Scientific paper

Low Nucleotide Variability of *CYP51A1* in Humans: Meta-analysis of Cholesterol and Bile Acid Synthesis and Xenobiotic Metabolism Pathways

Monika Lewiñska,[1] Ursula Prosenc Zmrzljak[1,2] and Damjana Rozman[1,*]

[1] *Center for Functional Genomics and Bio-Chips, Faculty of Medicine, University of Ljubljana, SI-1000 Ljubljana, Slovenia*

[2] *Department of Molecular Diagnostics, Institute of Oncology, SI-1000 Ljubljana, Slovenia*

** Corresponding author: E-mail: damjana.rozman@mf.uni-lj.si*
Tel.: (+386) 1-543-7591, Fax: (+386) 1-543-7588,

Received: 07-08-2013

Abstract

Lanosterol 14α-demethylase CYP51 is the most conserved cytochrome P450 (CYP) and is a part of hepatic cholesterol synthesis. Other liver CYPs contribute to cholesterol detoxification through bile acids or to xenobiotic detoxification (DM). To get novel insights into characteristics of the *CYP51A1* locus that was so far not linked to human disorders we performed a meta-analysis of *CYP51A1* gene polymorphisms in comparison to other liver *CYP*s and other genes of cholesterol synthesis. Cholesterol linked genes are generally less polymorphic than DM CYPs, with less coding variants, indicating differences in selection pressure between cholesterol and xenobiotic pathways. Among the studied liver CYPs, *CYP51A1* has the lowest number of coding variants, and less common variants compared to average for cholesterol synthesis. We were not able to detect other functional molecules within the *CYP51* gene (such as lincRNA or miRNA), so we looked into the entire gene locus. We found the *AL133568* sequence that overlaps with the *CYP51A1* promoter region. Our hypothesis was that the AL133568 transcript may have a role in regulating *CYP51A1* expression, but we were unable to prove this experimentally. The reason for the low population variability of the human *CYP51A1* thus remains uncertain.

Keywords: Cholesterol, polymorphism, Cytochrome P450, CYP51, lanosterol demethylase

1. Introduction

Although two unrelated humans share 99.9% of genetic information, the 0.1% can give very valuable knowledge about increased risk of particular diseases or response to drugs. The 0.1% of sequence that humans differ from each other are simple polymorphisms, consisting of single nucleotide polymorphisms (SNP) and small insertion-deletions. To date there are over 56 million of simple nucleotide polymorphisms reported in human DNA (ftp://ftp.ncbi.nih.gov/snp/) consisting ≈1.8% of human DNA sequence. About 25% of all variants are *common,* meaning that they are found in >1% of population. The 3% of all SNPs are variants in coding regions. These can by either synonymous resulting in no change in amino acid sequence, or non-synonymous that can affect the protein structure or function.

The liver is responsible for up to 500 separate catabolic and anabolic reactions, usually in combination with other systems and organs. One of its most important functions is maintaining the physiological levels of cholesterol, the essential metabolite and membrane structural lipid. Many genes of cholesterol synthesis or metabolism associate with Mendelian disorders or complex traits. *CYP51A1* may play an important role in development of pediatric cataract[1] or cerebral cavernous malformations,[2] however, *CYP51A1* mutations were not yet directly associated with human pathologies. CYP51 is evolutionarily the most conserved cytochrome P450 of the huge CYP superfamily.[3,4] The *Cyp51a1* knock-out mouse model is embryonic lethal with features of Antley-Bixler syndrome.[5] According to the best of our knowledge, no other mouse *Cyp* knockout is lethal, even if linked to cholesterol metabolism.[6] On the other hand, in addition to *Cyp51*, other (non-Cyp) genes from cholesterol synthesis genes have proven to be essential. For example, *Hmgcr, Mvk* and *Fdft1* present early embryonic lethal phenotype in the

mouse.[7] It was found that earlier the reaction takes place in cholesterol synthesis, more severe is the knockout phenotype. Consequences of mutations in early cholesterol synthesis genes in humans, as exemplified by *HMGCR* variations, range from preterm delivery,[8,9] to associations with Alzheimer disease[10,11] or blood lipid levels.[12,13] Mutations in genes of the later, post-lanosterol cholesterol synthesis also associate with rare diseases[14,15] such as Smith-Lemli-Opitz or CHILD syndrome, CDPX2, desmosterolosis and lathosterolosis and also with complex traits, like preterm delivery, low birth weight,[8,9] Alzheimer disease[10,16,17] and migraine.[18,19] We do not know, however, how many spontaneous abortions result from fatal mutations in the genes from this pathway.

Bile acid synthesis is the main cholesterol detoxification pathway. Majority of cholesterol that is toxic is excreted as bile acids that are formed with enzymes of the CYP superfamily. As stated before, mouse knockouts in the bile acid synthesis are not embryonic lethal,[6] but human polymorphisms in bile acid *CYPs* associate with some similar non-lethal pathologies as genes from cholesterol synthesis, including blood cholesterol levels (*CYP7A1*) or Alzheimer disease (*CYP46A1*), in addition to other bile acid-related defects (reviewed by *Lorbek et al*[6]). Cholesterol detoxification is frequently compared to xenobiotic detoxification where lipophilic drugs are converted to more soluble products by drug metabolizing CYPs. Key genes of bile acid synthesis and drug detoxification are regulated by identical nuclear receptors, such as CAR or PXR[20–23] and both processes take place in the liver, which is the major organ of cholesterol synthesis.

To get novel insights into characteristics of the *CYP51A1* gene from cholesterol synthesis that was so far not linked to human disorders, we performed a meta-analysis of *CYP51A1* gene polymorphisms in comparison to other human liver *CYP*s and other genes of cholesterol synthesis. In Fig 1 we presented the analyzed metabolic pathways. The drug metabolizing CYPs[24–26] were taken as a biochemical out-group since their major role is not in steroid metabolism. Polymorphisms in drug metabolizing CYPs associate with phenotypes that are not directly linked to cholesterol, but rather to individuals' response to drug treatment.[27,28] We considered possibilities that the major drive of the human *CYP51A1* sequence conservation is the functional conservation of cholesterol synthesis. That conservation might be a general feature of cholesterol synthesis and detoxification through bile acids, or that some common rules exist for the hepatic CYPs.

