2016级生物信息学房煜

【原文对照报告-研究生版】

报告编号：123e97c19b8cb74a　　检测时间：2020-05-25 12:12:56　　检测字数：20,550字

作者名称：佚名　　所属单位：南方医科大学

检测范围：

◎ 中文科技期刊论文全文数据库
◎ 中文主要报纸全文数据库
◎ 中国专利特色数据库
◎ 博士/硕士学位论文全文数据库
◎ 中国主要会议论文特色数据库
◎ 港澳台文献资源
◎ 外文特色文献数据全库
◎ 维普优先出版论文全文数据库
◎ 互联网数据资源/互联网文档资源
◎ 高校自建资源库
◎ 图书资源
◎ 古籍文献资源
◎ 个人自建资源库
◎ 年鉴资源
◎ IPUB原创作品

时间范围：1989-01-01至2020-05-25

检测结论：

全文总相似比		复写率		他引率		自引率		专业术语
12.53%	=	12.53%	+	0.0%	+	0.0%	+	0.0%

其他指标：

自写率：87.47%

专业术语：0.0%

高频词：基因,位点,表达,数据,调控

典型相似性： 无

指标说明：

复写率：相似或疑似重复内容占全文的比重

他引率：引用他人的部分占全文的比重，请正确标注引用

自引率：引用自己已发表部分占全文的比重，请正确标注引用

自写率：原创内容占全文的比重

专业术语：公式定理、法律条文、行业用语等占全文的比重

典型相似性：相似或疑似重复内容占互联网资源库的比重，超过30%可以访问

总相似片段：68

期刊：2 博硕：29 外文：0 综合：0 自建库：1 互联网：36

南方医科大学
本科毕业论文

利用等位基因表达不平衡分析寻找人类 CYP3A4、5 、7 基因 eQTL

Searching for cis-eQTL in human gene CYP3A4、5、7 using the allele expression imbalence analysis

房煜

指导教师姓名	李亮
单位名称及地址	南方医科大学基础医学院
专业名称	生物信息学
论文提交日期	2020 年 5 月
论文答辩日期	
答辩委员会主席	

论文评阅人

2020 年 5 月 31 日

摘 要

研究目的

探究对于CYP3A4、5、7基因表达不平衡有调控作用的顺式调控因子。

研究方法

综合使用R语言，实验所得的表达数据，以及卡方检验的方法并作图。最后检索所有有关的调控性SNP所位于的组织和地点。

成果

通过R语言进行作图，并且利用统计学检验的方法，筛选出了对于CYP3A4、5、7基因表达有影响的顺式eQTL。最后用R语言抓取NCBI上的数据并列出所有的SNP以供后续研究。

结论

1、确定了相比于小肠，肝脏更有可能存在调控CYP3A4\5\7等位基因表达不平衡的cis-eQTL。

2、找出了对于等位基因表达不平衡最可能有作用的调控性SNP。

关键词：CYP3A4、5、7 cis-eQTL GTEx

Name:Yu Fang

Supervisor:Liang Li

ABSTRACT

Objective

Research on the cis-eQTL that regulate the expression the allele site of the CYP3A4\5\7 gene

Method

Combing the method of biostatistics knowledge and R package knowledge, analyzing the result from the GTEx.Finally list all the probable SNPs that locate in the

Nearby area of the gene.

Result

By using the R programme, i draw the figure of the cis-eQTL along the upper and down stream of the gene transcription start site controlling the allele expression imbalence of the marker site. Then i use the R to grab the data from the NCBI and list them out to provide for further research.

Conclusion

This experiment finds that compared to small intestine, liver is more likely to contain cis-eQTL that controls the expression imbalance. We find the most probable eQTL with the lowest p value,suggesting what regulatory gene in the upper and lower stream may play the role in expression imbalance.

KEY WORDS:CYP3A7、5、4； cis-eQTL GTEx

目 录

一 引 言

（一）背景介绍

1、 CYP3A亚家族

CYP3A是一种重要的CYP450酶系，它在肝脏、肠道中含量最丰富。CYP3A基因家族中的CYP450约占成年人肝脏CYP450酶总量的25%，临床中约有60%的药物经CYP3A 催化代谢。CYP3A在人体之中的表达存在30倍以上的差异，这些差异造成生物口服利用度和清除率不同。个体间的基因差异导致个体之间的功能存在差异。CYP3A亚家族对于FK506,他克莫司，钙离子拮抗剂等药物都有代谢作用，并且和肝脏癌症、免疫病等息息相关。

2、 CYP3A4、5、7

CYP3A4\5\7存在于人体的第七条染色体上，CYP3A4主要是表达水平的差异，而CYP3A5的差异主要由单核苷酸多态性（SNPs）造成。CYP3A4酶会让许多药物失活，也能使得很多药物活性增加，不同种族间的CYP3A4、5、7的基因型频率各不相同。CYP3A4、5、7中的SNP以及转录水平出现的变化是产生酶活性区别的主要原因。以CYP3A5为例，研究表明，不同的CYP3A5基因型造成的肾脏中CYP3A5含量差异，并且在CYP3A4、5、7以及CYP3A43中的多个SNP存在着连锁的表达不平衡,例如，CYP3A4的启动子区域和CYP3A5、CYP3A7、ZSCAN25的启动子区域存在着启动子之间的作用。已经确定的可能的调控SNP是rs62471956。[1]

3、 CYP3A亚家族等位基因表达不平衡及调控

基因多态性对于生理功能的影响来自于不同的等位基因状态下，mRNA的表达差异。来自调控区域的杂合子，因为等位基因两个基因座

上的碱基不同，所以它们对于同一条基因下游基因的表达调控也不一样，这样就会导致mRNA的表达出现差异。

（二）方法引入

1、 数据库来源

GTEx是第一个收集了多个人体器官mRNA测序的数据库，并提供了跨器官的eQTL研究平台。The Genotype－Tissue Expression (GTEx)计划目的是研究个人的基因组变异如何影响基因表达，导致生物学差异(人体组织和细胞的健康状态和患病状态)。GTEx报告了组织和个体之间基因调控的重要差异，主要包括tissue-specific的基因表达和鉴定许多组织中的基因表达水平的遗传联系（表达数量性状基因座eQTL）。eQTL有助于寻找基因之间和个体之间基因表达的差异。通过对不同的个体/人体组织的基因组和转录组进行测序来鉴定eQTL。GTEx数据主要用作eQTL分析，包括(cis-eQTL)和(trans-eQTL)，其转录组数据需向dbGAP申请。GTEx数据中，GTEx-YYYY表示捐献者。

2、 AEI分析技术和marker SNP

在本实验中，使用等位基因表达不平衡分析寻找CYP3A4,CYP3A5,CYP3A7调控性SNPs位点。杂合子个体中，来自父本和母本的等位基因，它们在同一细胞中，处于相同的外部环境中，在没有顺式eQTL的情况下，他们的表达量应该是一样的。而个体的cis多态性会影响基因的表达及mRNA的加工，导致等位基因有不同的mRNA表达量水平，即AEI(等位基因表达不平衡)。它可以作为一个综合所有顺式作用因子的定量测量。另外，当目的SNP位点位于转录区时，可以直接在总的RNA反转录cDNA中观察到SNP 不同等位基因的特异性表达差异，也就是用这个SNP作为Marker（标志）对两条不同等位基因分别进行表达定量。从而观察在同一杂合子个体中两种不同的allele是否对expression level造成了影响，也就是说排除了个体差异和环境影响之后是否存在由该位点不同基因型所造成的表达差异。该实验需要此SNP位点分型为杂合子的个体的基因组DNA和目的组织细胞抽提得到的总RNA（或者已经反转录好的cDNA）。实验中我们将用基因组DNA的两条allele作为1:1的校正内参，观察组织细胞中的RNA中两条等位基因的比例是否与1:1相差很大，最终判断是否存在AEI的现象，进一步确定该位点与表达水平的关系。

