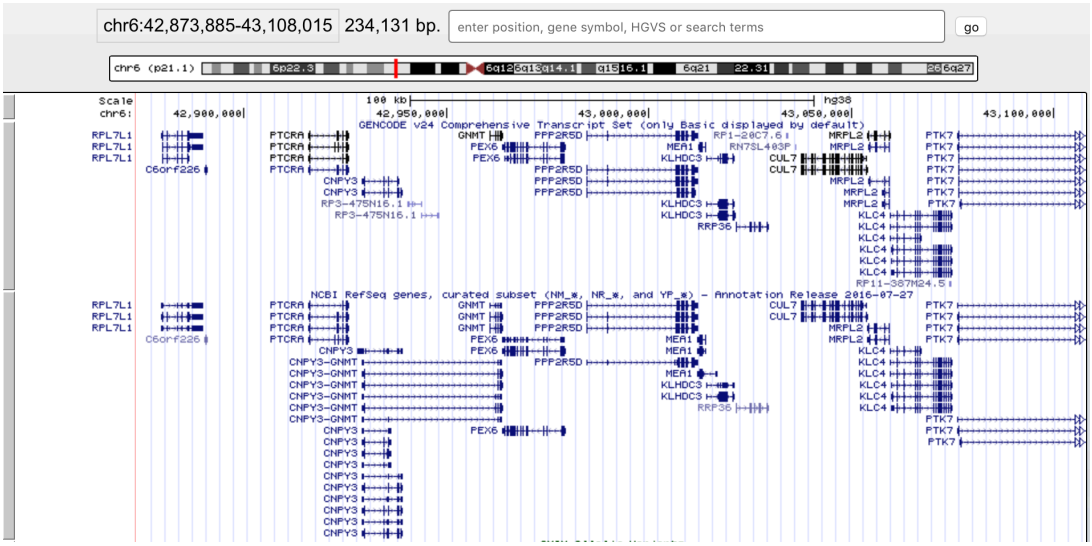


Figure 7. The local Manhattan plots and LD heatmaps for ERAP2, PEX6 and FAHD1 with linear mixed model.

A. For ERAP2 expression:



B. For PEX6 expression:



C. For FAHD1 expression:

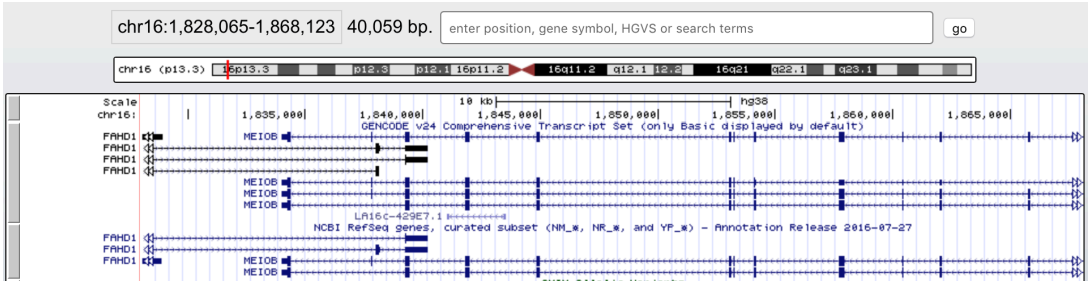


Figure 8. The region covered by the significant SNPs predicated from linear regression model on USCD genome browser.