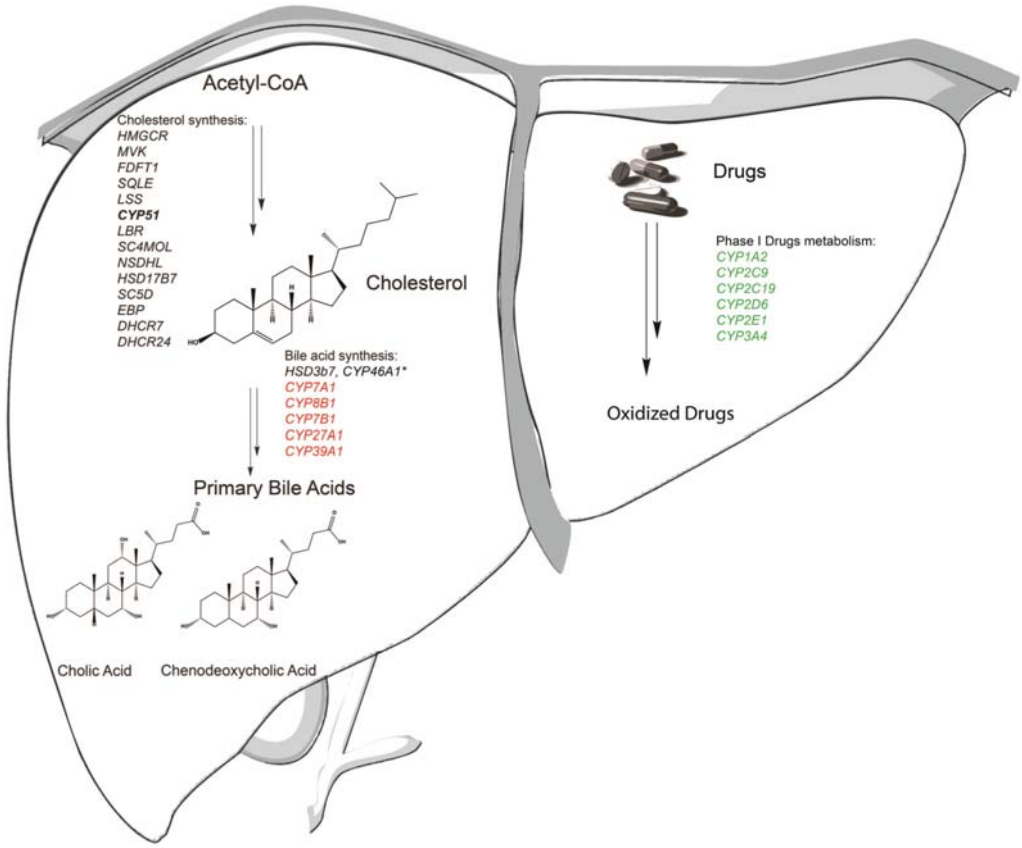

Figure 1 Cholesterol synthesis, Bile Acid synthesis and Drug Metabolizing genes in liver. *CYP51A1* from CS, was compared to CYPs from liver bile acid synthesis (left) and liver drug metabolizers (right).
* The brain specific *CYP46A1* was not compared with other liver CYPs, but was included when comparing the pathways.

2. Materials and Methods

2. 1. Genes Included in Meta-analysis.

The list of genes is provided in Supplementary table 1. The table includes also chromosomal location, information about the transcript variants and absolute numbers of polymorphisms. Genes have been divided into three groups:

- (CS) Cholesterol synthesis genes (*HMGCR, MVK, FDFT1, SQLE, LSS, CYP51, LBR, DHCR14, SC4MOL, NSDHL, HSD17B7, SC5D, EBP, DHCR7* and *DHCR24*),
- (BA) Bile acids synthesis genes – cholesterol detoxification (*HSD3B7, CYP7A1, CYP8B1, CYP7B1, CYP27A1, CYP39A1,* and *CYP46A1*)
- (DM) Drug metabolizers – xenobiotic detoxification (*CYP1A2, CYP2C9, Cyp2C19, CYP2D6, CYP2E1* and *CYP3A4*).

2. 2. Terminology of Investigated Polymorphisms and Applied Databases

A. The list of **ALL** polymorphisms for each investigated region was obtained from UCSC Genome Browser All SNP track (dbSNP build 137, available from ftp.ncbi.nih.gov/snp[29]), that contains information about SNPs and small insertions and deletions (indels) – collectively Simple Nucleotide Polymorphisms.

B. List of the **COMMON** polymorphisms was obtained at ftp.ncbi.nih.gov/snpfor the same region as a subset of SNPs and indels. Only SNPs that have a minor allele frequency **of at least 1%** and are mapped to a single location in the reference genome assembly are included in this subset.

C. List of the **CODING** polymorphisms was obtained from snp137CodingDbSnp track at UCSC Genome Browser (http://genome-euro.ucsc.edu). Here all coding variants resulting in synonymous substitution, missense substitution, premature stop codon substitution or amino-acid residue insertion/deletion are listed.

D. List of all polymorphisms (A) was then investigated for the number of **NON-SYNONYMOUS (missense and nonsense) MUTATIONS** occurring within the region.

2. 3. Amino Acid Sequence Comparison Between Human and Mouse Proteins

E. The **IDENTITY of HUMAN TO MOUSE** proteins were obtained by pairwise alignments generated by BLAST using HomoloGene at http://www.ncbi.nlm.nih.gov/homologene. The identity of *CYP3A4* to mouse homolog *Cyp3a11* (*Q64459)* was generated with Clustal Omega tool implemented with UniProt website www.uniprot.org.

2. 4. Nucleotide Variation Measures

To evaluate nucleotide sequence variation of selected genes we calculated **nucleotide polymorphism** as the relative number of SNPs with respect to the gene size[30] and number of coding polymorphisms in respect of protein size (Supplementary table 2):

F. **Nucleotide polymorphism**, defined as proportion of the number of segregating sites and the total number of sites compared.[30] For our analysis we calculated proportion of ALL SNPs in analyzed genes to all compared sites meaning the length of analyzed gene [kb].