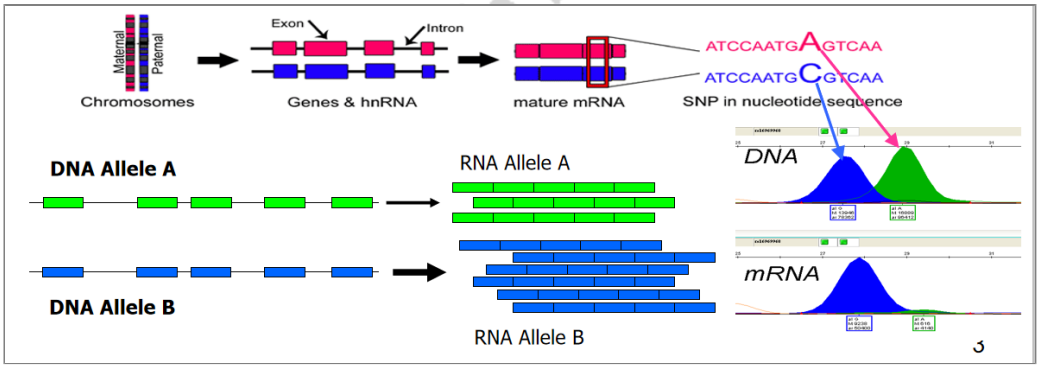


图1：来自两条等位基因座的所转录的mRNA出现了显著的等位基因表达不平衡

3、 卡方检验与关联性分析

卡方检验可以用来检验两个实验对象有没有相关性。例如：卡方检验可以检验男性或者女性对线上买生鲜食品有没有区别；不同城市级别的消费者对买SUV车有没有什么区别。在本实验中，通过对CYP3A4\5\7的位点和基因转录起始位置上下游250kb的野生\突变型（分别是纯合子和杂合子）是否对于表达不平衡有着调控作用。

4、 eQTL

eQTL指达数量形状基因座。这些基因做控制着单个mRNA的表达量。对于这一部分形状而言，它们和mRNA和蛋白质的表达量之间存在一个比例关系。比如身高，化学物质的分泌等等。cis-eQTL是指近距离相关的eQTL。表示这些调控性位点位于所调控的基因附近。实际上，一个完整的调控因子分析流程包括RNA-seq到eQTL分析到等位基因不平衡。对于cis-eQTL和trans-eQTL的分析。[2]

AF GWAS Locus	Nominal Gene Attribution	AF GWAS SNP*	Top eQTL Gene	Top eQTL Symbol	eQTL β†	eQTL *P* Value	eQTL *Q* Value	GWAS SNP Evidence AEI (*P* Value)	AEI Indicator SNP
1q21.3	KCNN3, PMVK	rs6666258	ENSG00000173207	CKS1B	−0.05	4.53E-03‡	5.87E-02	NA§	NA

1q24.2	METTL11B	rs12044963	ENSG00000231437	RP11-88H9.2	0.17	1.51E-02‡	7.36E-01	NA	NA
1q24.2	PRRX1	rs3903239	ENSG00000116132	PRRX1	−0.16	2.86E-05‡	9.33E-04‡	NA	NA
2p13.3	ANXA4, GMCL1	rs3771537	ENSG00000124380	SNRNP27	−0.05	6.22E-06‡	3.11E-04‡	NA	NA
2p14	CEP68	rs2540953	ENSG00000011523	CEP68	−0.10	4.98E-13‡	7.13E-11‡	NA	NA
2q31.2	TTN, TTN-AS1	rs2288327	ENSG00000079150	FKBP7	0.10	2.40E-06‡	1.33E-04‡	NA	NA
3p25.1	CAND2	rs4642101	ENSG00000144712	CAND2	−0.04	8.10E-03‡	8.85E-02	NA	NA
4q25	PITX2	rs6817105	ENSG00000250103	PANCR	0.02	7.44E-01	7.76E-01	NA	NA
5q22.3	KCNN2	rs337711	ENSG00000080709	KCNN2	−0.08	1.55E-03‡	3.35E-02‡	NA	NA

图2：left atrial基因高可能性eQTL和潜在的AEI indicator（marker）

（三）前沿研究

1、 来自内含子区域等位基因不平衡无法被探测，可以用外显子区域的代理SNP，来观测mRNA表达量的差异，从而侦测到内含子区域的不平衡。65个loci中，只有18个找到了代理。使用qtPCR可以测量risk allele上的代理SNP的表达量，并且和other allele上进行比较。[3]

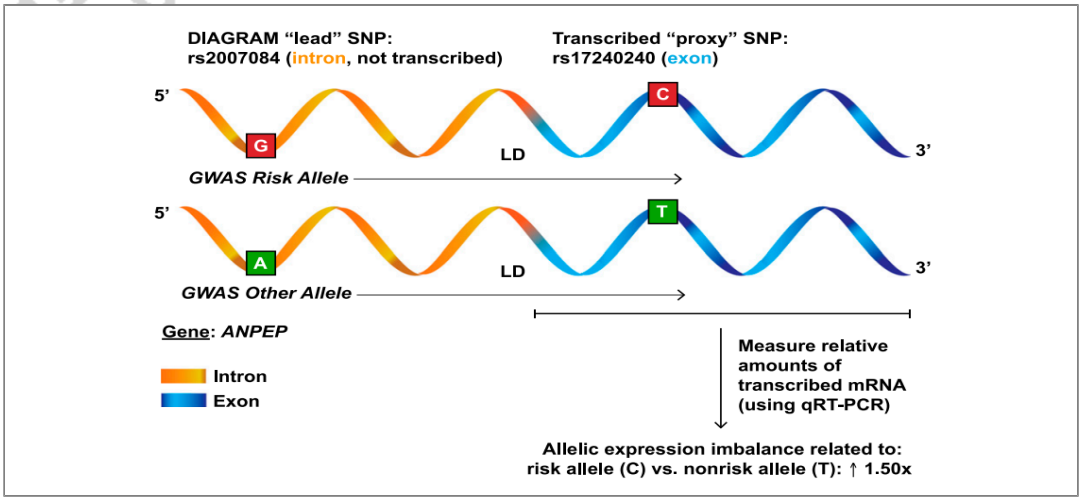


图4：lead SNP可由代理SNP的相对转录量替代。

2、应用等位基因表达不平衡分析寻找潜在致病基因（risk gene）的研究。Mahdi等人通过高通量的等位基因表达不平衡分析发现了潜在的breast cancer致病基因，这篇文章发表在非正式的网站bioRxiv上。使用AEI分析，作者发现了14个潜在的基因。使用正常状态和肿瘤的样本的TCGA和GTEx数据。在文献中，对于每个基因，把样本分成纯合子和杂合子的部分。对杂合子和纯合子的组。可以用双边t检验的方法来检验基因表达是否在杂合子和纯合子之间有显著的区别。如图所示是major transcribed allele 的比例范围。只有一小部分的杂合子是等位基因表达不平衡的。比例出现异常的可以排除出进一步分析的范围。文献中分析研究了不同基因的AEI情况以及致病基因。这进一步为我们研究AEI提供了理论依据。[4]

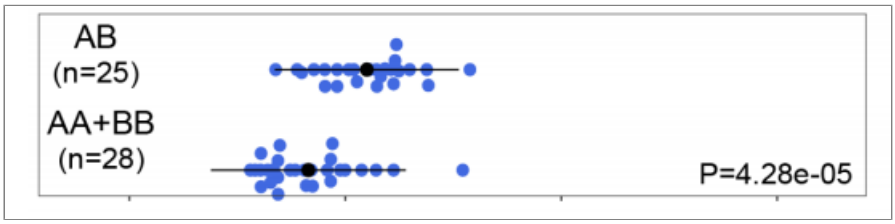


图5：某个基因中，不同样本中杂合子和纯合子的比例

3、LI等人研究了rs9910355对于PRKCA的调控性作用，这个SNP对于PRKCA基因的表达出现AEI（allele expression imbalance，等位基因表达不平衡现象）有着一定的影响。通过引入一个可供作为标志的marker SNP，可以观察在不同的样本或者组织之中，一个基因是否出现了不平衡的现象，从而根据在这些组织样本中有没有出现AEI来判断调控性的eQTL对于转录不平衡有无调控作用。同样的方法可以被应用于CYP3A基因。[5]

4、 其它关于AEI分析技术的研究主要围绕与AEI的分析技术等方面，包括使用GATK进行RNA-seq表达数据的处理，Stephane E. Castel等人认为通过改进方法可以使得等位基因表达数据更加精确[6]。Graham A. Heap等人则通过引入干细胞测序来排除其它环境的干扰，得到更好的表达不平衡数据。[7]

二 材料

（一）表达数据

GTEx838个样本位点等位基因表达数据，包括各个位点出现的基因多样性。

SAMPLE_ID	SUBJECT_ID	CHR	POS	VARIANT	REF_ALL	ALT_ALL	TISSUE_I	REF_COU	ALT_COU	TOTAL_C	REF_RAT	OTHER_A	NULL_RA	BINOM_P	BINOM_P	MAMBA_	MAMBA_	GENOTY	VARIA
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99648291	chr7_9964	A	G	LIVER	245	303	548	0.44708	0	0.496874	0.021002	0.291611	NA	NA	GT:01	3_prime
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99649369	chr7_9964	A	G	LIVER	6	11	17	0.352941	0	0.496874	0.332468	0.86237	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99650681	chr7_9965	C	T	LIVER	23	27	50	0.46	0	0.496874	0.672115	1	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99654245	chr7_9965	A	C	LIVER	47	29	76	0.618421	0	0.496874	0.038599	NA	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99656354	chr7_9965	G	A	LIVER	169	123	292	0.578767	0	0.496874	0.005859	NA	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99657897	chr7_9965	G	T	LIVER	169	142	311	0.543408	1	0.496874	0.112195	NA	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99658137	chr7_9965	A	G	LIVER	182	126	308	0.590909	0	0.496874	0.001128	NA	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99658262	chr7_9965	T	G	LIVER	116	95	211	0.549763	0	0.496874	0.130105	NA	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99659221	chr7_9965	A	G	LIVER	104	87	191	0.544503	1	0.496874	0.193377	NA	NA	NA	GT:01	non_co
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99659796	chr7_9965	T	C	LIVER	64	29	93	0.688172	0	0.496874	0.000245	NA	NA	NA	GT:01	non_co
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99662469	chr7_9966	T	C	LIVER	21	9	30	0.7	0	0.496874	0.02854	0.353467	NA	NA	GT:01	intron_
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99663296	chr7_9966	C	G	LIVER	264	216	480	0.55	0	0.496874	0.022378	0.296945	NA	NA	GT:01	3_prime
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99665212	chr7_9966	C	T	LIVER	180	187	367	0.490463	0	0.496874	0.834658	1	NA	NA	GT:01	synonyi
GTEX-11DXZ-0126-SM-5EGGY	GTEX-11DXZ	chr7	99668312	chr7_9966	C	T	LIVER	62	64	126	0.492063	0	0.496874	0.929234	1	NA	NA	GT:01	intron_

*表达数据来源

- Illumina TrueSeq RNA sequencing
- Affymetrix Human Gene 1.1 ST Expression Array

（二）分型数据

GTEx v8表达数据，主要包括第七条染色体上各位点的基因分型数据

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GTEX-111G	TEX-111G	TEX-111G	TEX-111G	TEX-111G	TEX-111G	TEX-112G	TEX-112G	TEX-113G	TEX-113G	TEX-113G	TEX-117G	TEX-117G	TEX-117G	TEX-118G	TEX-119G	TEX-11DG	TEX-11DG	TEX-11DG	TEX-11DG	TEX-11EG	TEX-11E
chr7	98648055	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648106	chr7_98EA	C	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648166	chr7_98ET	C	.		PASS	AN=1676;GT		1|1	1|1	1|0	1|1	1|1	1|1	1|1	1|1	1|1	1|1	0|1	1|1	1|1	1|0	1|0	1|1	1|0	1|1	1|1	1|1	0|1	1|1
chr7	98648177	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648178	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648234	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648257	chr7_98EG	A	.		PASS	AN=1676;GT		1|0	0|1	1|0	1|1	0|1	1|1	1|1	1|1	1|1	0|1	0|1	1|1	1|0	1|0	1|0	0|0	0|0	0|0	0|1	1|1	0|0	1|1
chr7	98648260	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648409	chr7_98ET	C	.		PASS	AN=1676;GT		0|1	1|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0	0|0	1|1	1|0	1|0	0|0	0|0	0|0	0|0
chr7	98648410	chr7_98EG	A	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648435	chr7_98EA	G	.		PASS	AN=1676;GT		1|1	1|1	1|0	1|1	1|1	1|1	1|1	1|1	1|1	1|1	0|1	1|1	1|1	1|0	1|0	1|1	1|0	1|1	1|1	1|1	0|1	1|1
chr7	98648443	chr7_98EG	A	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648577	chr7_98EAAT	A	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648785	chr7_98EA	G	.		PASS	AN=1676;GT		0|1	1|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0	0|0	1|1	1|0	1|0	0|0	0|0	0|0	0|0
chr7	98648804	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0	0|0	0|1	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98648848	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0
chr7	98649005	chr7_98EC	T	.		PASS	AN=1676;GT		0|0	0|0	0|0	0|0	1|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|1	0|0	0|0	0|0	0|0	0|0	0|0	0|1	0|0

* 分型数据来源

- Whole genome sequencing
- Whole exome sequencing
- Illumina OMNI 5M Array or 2.5M SNP Array
- Illumina Human Exome SNP Array

（三）R语言及excel

R语言是一款用于统计、作图以及完成其它丰富功能的软件，R语言进行生物信息学分析非常常见，比如R语言里的ggplot2包、rvest包，统计检验等等，R语言的文件保存于excel中后，可以利用excel更方便地进行筛选和处理，比一般的程序语言的筛选语句更加有效率。另外除了excel，还可以使用金山提供的WPS来进行分析，其功能相比较来说更贴合国人使用习惯。

（四）在线数据库

1、 NCBI（美国国家生物信息中心数据库）

https://www.ncbi.nlm.nih.gov/

2、 ensembl（欧洲生物信息研究所数据）

http://asia.ensembl.org/index.html

三 实验方法

（一）代表等位基因表达不平衡的样本marker的筛选。

我们已经有了838个GTEx样本的位点特异性表达数据数据，包括reference表达量和alternative表达量counts值。这些数据分别来自不同个体的不同组织，首先我们从838个GTEx数据中找到符合我们要求的部分。表格中存放着838个文件，都是表格形式。我们选中其中所有的文件，并且将它们的名字储存在一个列表里，并且从列表读取依次读取excel文件，每个组织的GTEx表格中中选择位于LIVER（肝脏）和SNTTRN（小肠）的CYP3A4、5、7的表达数据。并且筛选出了其中位于CDS区域，3 prime和5 prime区域的数据。