G. **Proportion of polymorphic loci** is defined as number of polymorphic loci (where frequency of most common allele is <0.99) in all loci examined in the population.[31] For our analysis we compared number of polymorphic sites (COMMON SNPs) in analyzed genes to number of all sites in the gene (gene length in kb);

H. Proportion of CODING variants – number of coding variants per number of amino-acids residues in a protein

I. Proportion of NON-SYNONYMOUS mutations – number of missense mutations per number of amino acids in the protein.

We compared the average nucleotide variations within the studied groups of genes (cholesterol synthesis, bile acid synthesis, drug metabolism) to the overall variation of all genes in the genome, applying **GENOME** and **ENCODE**. **GENOME** is defined as the entire DNA of an organism, including its genes (for humans it is ~3 billion base pairs of DNA sequence). The data, for entire human genetic material (GENOME), were obtained from UCSC Genome Browser tracks **ALL**, **COMMON**, **CODING** and filtered for **NON-SYNONYMOUS** mutations (dbSNP build 137, available from ftp.ncbi.nih.gov/snp[29]).

Approximately 1% of human genome sequence represents coding regions, genes and functional elements, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. The built of these sequence is defined as Encyclopedia of DNA Elements (**ENCODE)** pilot regions.[32] Data regarding polymorphisms in ENCODE were obtained as described above.

2. 5. Comparative Analyses

To compare the pathways, the numbers of polymorphisms (common, synonymous, non-synonymous, et.) were calculated for each gene group (cholesterol synthesis, bile acid synthesis and drug metabolism) were compared by One-way ANOVA testing followed by Fishers Least Significant Difference (LSD) post-hoc test using IBM SPSS Statistics 21.

To compare *CYP51A1* to other *CYP*s from bile acid synthesis and drug metabolism, we applied average values

for different polymorphisms of liver CYPs from BA (*CYP7A1, CYP8B1, CYP7B1, CYP27A1* and *CYP39A1)* and DM (*CYP1A2, CYP2C9, Cyp2C19, CYP2D6, CYP2E1* and *CYP3A4*). We also compared the average values to entire GENOME and ENCODE.

2. 6. Expression of *AL133568*

We screened the NCBI database for additional candidate genes that share the same locus with *CYP51* and found a gene LOC 613126 (gene ID 613126) positioned at chr7:91.763.176-91.771870. The two genes are oriented head to head and AL133568 transcript overlaps with 5ž-UTR region of *CYP51* (Supplementary Figure 1). We measured the expression of both transcripts (CYP51 and AL133568) in 20 normal human tissues (Applied Bioscience - Frist Choice Total RNA). Primers for the expression measurements were designed in regions that do not overlap: for AL133568 Fw – CCGTCCACTCCAACTAAAA, Rev – GCTGCAGACCTTCGCAAC and for *CYP51* as previously described.[33] Expression data were normalized according to expression of *18s* (Fw - ACCGCAGCTAGGAATAATGGA, Rev – GCCTCAGTTCCGAAAACCA), *Ppib* (Fw – GGAGATGCACAGGAGGAAA, Rev – CCGTAGTGCTTCAGTTTGAAGTTCT) and *Rplp* (Fw – TGCATCAGTACCCCATTCTATCA, Rev – AAGGTGTAATCCGTCTCCACAGA) as described in.[34]

2. 7. Cloning of *AL133568* and Transfection Measurements

Clone p434N197 with AL133568 transcript was obtained from German cancer research center DKFZ. Originally this clone was produced by a group of S. Wiemann and sequence was submitted at NCBI (http://www.ncbi. nlm.nih.gov/nuccore/6599146). We recloned the AL133568 sequence from pSPORT1 to mammalian expression vector pEGFP-N1 (Clontech) using restriction endonucleases *Kpn*I and *Bam*HI (New England Biolabs).

Transfections were performed with Lipofectamine 2000 (Invitrogen) in Hek293 cell line (ATCC, CRL-1573) following the manufacturers recommendations with some changes. In brief: to observe the expression level of endogenous *CYP51,* cells were switched to the serum free medium when they reached the desired confluence. In this manner endogenous *CYP51* expression was upregulated, to enable observation of possible effects of AL133568. Transfection was performed after 8h in serum free medium. Control transfections were performed with p-CAT basic vector.

For mRNA quantity measurements we performed experiment in 12-well plate. For each treatment and each time point 4 replicas were sampled. Cells were washed with ice-cold PBS and collected with addition of TRI-reagent (Sigma Aldrich). RNA was isolated according to manufacturer's recommendations, cDNA was prepared as described[34] and expression of *CYP51* was measured. Statistically significant difference was calculated with Student T-test, $p < 0.05$ was considered as significant.

For protein detection we performed transfections in petri-dishes. Collection of total proteins, detection and quantification was performed as described.[35]

3. Results

3. 1. Analysis of Cholesterol Synthesis, Bile Acid Synthesis and Xenobiotic Metabolism Pathways

The initial task was to evaluate what is the nucleotide variation in the human population in genes from the three studied pathways, where *CYP51A1* belongs to cholesterol synthesis. Numbers for individual genes are presented in Supplementary Table 1. The average values for each pathway are presented in Table 1. One-way ANOVA followed by LSD revealed significant differences between the pathways on different levels. They differ in number of all nucleotide variants per 1 kb of the sequence ($p = 0.013$) as well as in number of non-synonymous variants with respect to protein length ($p = 0.006$). Interestingly, the post-hoc test that allows comparison between two groups shows no statistically significant differences between cholesterol synthesis genes and bile acids synthesis genes, but a statistically significant difference between drug metabolizing CYPs and the two pathways linked to cholesterol (Table 1). This finding would suggest that both pathways of cholesterol metabolism are not as polymorphic as the pathway of xenobiotic metabolism, even though bile acid synthesis genes are not as conserved as genes of cholesterol synthesis. This part of analysis would suggest that the low nucleotide variation of *CYP51A1* in humans is due to its contribution to cholesterol synthesis.

If we look to the evolutionary conservation of these pathways by comparing the amino acid identity of mouse and human proteins (Supplementary Table 1) and the average values for each pathway (Table 1), we see that cholesterol synthesis shows 86% identity between mouse and the human, bile acid synthesis 79%, and drug metabolism proteins only 74%. This underlines the mouse to human differences in cholesterol and xenobiotic detoxification pathways, while cholesterol synthesis is well conserved. In this pathway there are two genes even more conserved than *CYP51A1* (91%): the *HMGCR* (93%) and *DHCR24* (97%).

3. 2. Analysis of the Human Hepatic Cytochromes P450

CYP51A1 is the only CYP from cholesterol synthesis while several CYPs exist in cholesterol and xenobiotic

Table 1 The one-way analysis of variance using F distribution to compare means between three metabolic pathways – cholesterol synthesis, cholesterol detoxification and xenobiotic detoxification. The p<0.05 was considered as significant.

Pathway		E Mouse/Human Protein Identity	F Nucleotide Polymorphism (ALL/gene length kb)	I Non-synosymous variants/amino acid
Cholesterol synthesis (CS)		86%	20.0	0.086
Bile Acids synthesis pathway (BA)		79%	19.4	0.079
Drug Metabolizers (DM)		74%	35.0	0.140
One-Way Anova p value <0.05		**0.003**	**0.013**	**0.006**
LSD Post-hoc	CS vs BA	**0.038**	0.907	0.100
	CS vs DM	**0.001**	**0.006**	**0.021**
	BA vs DM	0.158	**0.011**	**0.002**

detoxification pathways. Due to a single representative (*CYP51A1*) in one group, the statistical analysis was not possible. We thus compared nucleotide variations of *CYP51A1* with average values for CYPs from liver bile acid synthesis (*CYP7A1, CYP7B1, CYP8A1, CYP27A1* and *CYP39A1*) and drug metabolism (*CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1* and *CYP3A4*). Among the liver CYPs, the human *CYP51A1* has the highest identity (91%) with the mouse homolog (Table 2). Outside the liver, the human to mouse identity is highest (95%) for the brain *CYP46A1* (Supplementary Table 1) which on the body level contributes to the alternative pathway of bile acid synthesis. However, *CYP51A1* is widely spread across kingdoms which is not true for CYP46, thus *CYP51* remains the most evolutionarily conserved cytochrome P450 of the superfamily.[3,36,37] On the human population level, *CYP51A1* contains relatively few nucleotide polymorphisms compared to other studied *CYP* genes. On average genes from xenobiotic detoxification have over twice as many variants residing per 1 kb of sequence com-

Table 2 Liver cytochromes P450 - analysis of all SNPs, COMMON polymorphisms and CODING and Non-Synonymous mutations in respect of gene/protein length CYPs from bile acid synthesis (red) and drug metabolizers (green) and the average for groups compared to only cholesterol synthesis CYP (*CYP51A1*).

Genes/ pathways	Protein size [aa]	Analyzed region length [bp]	E Mouse/Human Protein Identity	F Nucleotide Polymorphism (All/gene length kb)	G Proportion of polymorphic loci (COMMON/ gene length)	H Coding Variants/ amino acid	I Non-synonymous Variants/ amino acids
CYP51	**509**	**22597**	**91%**	**15.05**	**2.66**	**0.09**	**0.06**
Bile acid synthesis genes (Average)	**502**	**70676**	**75%**	**18.86**	**3.89**	**0.12**	**0.09**
CYP7A1	504	9984	82%	19.93	4.21	0.12	0.08
CYP8B1	501	3950	75%	25.32	4.30	0.12	0.07
CYP7B1	506	202820	67%	14.96	2.88	0.09	0.05
CYP27A1	531	33545	74%	16.72	3.85	0.17	0.11
CYP39A1	469	103079	75%	17.4	4.2	0.12	0.09
Drug Metabolizers (Average)	**498**	**32011**	**74%**	**35.0**	**6.34**	**0.21**	**0.14**
CYP1A2	516	7758	73%	30.94	2.96	0.19	0.14
CYP2C9	490	50734	74%	28.38	5.20	0.20	0.13
CYP2C19	490	90209	76%	26.06	4.20	0.22	0.16
CYP2D6	497	4383	71%	75.06	15.51	0.36	0.19
CYP2E1	493	11754	78%	28.50	7.49	0.15	0.10
CYP3A4	503	27229	73%*	21.23	2.68	0.17	0.12
GENOME	N/A	**3137161264**	N/A	**17.93**	**4.43**	N/A	N/A
ENCODE	N/A	**29955196**	N/A	**19.10**	**4.67**	N/A	N/A

* The *CYP3A4* identity was assessed to murine homolog *Cyp3a11using Clustal Omega tool implemented in UniProt website http://www.uniprot.org/*

pared to *CYP51A1* (Table 2). *CYP51A1* is also less polymorphic than the bile acid synthesis *CYP*s. Relatively few *CYP51A1* polymorphisms are common variants (2.66 *vs* 3.89 in bile acids and 6.34 in drug metabolism *CYPs*), indicating that majority of *CYP51A1* variations are rare and only few variants/per amino acid reside in the coding region (0.09 for *CYP51A1* compared to 0.12 for bile acids and 0.21 for drug metabolizers).

Another aspect of our analysis was to evaluate how polymorphic are human hepatic *CYP*s of the three studied metabolic pathways compared to the average nucleotide variation of the entire GENOME or to the genome coding and regulatory parts that are described as ENCODE (Table 2). Here, only the number of all and common polymorphisms/kb of sequence can be compared. Bile acid synthesis CYPs are about as polymorphic as average of GENOME or ENCODE sequences (18.86 polymorphisms/kb of gene for BA *vs* 17.93 and 19.10 for GENOME and ENCODE). Drug metabolizing *CYPs* are substantially more polymorphic (average number of 35 polymorphisms residing per 1kb of sequence) while *CYP51A1* exhibits only 15 polymorphisms/kb. Relations stay similar if we look at the proportion of common polymorphisms. The average for GENOME and ENCODE is 4.43 and 4.67 while *CYP51A1* shows only 2.66 common polymorphisms/kb which is the lowest number of all studied *CYP* genes.

3. 3. Does the Human *CYP51A1* Locus Encode Other Functional Sequences?

The data above (Table 2) indicate that the majority of human *CYP51A1* nucleotide variations are rare and that this gene is less polymorphic than other studied *CYP* and the GENOME and ENCODE. We can deduce from Supplementary Table 2, that *CYP51A1* is less polymorphic compared the average of cholesterol synthesis genes with respect to all, common, coding and non-synonymous variants. *CYP51A1* is generally among the least variable human genes of the cholesterol synthesis pathway. We thus tested the hypothesis that the reason for low variation is not only the essentiality of the gene for cholesterol synthesis, but that potentially other functional molecules might be encoded within this gene locus. The analysis of *CYP51A1* gene locus did not show any precursor forms of microRNAs (pre-miRNAs), C/D box small nucleolar RNAs (C/D box snoRNAs), H/ACA box snoRNAs, nor small Cajal body-specific RNAs (scaRNAs)[38–42] nor lincRNAs (large intergenic non coding RNAs) and TUCPs (transcripts of uncertain coding potential).[43,44] However, we found the *AL133568* (Entrez Gene LOC613126) transcript that is oriented head-to-head with *CYP51A1,* and is overlapping partially with the 5' untranslated region of *CYP51A1* transcript variant 2 (NM_001146152) that differs in the 5' UTR and coding sequence compared to variant 1 (Supplementary Figure 1) and with *CYP51* promoter region.