（二）使用R语言作marker基因ref-alt散点图。

如图1-图6所示，为小肠和肝脏中CYP3A4 5 7的ref-alt散点图，p值小于0.05的即为AEI，用红色表示，其余的为NON-AEI，用蓝色表示。

3、利用线性回归的方法确定小肠中更可能含有顺式调控因子。

从上图中我们可以判断出，尽管在小肠中也存在CYP3A4、5、7的显著不平衡，但是大致上仍然处于一条直线范围内而不存在明显的REF/ALT差异，而对于肝脏，有了cis-eQTL杂合子调控的等位基因表达不平衡，则位于不同的直线上，这表明经过调控后的等位基因出现了明显的表达不平衡，因此我们暂时撇弃小肠，主要寻找肝脏中的等位基因表达不平衡调控分子。本实验对于上述六幅图进行了线性拟合，用来进一步确认肝脏比小肠更可能含有显著的等位基因表达不平衡，这是因为，如果因为调控出现了等位基因表达不平衡，会出现有2条或者多条拟合线，从而导致简单线性回归的准确度（统计学上用p值衡量）下降。

(R 用的 Null hypothesis 默认是 $\beta_0=0$ 和 $\beta_1=0$，所以 p 值越小，代表线性程度和模型可信度越高)

表格1：对肝脏和小肠中等位基因表达数据的ref-alt值的简单线性拟合的p值

组织 基因	CYP3A4	CYP3A5	CYP3A7
肝脏	0.005622	0.0005647	<2.2e-16
小肠	<2.2e-16	4.489e-16	<2.2e-16

由上图可知，肝脏之中不同基因的表达数据的ref-alt数量的线性拟合p值普遍大于小肠，印证了关于肝脏中更有可能存在调控性的顺式eQTL的猜想。

（三）利用marker数据CYP3A4、5、7判断各样本中是否存在AEI

顺式调控性eQTL影响目的基因表达，从而导致表达存在差异的现象。被称为等位基因表达不平衡。

由第二步已经知道，本实验筛选出了GTEx数据中的每一个3-UTRO 5-UTR 和CDS中的位点作为marker，并且对于其中的CYP3A4、5、7中的SNP位点，分别从各自基因起始位点上下游的250kb的分型数据中内寻找对应的基因的野生型和突变型的情况。

为此我们进行AEI分析，其目的在于判断某基因在某组织中是否存在顺式数量形状基因座影响表达。Cis-eQTL应该在有AEI的样本中为纯合子，在没有AEI的样本中为杂合子。

AEI的判断有两种方法，第一种是用测量的办法判断，第二种是AEI的判断应该用minor allele counts/major allele counts的比值来判断。在GTEx的表达数据中已经列出了因此这一部分的位点应该是等位基因表达不平衡的位点。

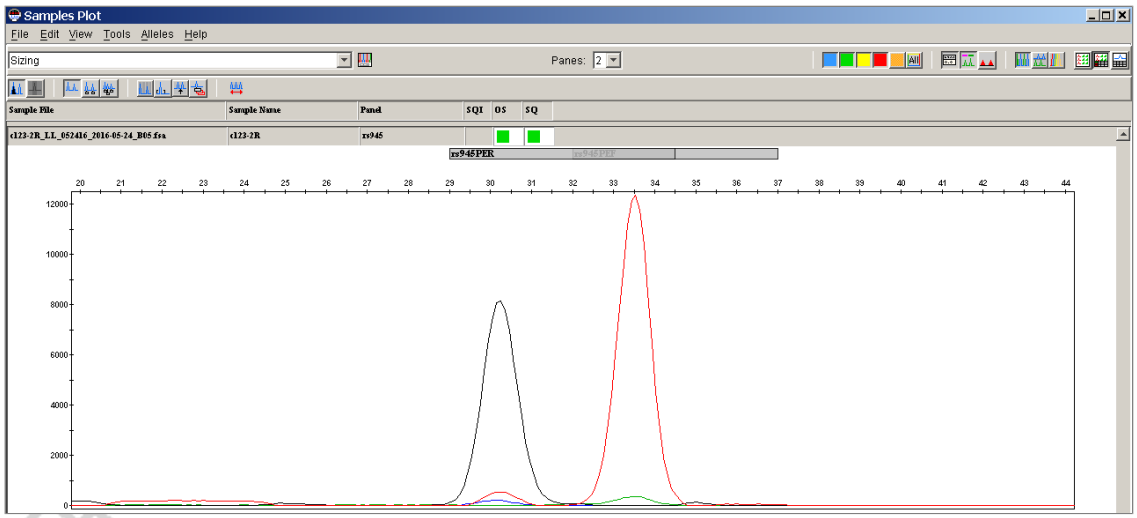


图8：测量法判断AEI，需要对样本cDNA和基因组DNA作3碱基延伸反应

图 9：高通量法AEI可以用major allele counts/minor allele counts的比值偏离1:1的程度来判断，绿色为major等位基因，蓝色为minor等位基因。

在本实验中，我们使用第二种方法判断AEI。以CYP3A4为例，已经筛选出的16个组织的表现CYP3A4表达不平衡的marker位点的情况如下表所示

表格2：CYP3A4基因在16个样本的marker中的AEI/NON-AEI分布

NON	AEI	AEI	NON	NON	NON	NON	NON	NON	NON	NON	NON	AEI	NON	AEI	NON

（四）利用卡方检验找出与marker位点最吻合的位点。

对于基因上下游的分型数据，首先找到已经得到的CYP3A4\5\7的位点数据所对应的样本，并且列出转录起始位点上下游250kb（±250*1000）范围内所有的位点的分型情况，其中0|0,和1|1代表样本在某一位点是野生的纯合子(HOM)，而1|0,0|1代表了样本在某一位点为突变的杂合子（HET）。纯合子中，两个等位基因座上的碱基相同，一般来说对于等位基因表达的作用也相同，因此一般来说只有杂合子调控等位基因表达不平衡，所以样本应该在有AEI的情况下为杂合子，在没有AEI的时候为纯合子.对于CYP3A4、5、7。

另外，已经得知，顺式调控因子应该定位于基因附近，由文献可知，使用基因起始位点上下游250kb（250*1000 base pairs）最为合适，使用ensemble数据库查找可得到CYP3A4、5、7基因的起始位点，可以列出上下游全部的位点的分型数据。仍然以CYP3A4为例。

表格3：CYP3A4基因起始位点上下游分型数据

GTEX-12KS4	GTEX-131YS	GTEX-139TU	GTEX-13O3O	GTEX-14AS3	GTEX-14DAQ	GTEX-14JG1	GTEX-18A7A	GTEX-1QP6S	GTEX-1RQEC	GTEX-U3ZN	GTEX-U8XE	GTEX-WK11
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM

HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM	HOM
...	...	...	...	...	...	...	...	...	...	...	...	
NON	AEI	AEI	NON	NON	NON	NON	NON	NON	NON	NON	NON	AEI

注：该表中，第一行为marker所在的组织，第二行到倒数第二行为marker所在组织在基因起始位点上下游所有位点的分型情况，最后一行为marker SNP的分类。

如表所示，最后一行的NON/AEI为marker的AEI情况，在上下游位点里找到对应的位点在对应组织中的分型情况，如果该位点是潜在的调控性SNP，那么在某一位点下，杂合子的位点在同一样本中应该对应于marker的AEI（等位基因表达不平衡）而HOM对应于NON-AEI。因此我们使用卡方检验的方法。来判断上下游位点和marker的吻合度。