Our expression analysis of the 20 normal human RNA tissue panel showed that *AL133568* transcript is present only in testes (Figure 2). To get some insights into potential functional role of *AL133568*, we investigated whether the *AL133568* affects the *CYP51A1* expression. Even

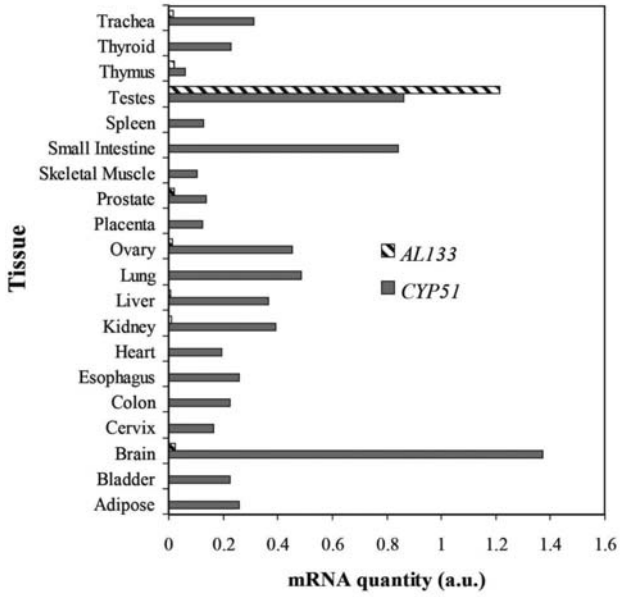

Figure 2 Expression profiles of AL133568 and *CYP51A1* in different normal human tissues.

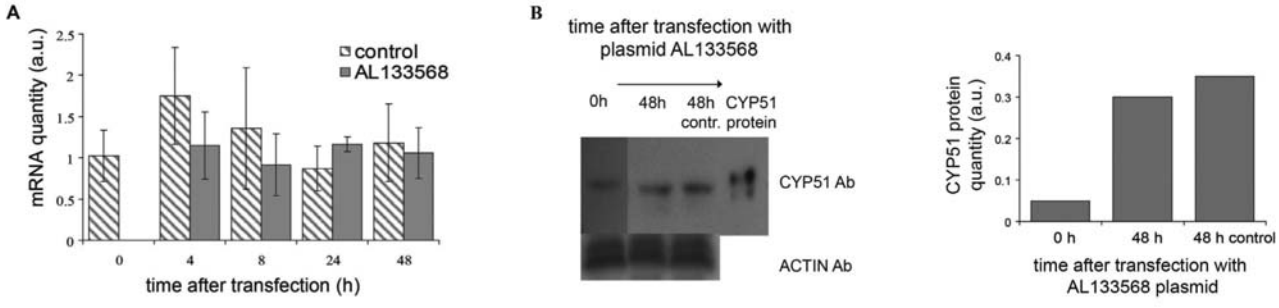

Figure 3 AL133568 transfections. Hek293 cell line was transfected with vector expressing AL133568 and A) an endogenous *CYP51*expression was measured (n = 4) or B) endogenous CYP51 protein was detected.

if there is some trend of *CYP51A1* mRNA diminishment after 4 and 8 hours of *AL133568* expression (Figure 3A), and a trend of CYP51 protein diminishment at 48h (Figure 3B), the differences are not statistically significant. Thus, the functional role of *AL133568* in regulation of *CYP51A1* is not yet conclusive.

4. Discussion

The next generation sequencing technology made human exome and genome easy, fast and affordable to screen the entire populations. Completing the International HapMap Project[45] and 1000 Genomes[46] estimated number of SNPs at the level of 17 million, however today this number is tripled and more variants are being deposited in dbSNP database. While discovery of novel variants in whole genome is in progress, the question is whether all regions of the genome are equally sensitive to discovery of new variants. We asked whether in housekeeping genes of well-conserved cholesterol synthesis pathway polymorphism occur less frequently than in the whole genome, ENCODE, bile acids synthesis pathway and drugs metabolizing genes, and how often the coding and missense mutations are occurring in these genes.

Many of polymorphisms in genes of cholesterol synthesis and bile acid synthesis result in serious phenotypes including lethality. While looking into the ENCODE sequence, which is only about 1% of human DNA sequence, both all and common variations remain on level of 2% and 0.5% (ALL and COMMON variants/kb are 19,10 and 4,67). Not surprisingly, the coding variants represent almost 8% of all variants in the ENCODE sequence. It means that on average in every 1 kb of human DNA we expecting 18-19 variants, out of which 4.5-4.7 (~25% of all SNPs) will be common and much less (3-8%) would carry information resulting in change in the amino acid coding sequence.

In the analyzed pathways, the coding variants represent 13.52.4% (standard error mean SEM) of ALL SNPs in cholesterol synthesis when in the case of bile acids and drug metabolism; the coding variants represent about 24% (24.19.3% in BS and 23.98.1% in DM) of all variants in analyzed regions. This would suggest that it is less likely to find a coding variant in cholesterol synthesis genes compared to the genes involved in cholesterol or xenobiotic detoxification pathways. Mutations in cholesterol synthesis genes can result in lethal phenotype which could explain the lower number of reported coding mutations. Additionally, drug metabolizing CYPs have almost twice as many polymorphisms residing per 1 kb compared to genes of cholesterol synthesis, bile acid synthesis, and genome/ENCODE sequences. We found that *CYP51A1* contains relatively few coding variants with respect to protein size (0.09) compared with other liver CYPs (0.12 in BA and 0.21 in DM). *CYP51A1* is also widely spread across kingdoms and an evolutionarily well conserved gene, possibly due to a crucial role in cholesterol synthesis pathway. It seems that deleterious polymorphisms in the human *CYP51A1* are less likely to be found, since the loss of function of this gene may result in a lethal phenotype, similarly as shown in the mouse model.[5] That can explain the selection pressure of *CYP51* coding regions and higher constraints on CYP51 protein structure relevant for protein-protein interactions; however it does not explain the low variability of intronic regions. Thus, we investigated the hypothesis that the low variability of human *CYP51A1* in the human population might result also from essentiality of other functional molecules that would reside within the *CYP51A1* chromosomal locus. We did not find any miRNA or lincRNA molecules encoded within *CYP51A1* gene. However, we found *AL133568* that is oriented head to head to *CYP51A1* with partially overlapping regulatory regions. Our experimental analyses failed to show the role of *AL133568* on the *CYP51A1* promoter activity, and the mRNA and protein levels. However, we cannot exclude the possibility that *AL133568* contributes to regulation of some more distant genes.