卡方检验p值小于0.05的就认定为这个位点符合预计的相关。也就是说这个位点是调控等位基因表达不平衡的顺式调控因子。检验后卡方检验的p值如下表。

表格4：对上下游位点和marker位点进行卡方检验后所得的结果。

我们将结果整理成如上表所示的表格，显示了上下游250kb范围内所有位点和CYP3A4、5、7作关联检验的p值结果，筛去NA值后所得的

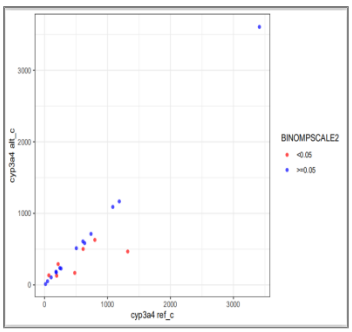
结果。

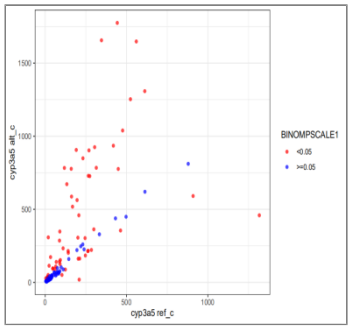
注：该表为CYP3A4每个位点和marker SNP关联程度的p值表

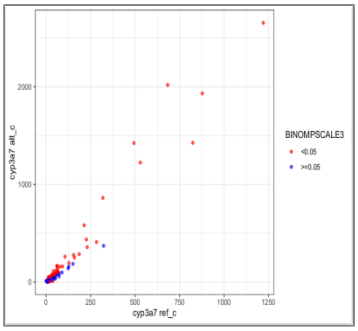
pval	POS	GENE
0.412783208	99535987	CYP3A4
1	99536200	CYP3A4
1	99536585	CYP3A4
0.412783208	99540251	CYP3A4
1	99540980	CYP3A4
1	99541738	CYP3A4
1	99542493	CYP3A4
1	99542685	CYP3A4
...	...	...

四 结果

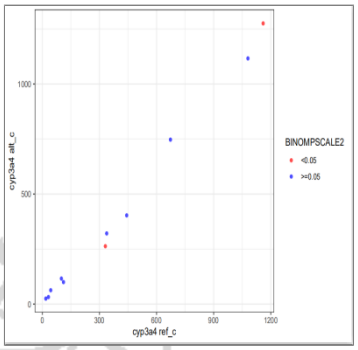
（一）肝脏标志SNP位点的ref-alt counts散点图

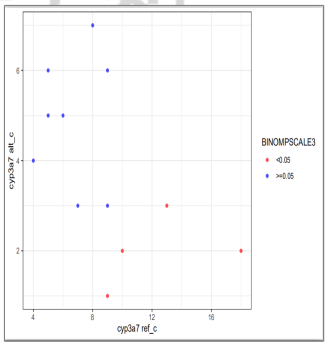


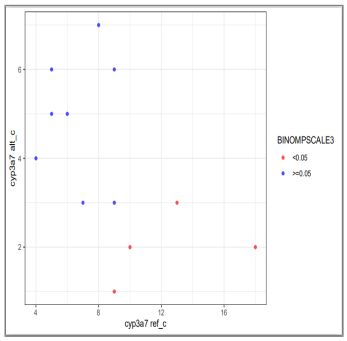




（二）小肠标志SNP位点的ref-alt counts散点图







注明：从左到右分别为CYP3A4、5、7

（三）调控性SNP

对CYP3A4、5、7,分别对分型位点和marker进行卡方检验。所得的p值取负log10，并且以所得的p值的负对数和这些分型位点的坐标（hg38坐标系下）作散点图，p值小于0.05的以红色表示，p值大于0.05的用蓝色点表示，基因坐标下反向箭头的起始和终止位置分别表示CYP3A4\5\7基因的起始位点和终止位点。黄色边框标注出CYP3A4、5、7中p值最小（-log10最大）的位点（如果p值不同选前五名，相同则全选）。利用NCBI中的dbSNP数据库（网址：https://www.ncbi.nlm.nih.gov/snp/），在搜索栏中输入“7：”和位置，

可以搜索到位置对应的SNP的rs序号。

1、CYP3A4基因检验曼哈顿图及可能的调控性SNP位点

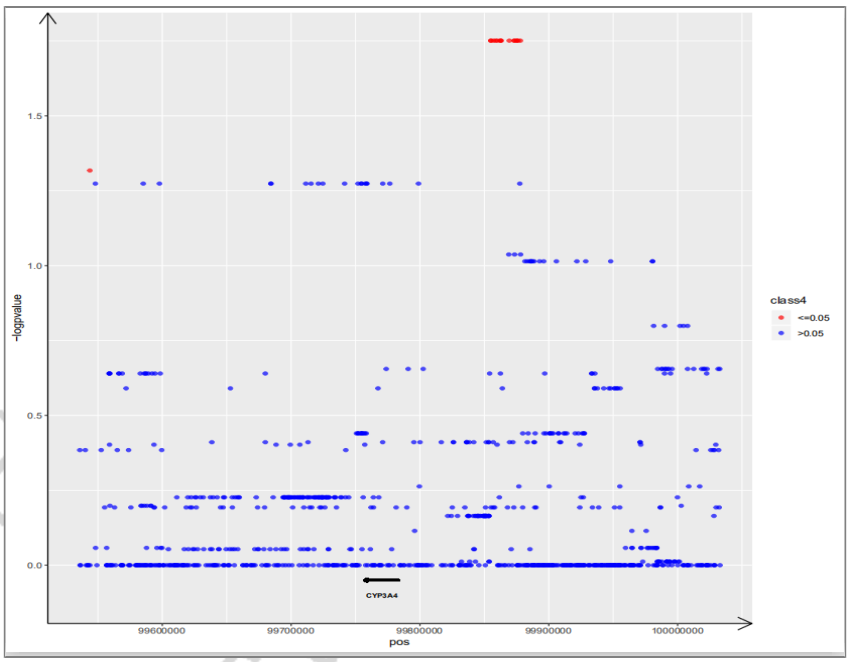


图10：肝脏CYP3A4 卡方检验（-logp）-position图

表格5：CYP3A4卡方检验p值最小的位点以及对应的SNP

pval	POS	GENE	class	logp	SNP
0.017750318	99855044	CYP3A4	<=0.05	1.750793862	rs35987562（CYP3A43）
0.017750318	99855049	CYP3A4	<=0.05	1.750793862	NA
0.017750318	99857132	CYP3A4	<=0.05	1.750793862	rs517284
0.017750318	99858788	CYP3A4	<=0.05	1.750793862	rs480596
0.017750318	99859982	CYP3A4	<=0.05	1.750793862	rs680055
0.017750318	99862065	CYP3A4	<=0.05	1.750793862	rs10654296
0.017750318	99862835	CYP3A4	<=0.05	1.750793862	rs580123
0.017750318	99862988	CYP3A4	<=0.05	1.750793862	rs11981167
0.017750318	99869375	CYP3A4	<=0.05	1.750793862	rs559239

0.017750318	99873190	CYP3A4	<=0.05	1.750793862	rs613963
0.017750318	99873228	CYP3A4	<=0.05	1.750793862	rs516481
0.017750318	99875063	CYP3A4	<=0.05	1.750793862	rs35649099
0.017750318	99875234	CYP3A4	<=0.05	1.750793862	rs35405904
0.017750318	99876235	CYP3A4	<=0.05	1.750793862	rs17161997
0.017750318	99877964	CYP3A4	<=0.05	1.750793862	rs66629054
0.048151582	99543750	CYP3A4	<=0.05	1.31738944	rs10278040