In conclusion, our analysis shows that genes of the well conserved housekeeping pathway of cholesterol synthesis presents on average less polymorphisms and less coding variants in respect of protein size compared to drug metabolizing *CYP*s. Interestingly, this was true also for bile acid synthesis pathway, suggesting that both pathways of steroid metabolism are not so polymorphic than drug metabolizing cytochromes P450 in the liver. *CYP51* as the evolutionarily most conserved cytochrome P450 shows in the human population less than average nucleotide variations compared to other liver *CYP*s. Even if another gene *AL133568* resides at the same chromosomal locus, we were unable to prove its functional role regarding *CYP51A1* expression.

5. Acknowledgements

This work was funded by the FP7 FightingDrugFailure ITN Marie Curie grant #238132 to support M. Lewinska Ph.D. thesis, and by the Slovenian Research Agency program grant P1-0104, grant P1-0010 and project J7-4053.

6. References

1. M. A. Aldahmesh, A. O. Khan, J. Y. Mohamed, H. Hijazi, M. Al-Owain, A. Alswaid and F. S. Alkuraya, *Genet Med* **2012**, *14*, 955–62.
2. L. A. Muscarella, V. Guarnieri, M. Coco, S. Belli, P. Parrella, G. Pulcrano, D. Catapano, V. A. D'Angelo, L. Zelante and L. D'Agruma, *J Biomed Biotechnol* **2010**, *2010*.
3. T. Rezen, N. Debeljak, D. Kordis and D. Rozman, *J Mol Evol* **2004**, *59*, 51–8.

4. D. R. Nelson, J. V. Goldstone and J. J. Stegeman, *Philos Trans R Soc Lond B Biol Sci* **2013**, *368*, 20120474.
5. R. Keber, H. Motaln, K. D. Wagner, N. Debeljak, M. Rassoulzadegan, J. Acimovic, D. Rozman and S. Horvat, *J Biol Chem* **2011**, *286*, 29086–97.
6. G. Lorbek, M. Lewinska and D. Rozman, *Febs J* **2012**, *279*, 1516–33.
7. S. Horvat, J. McWhir and D. Rozman, *Drug Metab Rev* **2011**, *43*, 69–90.
8. E. N. Bream, C. R. Leppellere, M. E. Cooper, J. M. Dagle, D. C. Merrill, K. Christensen, H. N. Simhan, C. T. Fong, M. Hallman, L. J. Muglia, M. L. Marazita and J. C. Murray, *Pediatr Res* **2013**, *73*, 135–41.
9. K. M. Steffen, M. E. Cooper, M. Shi, D. Caprau, H. N. Simhan, J. M. Dagle, M. L. Marazita and J. C. Murray, *J Perinatol* **2007**, *27*, 672–80.
10. C. R. Simmons, F. Zou, S. G. Younkin and S. Estus, *Mol Neurodegener* **2011**, *6*, 62.
11. E. Porcellini, E. Calabrese, F. Guerini, M. Govoni, M. Chiappelli, E. Tumini, K. Morgan, S. Chappell, N. Kalsheker, M. Franceschi and F. Licastro, *Neurosci Lett* **2007**, *416*, 66–70.
12. Y. Tong, S. Zhang, H. Li, Z. Su, X. Kong, H. Liu, C. Xiao, Y. Sun and J. J. Shi, *Lipids* **2004**, *39*, 239–41.
13. Y. Tong, S. Z. Zhang, Z. G. Su, X. D. Kong, J. J. Shi, L. Zhang, H. Y. Zhang and K. L. Zhang, *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* **2003**, *20*, 207–10.
14. R. I. Kelley, *Adv Pediatr* **2000**, *47*, 1–53.
15. F. D. Porter and G. E. Herman, *J Lipid Res* **2011**, *52*, 6–34.
16. I. Greeve, I. Hermans-Borgmeyer, C. Brellinger, D. Kasper, T. Gomez-Isla, C. Behl, B. Levkau and R. M. Nitsch, *J Neurosci* **2000**, *20*, 7345–52.
17. L. J. Sharpe, J. Wong, B. Garner, G. M. Halliday and A. J. Brown, *J Alzheimers Dis* **2012**.
18. B. H. Maher and L. R. Griffiths, *Mol Genet Genomics* **2011**, *285*, 433–46.
19. B. H. Maher, M. Kerr, H. C. Cox, J. C. Macmillan, P. J. Brimage, T. Esposito, F. Gianfrancesco, L. M. Haupt, D. R. Nyholt, R. A. Lea and L. R. Griffiths, *Neurogenetics* **2012**.
20. C. Handschin, M. Podvinec, R. Amherd, R. Looser, J. C. Ourlin and U. A. Meyer, *J Biol Chem* **2002**, *277*, 29561–7.
21. M. Hafner, T. Rezen and D. Rozman, *Curr Drug Metab* **2011**, *12*, 173–85.
22. T. Rezen, D. Rozman, J. M. Pascussi and K. Monostory, *Biochim Biophys Acta* **2011**, *1814*, 146–60.
23. T. Rezen, *Expert Opin Drug Metab Toxicol* **2011**, *7*, 387–98.
24. M. J. Kim, H. Kim, I. J. Cha, J. S. Park, J. H. Shon, K. H. Liu and J. G. Shin, *Rapid Commun Mass Spectrom* **2005**, *19*, 2651–8.
25. S. Zhou, S. Yung Chan, B. Cher Goh, E. Chan, W. Duan, M. Huang and H. L. McLeod, *Clin Pharmacokinet* **2005**, *44*, 279–304.
26. K. S. Lee and S. K. Kim, *J Appl Toxicol* **2013**, *33*, 100–8.
27. A. Gunes and M. L. Dahl, *Pharmacogenomics* **2008**, *9*, 625–37.
28. D. Van Booven, S. Marsh, H. McLeod, M. W. Carrillo, K. Sangkuhl, T. E. Klein and R. B. Altman, *Pharmacogenet Genomics* **2010**, *20*, 277–81.
29. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski and K. Sirotkin, *Nucleic Acids Res* **2001**, *29*, 308–11.
30. M. Nei and S. Kumar, Book Molecular evolution and phylogenetics, Oxford University Press, New York, **2000**.
31. P. W. Hedrick, Book Genetics of populations, Jones and Bartlett Publishers, Sudbury, Mass., **2011**.
32. ENCODE Project Consortium, *PLoS Biol* **2011**, *9*, e1001046.
33. M. Fink, J. Acimovic, T. Rezen, N. Tansek and D. Rozman, *Endocrinology* **2005**, *146*, 5321–31.
34. R. Kosir, J. Acimovic, M. Golicnik, M. Perse, G. Majdic, M. Fink and D. Rozman, *BMC Mol Biol* **2010**, *11*, 60.
35. R. Kosir, U. P. Zmrzljak, T. Bele, J. Acimovic, M. Perse, G. Majdic, C. Prehn, J. Adamski and D. Rozman, *Febs J* **2012**, *279*, 1584–93.
36. G. I. Lepesheva and M. R. Waterman, *Biochim Biophys Acta* **2011**, *1814*, 88–93.
37. N. Debeljak, S. Horvat, K. Vouk, M. Lee and D. Rozman, *Arch Biochem Biophys* **2000**, *379*, 37–45.
38. S. Griffiths-Jones, *Nucleic Acids Res* **2004**, *32*, D109–11.
39. S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman and A. J. Enright, *Nucleic Acids Res* **2006**, *34*, D140–4.
40. S. Griffiths-Jones, H. K. Saini, S. van Dongen and A. J. Enright, *Nucleic Acids Res* **2008**, *36*, D154–8.
41. L. Lestrade and M. J. Weber, *Nucleic Acids Res* **2006**, *34*, D158–62.
42. M. J. Weber, *Febs J* **2005**, *272*, 59–73.
43. M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev and J. L. Rinn, *Genes Dev* **2011**, *25*, 1915–27.
44. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter, *Nat Biotechnol* **2010**, *28*, 511–5.
45. International HapMap Consortium, *Nature* **2003**, *426*, 789–96.
46. G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles and G. A. McVean, *Nature* **2010**, *467*, 1061–73.