CYP3A4可能的调控因子主要位于CYP3A43和一些基因间隔区。

2、CYP3A5基因检验曼哈顿图及可能的调控性SNP位点

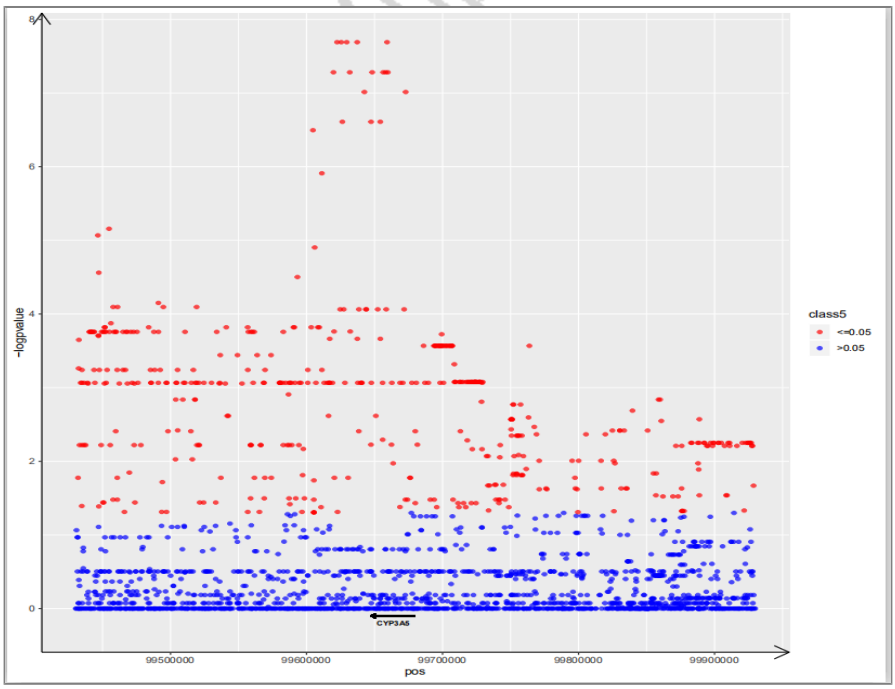


图11：肝脏CYP3A5 卡方检验（-logp）-position图

表格6：CYP3A5卡方检验p值最小的位点以及对应的SNP

pval	class	POS	gene	logp	SNP
2.04E-08	<=0.05	99622460	CYP3A5	7.690369833	rs780822(ZSCAN5)

2.04E-08	<=0.05	99625524	CYP3A5	7.690369833	rs13362853(ZSCAN5)
2.04E-08	<=0.05	99629549	CYP3A5	7.690369833	rs6859590(ZSCAN5)
2.04E-08	<=0.05	99637276	CYP3A5	7.690369833	rs10229552(ZSCAN5)
2.04E-08	<=0.05	99659221	CYP3A5	7.690369833	rs7780328(CYP3A5 ZSCAN5)

最可能的5个位点来自于ZSCAN5/CYP3A5上。来自这些区域的SNP可能对CYP3A5有调控作用。

3、CYP3A7基因检验曼哈顿图及可能的调控性SNP位点

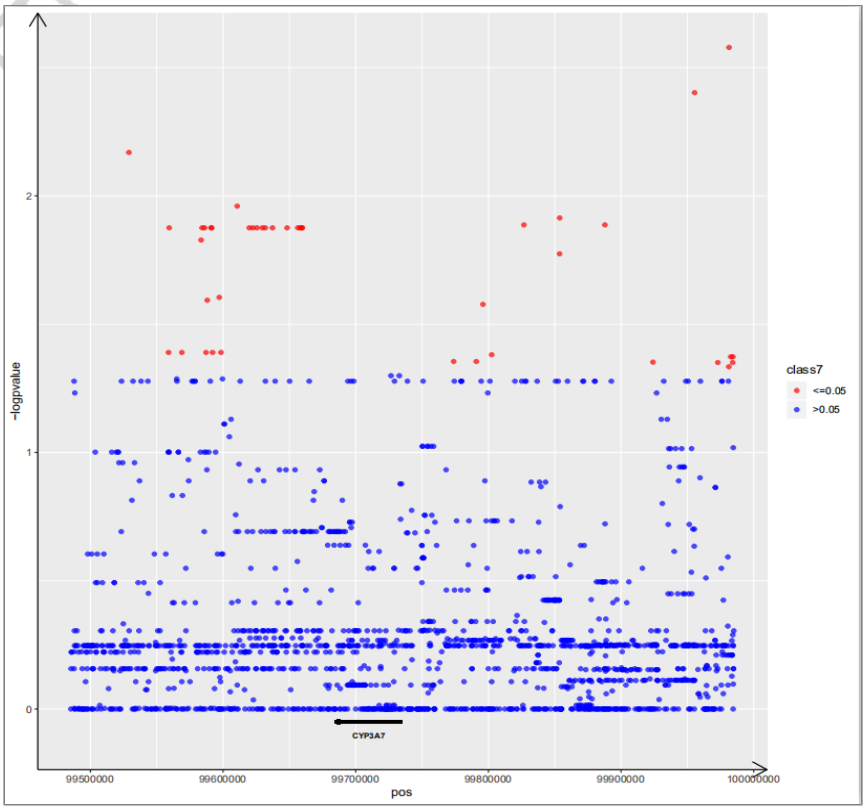


图12：肝脏CYP3A7 卡方检验（-logp）-position图

表格7：CYP3A7卡方检验p值最小的位点以及对应的SNP

pval	class	POS	gene	logp	SNP
0.002640205	<=0.05	99981409	CYP3A7	2.578362351	rs10953293 AZGP1P1
0.003960764	<=0.05	99955443	CYP3A7	2.402221034	rs117268080

0.006774276	<=0.05	99529017	CYP3A7	2.169137113	rs3843540 ZKSCAN5
0.010968751	<=0.05	99610679	CYP3A7	1.959842822	rs10264022 TMEM225B
0.012196186	<=0.05	99853750	CYP3A7	1.913775961	rs522415 CYP3A43

CYP3A7最可能的调控位点分布不固定，说明CYP3A7更可能是多种因素调控的。

五 讨论

（一）CYP3A4、5、7比较

从最终的结果来看，可能的影响CYP3A7表达的cis-eQTL多于CYP3A5,CYP3A5多于CYP3A4。其中CYP3A5有一部分集中于基因的down stream 而CYP3A7则位于基因的 upper stream。CYP3A4\5\7的marker数量分别为

CYP3A4	CYP3A5	CYP3A7
17(+1NA)	109(+0)	116(+1NA)

不过，符合卡方检验p值小于0.05的位点数量和比率则为

CYP3A4	CYP3A5	CYP3A7
16	467	42
16/6747	467/6747	42/6726

这表明，CYP3A5转录起始位点上下游调控等位基因表达不平衡的基因座最为密集。

（二）使用adjusted_p作为划分AEI标准的合理性。

在GTEX数据中，p值分为p和adjust后的p值，在划分AEI的时候，应该使用adjust后的还是选择adjust前的作为p值需要经过检测。使用excel的筛选功能（或者使用程序语言）。可以知道，在筛选出0.43到0.56之间的ref_ratio（我们认为这之间的是non-AEI，计算方式为1/(1+0.77/1.3)）。p值仍然有两个是0.05-。而在替换为adjust_p后，全部的p值变为0.05+。可见，使用adjust_p是精确的。

（三）利用R语言抓取CYP3A5、CYP3A7剩余的位点进行分析。

（四）在CYP3A5、7中，p值小于0.05的比较多，但是在结果中只展示了一小部分p值最小的，使用R语言可以直接从CYP3A5和CYP3A7网站中抓取位点，对剩余的位点进行分析。有了这些SNP，利用数据库源码搜索进一步进入深入研究所有可能的SNP位点，并且还能够知道基于这些位点有没有相关的文献。具体使用的R包为rvest、xml2和dplyr。代码以及样表数据见于附录。因为方法雷同，只列举CYP3A5。从CYP3A5的结果可知。以下SNP在过去的文献中已经有了发现：rs776746、rs15524、rs4646450、rs2257401、rs12333983等。这些SNP位点的文献数已经大于5。最大的高达391。

六 结论

本实验结果讨论认为CYP3A4、5、7顺式调控因子主要存在于肝脏中，在小肠中的可能性较低。并且找到了CYP3A4、5、7等位基因表达不平衡的潜在的调控性位点（eQTL）。比如CYP3A4最可能的位点存在于CYP3A43（）以及一些暂时未报道的基因区域（）,CYP3A5最可

能的位点主要存在于CYP3A5（rs7780328）、CYP3A43（rs35987562、rs517284）、ZSCAN5（rs780822）等基因上，CYP3A7的基因位于TMEM225B（rs10264）、CYP3A43，提示这CYP3A4、5、7基因和这些基因上的调控性位点有相关性。

附 录

代码1

```
Setwd ("D:/data")
Getwd ()
list.files ("GTEx_Analysis_v8_ASE_WASP_counts_by_subject")
x <- list.files ("GTEx_Analysis_v8_ASE_WASP_counts_by_subject")
dir <- paste ("./GTEx_Analysis_v8_ASE_WASP_counts_by_subject/",x,sep="")
n <- length (dir)
```

获得文件长度 838

```
r.data <- read.table(file=dir[1],header=TRUE,sep=",")
r.data<-r.data[which(r.data$tissue%in% c('ENSG00000106258','ENSG00000160868','ENSG00000160870')),]
r.data <-r.data[which9result.data$gene_id %in% c('LIVER','SRTTRM')),]
for(i in 2:n){new.data <- read.table(file=dir[i], header=TRUE, sep=",")
e.data<-e.data[which(r.data$tissue%in% c('ENSG00000106258','ENSG00000160868','ENSG00000160870')),]
e.data <-e.data[which(r.data$gene_id %in% c('LIVER','SNTTRM')),]
r.data <-rbind(r.data,e.data) }
list1<-result.data [which(result.data$tissue=="LIVER"),]
list1<-result.data [which(result.data$tissue=="SNTTRM"),]
list11<-list1 [which(list1$GENE_ID=="ENSG00000106258"),]
list12<-list1 [which(list1$GENE_ID=="ENSG00000160868"),]
list13<-list1 [which(list1$GENE_ID=="ENSG00000160870"),]
list21<-list2 [which(list2$GENE_ID=="ENSG00000106258"),]
list22<-list2 [which(list2$GENE_ID=="ENSG00000160868"),]
list23<-list2 [which(list2$GENE_ID=="ENSG00000160870"),]
list11$BINOMP[list11$BINOM_P>=0.05]=">=0.05"
list11$BINOMP[list11$BINOM_P<0.05]="<0.05"
list12$BINOMP [list12$BINOM_P>=0.05]=">=0.05"
list11$BINOMP [list11$BINOM_P<0.05]="<0.05"
list13$BINOMP [list13$BINOM_P>=0.05]=">=0.05"
list13$BINOMP [list13$BINOM_P<0.05]="<0.05"
list21$BINOMP [list21$BINOM_P>=0.05]=">=0.05"
list21$BINOMP [list21$BINOM_P<0.05]="<0.05"
list22$BINOMP [list22$BINOM_P>=0.05]=">=0.05"
list22$BINOMP [list22$BINOM_P<0.05]="<0.05"
list23$BINOMP [list23$BINOM_P>=0.05]=">=0.05"
list23$BINOMP [list23$BINOM_P<0.05]="<0.05"
```

这里对于大于0.05的和小于0.05的位点进行划分。

```
ggplot(list21,aes(x=list21$REF_COUNT,y=list21$ALT_COUNT,color=BINOMPSCALE))+geom_point(alpha=.3)+labs(x="SI3 REF",
y="SI4 ALT")+theme_bw()+ geom_smooth(method="lm")
ggplot(list12,aes(x=list12$REF_COUNT,y=list12$ALT_COUNT,color=factor(BINOMPSCALE)))+geom_point(alpha=.3)+labs(x="
CYP3A4 in liver:REF",y="CYP3A4 in liver:ALT")+=theme_bw()
```

（重复代码，更换其中的子列表即可）我们可以作出肝脏和小肠中的CYP3A4\5\7的reference和alternative的count数的相对关系的散点图。

代码2

我们从所有的肝脏和小肠中筛选出那些属于stoplost synoymoos和missense varient 3’ 和5’的突变位点，这些位点都是潜在的marker。

```
list1 < -read.table("D:/data/list1.csv",header=TRUE,sep=",")
Ls <- list1[which( list1$VARIANT_ANNOTATION %in% c('stop_lost',' synonymous_variant' ,
'3_prime_UTR_variant','5_prime_UTR_variant','missense_variant')),]
library(ggplot2)
ls1 <- ls[which(ls$GENE_ID=="ENSG00000106258"),]
ls2 <- ls[which(ls$GENE_ID=="ENSG00000160868"),]
ls3 <- ls[which(ls$GENE_ID=="ENSG00000160870"),]
ls1$BINOMP[ls1$BINOM_P>=0.05]=">=0.05"
ls1$BINOMP[ls1$BINOM_P<0.05]="<0.05"
ls2$BINOMP[ls2$BINOM_P>=0.05]=">=0.05"
ls2$BINOMP[ls2$BINOM_P<0.05]="<0.05"
ls3$BINOMP[ls3$BINOM_P>=0.05]=">=0.05"
ls3$BINOMP[ls3$BINOM_P<0.05]="<0.05"
BINOMPSCALE1<-factor(ls1$BINOMP)
p1 <- ggplot(ls1,aes(x=ls1$REF_COUNT,y=ls1$ALT_COUNT,color=BINOMPSCALE1))+geom_point(alpha=.7)+labs(x="cyp3a5
ref_c",y="cyp3a5 alt_c")+theme_bw()+scale_color_manual(values = c("red", "blue"))
BINOMPSCALE2 <- factor(ls2$BINOMP)
p2 <- ggplot(ls2,aes(x=ls2$REF_COUNT,y=ls2$ALT_COUNT,color=BINOMPSCALE2))+geom_point(alpha=.7)+labs(x="cyp3a4
ref_c",y="cyp3a4 alt_c")+theme_bw()+scale_color_manual(values = c("red", "blue"))
p2
BINOMPSCALE3<-factor(ls3$BINOMP)
p3<-ggplot(ls3,aes(x=ls3$REF_COUNT,y=ls3$ALT_COUNT,color= BINOMPSCALE3))+geom_point(alpha=.7)+labs(x="cyp3a7 ref_c",
y="cyp3a7 alt_c")+theme_bw()+ scale_color_manual(values = c("red", "blue"))
p3
```