Povzetek

Lanosterol 14α-demetilaza CYP51 je najbolj ohranjen citokrom P450 (CYP) in sodeluje pri biosintezi holesterola v jetrih. Drugi encimi CYP v jetrih sodelujejo pri detoksifikaciji holesterola preko pretvorbe v žolčne kisline ali pri detoksifikaciji ksenobiotikov (DM). Da bi pridobili nova spoznanja o lastnostih lokusa gena *CYP51A1,* ki do sedaj ni bil povezan z boleznimi pri človeku, smo izvedli meta-analizo polimorfozmov gena *CYP51A1* v primerjavi z ostalimi CYPi v jetrih in geni, vključenimi v biosintezo holesterola. S holesterolom povezani geni so v splošnem manj polimorfni od DM CYP, z manj spremembami v kodirajočih delih, kar nakazuje na razlike v selekcijskem pritisku med potmi holesterola in presnovo zdravil. Med jetrnimi CYPi, vključenimi v študijo, ima gen *CYP51A1* najmanjše število kodirajočih variant in manj pogostih variant kot povprečje genov sinteze holesterola. Ker znotraj gena *CYP51A1* nismo uspeli zaznati ostalih funkcionalnih molekul (npr. lincRNA ali miRNA), smo pregledali celoten lokus gena. Odkrili smo, da se zaporedje *AL133568* prekriva s promotorsko regijo gena *CYP51A1*. Postavili smo hipotezo, da bi prepis AL133568 lahko imel vlogo pri uravnavanju izražanja gena *CYP51A1*, a tega nismo mogli eksperimentalno potrditi. Razlog za nizko populacijsko raznolikost človeškega gena *CYP51A1* tako še vedno ostaja nepojasnjen.

Supplementary Information

Table 1 The list of analyzed genes including transcript variants, length or regions and the number of SNPs obtained from UCSC Genome Browser for each of the gene. The letter numbering (A-E) refers to the listed datasets in materials and methods section.