重复之前的作图步骤。

代码3

```
storea<-c(rep(0,18))
for (i in 1:18)
{for (j in 1:length(names(ud3a4)))
{if(aa4[1,i]==names(ud3a4)[j]){storea[i]<-j}}}
```

```
ud3a4s<-ud3a4[storea,]
storeb<-c(rep(0,109))
for (i in 1:109)
{for (j in 1:length(names(ud3a5)))
{if(aa5[1,i]==names(ud3a4)[j]){storeb[i]<-j}}}
ud3a5s<-ud3a5[storeb,]
storec<-c(rep(0,117))
for (i in 1:117)
{for (j in 1:length(names(ud3a7)))
{if(aa7[1,i]==names(ud3a7)[j]){storec[i]<-j}}}
ud3a7s<-ud3a7[storea,]
ud3a5s<-as.matrix(ud3a5s)
ud3a4s<-as.matrix(ud3a4s)
ud3a7s<-as.matrix(ud3a7s)
Aqpvalue1<-c(rep(0,6747))
Aqpvalue2<-c(rep(0,6747))
Aqpvalue3<-c(rep(0,6726))
for(i in 1:6747)
{line1<-data.frame(aein=factor(ud3a4s[6748,],levels=c("AEI","NON")),homn=factor(ud3a4s[i,],levels=c("HET","HOM")))
aqpvalue<-chisq.test(table(line1)
Aqpvalue1[i]<-aqpvalue$p.value}
Aqpvalueframe1<-data.frame(Aqpvalue1,ud3a4$POS)
for(i in 1:6747)
{line1<-data.frame(aein=factor(ud3a5s[6748,],levels=c("AEI","NON")),homn=factor(ud3a5s[i,],levels=c("HET","HOM")))
aqpvalue<-chisq.test(table(line1)
Aqpvalue2[i]<-aqpvalue$p.value}
Aqpvalueframe1<-data.frame(Aqpvalue1,ud3a4$POS)
Aqpvalueframe2<-data.frame(Aqpvalue2,ud3a4$POS)
for(i in 1:6726)
{line1<-data.frame(aein=factor(ud3a4s[6727,],levels=c("AEI","NON")),homn=factor(ud3a7s[i,],levels=c("HET","HOM")))
aqpvalue<-chisq.test(table(line1)
Aqpvalue3[i]<-aqpvalue$p.value}
Aqpvalueframe3<-data.frame(Aqpvalue3,ud3a7$POS)
```

用excel整理为如正文中的形式

第一栏为p值，第二栏为基因hg38坐标下的位置，第三栏为位于的组织。然后我们画出p值的负对数和位置的点图，这样我们方便看到顺式调控因子相对于CYP3A4\5\7基因的位置。

```
graph4<-read.table("D:/data/4.csv",header=TRUE,sep=",")
graph5<-read.table("D:/data/5.csv",header=TRUE,sep=",")
graph7<-read.table("D:/data/7.csv",header=TRUE,sep=",")
```

```
graph4r<-graph4[!is.na(graph4$pval),]
graph5r<-graph5[!is.na(graph5$pval),]
graph7r<-graph7[!is.na(graph7$pval),]
graph4r$logp=(-log10(graph4r$pval))
graph5r$logp=(-log10(graph5r$pval))
graph7r$logp=(-log10(graph7r$pval))
CYP3A5LOG<-ggplot(graph5r,aes(x=graph5r$POS,y=graph5r$logp,color=class5))+labs(x="pos",y="-logpvalue")+theme(axis.
line = element_line(arrow = arrow(length = unit(0.5, 'cm'))))+geom_segment(aes(x=99679998,xend=99648194,y=-0.1,
yend=-0.1),color="black", size=1.1, arrow = arrow(type="closed",length=unit(0.1, "cm") ) ) + annotate("text", label
= "CYP3A5", x = 99664096, y = -log10(5)+0.5, size=2.5, colour = "black",fontface="bold")+geom_point(alpha=.7)
+scale_color_manual(values = c("red", "blue"))
CYP3A4LOG=ggplot(graph4r,aes(x=graph4r$POS,y=graph4r$logp,color=class4))+labs(x="pos",y="-logpvalue")+theme(axis.
line = element_line(arrow = arrow(length = unit(0.5, 'cm'))))+geom_segment(aes(x=99784248,xend=99756960,y=-0.05,
yend=-0.05),color="black", size=1.1, arrow = arrow(type="closed",length=unit(0.1, "cm") ) ) + annotate("text",
label = "CYP3A4", x =99770604, y = -log10(4)+0.5, size=2.5, colour = "black",fontface="bold")+geom_point(alpha=.7)
+scale_color_manual(values = c("red", "blue"))
CYP3A7LOG=ggplot(graph7r,aes(x=graph7r$POS,y=graph7r$logp,color=class7))+labs(x="pos",y="-logpvalue")+theme(axis.
line = element_line(arrow = arrow(length = unit(0.5, 'cm'))))+geom_segment(aes(x=99735196,xend=99684957,y=-0.05,
yend=-0.05),color="black", size=1.1, arrow = arrow(type="closed",length=unit(0.1, "cm") ) ) + annotate("text",
label = "CYP3A7", x = 99710077, y = -log10(4)+0.5, size=2.5, colour = "black",fontface="bold")+geom_point(alpha=.7)
+scale_color_manual(values = c("red", "blue"))
```

代码4（线性拟合）

```
fit1<-lm(x11$ALT_COUNT~x11$REF_COUNT)
summary(fit1)
```

代码5（网页数据爬取）

首先是根据位置搜索SNP的网址，然后是根据SNP搜索组织

加载xml2，rvest,dplyr

```
web<-read_html(DATA5N$SNP_HTML[i],encoding="utf-8")
w1<-web %>% html_nodes("#main_content") %>% html_nodes("main") %>% html_nodes("div.summary-box.usa-grid-full") %>%
html_nodes("dl:nth-child(2)") %>% html_nodes("dd:nth-child(4)") %>% html_nodes("div") %>% html_text()
w2<-web %>% html_nodes("#main_content") %>% html_nodes("main") %>% html_nodes("div.summary-box.usa-grid-full") %>%
html_nodes("dl:nth-child(2)") %>% html_nodes("dd:nth-child(6)") %>% html_text()
```

unlist然后合并为data frame输出。

在excel表中，按照cit从大到小排列，引用数在2以上的SNP可以进行重点关注。

参考文献

[1]Joseph M. Collins and Danxin Wang,Cis-acting regulatory elements regulating CYP3A4 transcription in human liver, Pharmacogenetics and Genomics ,2020

窗体底端

[2]Hsu J , Gore-Panter S , Tchou G , et al. Genetic Control of Left Atrial Gene Expression Yields Insights into the

Genetic Susceptibility for Atrial Fibrillation[J]. circulation genomic & precision medicine, 2018, 11(3):e002107.

[3]Mccormack S E , Grant S F A . Allelic Expression Imbalance: Tipping the Scales to Elucidate the Function of Type 2 Diabetes-Associated Loci[J]. diabetes, 2015, 64(4):1102-4.

[4]Mahdi Moradi Marjaneh1*, Haran Sivakumaran1*, Kristine M Hillman1 ,SusanneKaufmann1,High-throughput allelic expression imbalance analyses identify candidate breast cancer risk genes, 2020

[5]Li L, Zhang L, Binkley PF, Sadee W, Wang D*. Regulatory Variants Modulate Protein Kinase C α (PRKCA) Gene Expression in Human Heart. Pharm Res. 2017;34(8):1648-1657.

[6]Castel, Stephane E, Levy-Moonshine, Ami, Mohammadi, Pejman, Tools and best practices for data processing in allelic expression analysis[J]. Genome Biology, 16(1):195.

[7]Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing.Human Molecular Genetics, 2010, Vol. 19, No. 1

致 谢

感谢李亮老师在我实习期间对我的谆谆教诲，对我的毕业设计及论文给予的指导和帮助，感谢陆慧洁师姐对我实习上的教导，对我学习上的指导和关怀。感谢所有在我进行毕业设计及论文撰写时给予过我帮助和指导的人。本实验的所有数据皆为李亮老师提供。

本研究受国家级大学生创新创业训练项目（201812121005）资助。

本科毕业论文（设计）原创保证书

本人郑重声明：所呈交的毕业论文（设计）是本人在导师的指导下独立完成，如有抄袭、剽窃、雷同等现象，愿承担相应后果，接受学校的处理。

专业：

年级：

签名：

年 月 日

关注微信公众号