Pathway	Gene	Region	Transcript variant	Analyzed region length [bp]	Protein size [aa]	A All variants	B Common	C Coding variants	D Non-synonymous mutations	E Mouse/Human Protein Identity
Cholesterol synthesis pathway	*HMGCR*	chr5:74632993-74657926	NM_000859	24934	888	411	92	44	23	93%
	MVK	chr12:110011500-110035071	NM_000431	23572	396	542	109	117	75	81%
	FDFT1	chr8:11660190-11696818	NM_004462	36629	417	1309	398	77	45	89%
	SQLE	chr8:126010720-126034525	NM_003129	23806	574	414	126	25	14	84%
	LSS	chr21:47608360-47648738	NM_002340	40379	732	883	219	94	53	86%
	**CYP51*	chr7:91741463-91764059	*NM_000786	22597	509	340	60	47	32	91%
	LBR	chr1:225589204-225616557	NM_002296	27354	615	493	100	89	66	79%
	SC4MOL	chr4:166248818-166264314	NM_006745	15497	293	318	120	20	15	89%
	NSDHL	chrX:151999511-152037907	NM_015922	38397	373	626	121	55	39	83%
	HSD17B7	chr1:162760496-162782608	NM_016371	22113	341	465	114	25	14	79%
	SC5D	chr11:121163388-121184119	NM_006918	20732	299	288	36	30	23	84%
	EBP	chrX:48380164-48387104	NM_006579	6941	230	90	14	28	21	78%
	DHCR7	chr11:71145457-71159477	NM_001360	14021	475	367	55	117	79	88%
	DHCR24	chr1:55315300-55352921	NM_014762	37622	516	789	216	64	42	97%
Bile acid synthesis pathway	*HSD3b7*	chr16:30996519-31000473	NM_025193	3955	369	93	9	52	36	87%
	CYP7A1	chr8:59402737-59412720	NM_000780	9984	504	199	42	59	42	82%
	CYP8B1	chr3:42913684-42917633	NM_004391	3950	501	100	17	58	33	75%
	CYP7B1	chr8:65508529-65711348	NM_004820	202820	506	3035	585	46	33	67%
	CYP27A1	chr2:219646472-219680016	NM_000780	33545	531	561	129	92	64	74%
	CYP46A1	chr14:100150755-100193638	NM_006668	42884	500	770	187	33	16	95%
	CYP39A1	chr6:46517445-46620523	NM_016593	103079	469	1790	433	54	43	75%
Drug Metabolizers	*CYP1A2*	chr15:75041184-75048941	NM_000761	7758	516	240	23	99	74	73%
	CYP2C9	chr10:96698415-96749148	NM_000771	50734	490	1440	264	97	62	74%
	CYP2C19	chr10:96522463-96612671	NM_000769	90209	490	2351	379	108	78	76%
	CYP2D6	chr22:42522501-42526883	NM_000106	4383	497	329	68	178	92	71%
	CYP2E1	chr10:135340867-135352620	NM_000773	11754	493	335	88	73	50	78%
	CYP3A4	chr7:99354583-99381811	NM_017460	27229	503	578	73	85	62	73%
GENOME		N/A	N/A	3137161264	N/A	56248699	13894623	1922594	701454	N/A
ENCODE		N/A	N/A	29955196	N/A	572024	139975	44716	15985	N/A
Cholesterol synthesis (Average)		N/A	N/A	25328	476	524	127	59	39	86%
Bile Acid Synthesis (Average)		N/A	N/A	57174	483	935	200	56	38	79%
Drug Metabolizers (Average)		N/A	N/A	32011	498	879	149	107	70	74%
BA liver CYPs (Average) Excluding *CYP46A1*		N/A	N/A	70676	502	1137	241	62	43	75%

*The analyzed region for *CYP51A1* gene was extended for 219 bp to cover the untranslated region of transcript variant 2 that overlaps with AL133568. The determination whether SNP is coding was done in respect of the NM_000786 transcript.

Table 2 Nucleotide polymorphism, proportion of polymorphic loci in analyzed regions and proportion of CODING and NON-Synonymous loci in respect of protein length

Gene	Analyzed region lenght [bp]	Protein size [aa]	F Nucleotide Polymorphism (All/gene length kb)	G Proportion of polymorphic loci (Common/ gene length)	H Coding Variants/ amino acid	I Non-synonymous Variants/ amino acids
HMGCR	24934	888	16.5	3.7	0.050	0.026
MVK	23572	396	23.0	4.6	0.295	0.189
FDFT1	36629	417	35.7	10.9	0.185	0.108
SQLE	23,806	574	17.4	5.3	0.044	0.024
LSS	40379	732	21.9	5.4	0.128	0.072
CYP51	**22597**	**509**	**15.0**	**2.7**	**0.092**	**0.063**
LBR	27354	615	18.0	3.7	0.145	0.107
SC4MOL	15497	293	20.5	7.7	0.068	0.051
NSDHL	38397	373	16.3	3.2	0.147	0.105
HSD17B7	22113	341	21.0	5.2	0.073	0.041
SC5D	20732	299	13.9	1.7	0.100	0.077
EBP	6941	230	13.0	2.0	0.122	0.091
DHCR7	14021	475	26.2	3.9	0.246	0.166
DHCR24	37622	516	21.0	5.7	0.124	0.081
HSD3b7	3955	369	23.5	2.3	0.141	0.098
CYP7A1	9984	504	19.9	4.2	0.117	0.083
CYP8B1	3950	501	25.3	4.3	0.116	0.066
CYP7B1	202820	506	15.0	2.9	0.091	0.065
CYP27A1	33545	531	16.7	3.8	0.173	0.121
CYP46A1	42884	500	18.0	4.4	0.066	0.032
CYP39A1	103079	469	17.4	4.2	0.115	0.092
CYP1A2	7758	516	30.9	3.0	0.192	0.143
CYP2C9	50734	490	28.4	5.2	0.198	0.127
CYP2C19	90209	490	26.1	4.2	0.220	0.159
CYP2D6	4383	497	75.1	15.5	0.358	0.185
CYP2E1	11754	493	28.5	7.5	0.148	0.101
CYP3A4	27229	503	21.2	2.7	0.169	0.123
GENOME	3137161264	N/A	17.9	4.4	N/A	N/A
ENCODE	29955196	N/A	19.1	4.7	N/A	N/A
Cholesterol synthesis (Average)	25328	476	20.0	4.7	0.130	0.086
Bile Acid Synthesis (Average)	57174	483	19.4	3.7	0.117	0.079
Drug Metabolizers (Average)	32011	498	35.0	6.3	0.214	0.140
BA liver CYPs (Average) (Excluding *CYP46A1*)	70676	502	18.9	3.9	0.122	0.085

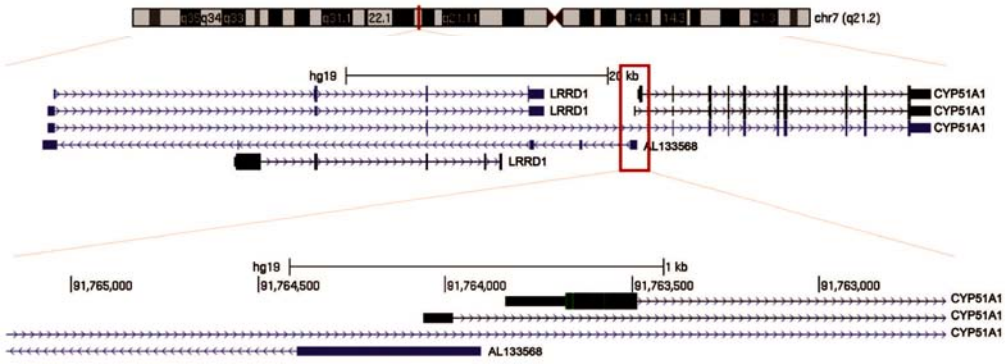

Figure 1 The *CYP51A1* 5' untranslated region with overlap with *AL133568* mRNA transcript