OPEN ACCESS Freely available online

PLoS one

Cis-Acting Polymorphisms Affect Complex Traits through Modifications of MicroRNA Regulation Pathways

Matthias Arnold[1☯], Daniel C. Ellwanger[1,2☯], Mara L. Hartsperger[1], Arne Pfeufer[3,4,5], Volker Stümpflen[1]*

1 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany, **2** Chair of Genome-Oriented Bioinformatics, Technische Universität München, Center of Life and Food Science, Freising-Weihenstephan, Germany, **3** Institute for Human Genetics, Technische Universität München, Munich, Germany, **4** Institute of Human Genetics, Helmholtz Zentrum München, German National Research Center for Environmental Health, Neuherberg, Germany, **5** Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Bolzano, Italy - Affiliated Institute of the University Lübeck, Germany

Abstract

Genome-wide association studies (GWAS) have become an effective tool to map genes and regions contributing to multifactorial human diseases and traits. A comparably small number of variants identified by GWAS are known to have a direct effect on protein structure whereas the majority of variants is thought to exert their moderate influences on the phenotype through regulatory changes in mRNA expression. MicroRNAs (miRNAs) have been identified as powerful posttranscriptional regulators of mRNAs. Binding to their target sites, which are mostly located within the 3′-untranslated region (3′-UTR) of mRNA transcripts, they modulate mRNA expression and stability. Until today almost all human mRNA transcripts are known to harbor at least one miRNA target site with an average of over 20 miRNA target sites per transcript. Among 5,101 GWAS-identified sentinel single nucleotide polymorphisms (SNPs) that correspond to 18,884 SNPs in linkage disequilibrium (LD) with the sentinels ($r^2 \geq 0.8$) we identified a significant overrepresentation of SNPs that affect the 3′-UTR of genes (OR = 2.33, 95% CI = 2.12–2.57, $P < 10^{-52}$). This effect was even stronger considering all SNPs in one LD bin a single signal (OR = 4.27, 95% CI = 3.84–4.74, $P < 10^{-114}$). Based on crosslinking immunoprecipitation data we identified four mechanisms affecting miRNA regulation by 3′-UTR mutations: (i) deletion or (ii) creation of miRNA recognition elements within validated RNA-induced silencing complex binding sites, (iii) alteration of 3′-UTR splicing leading to a loss of binding sites, and (iv) change of binding affinity due to modifications of 3′-UTR folding. We annotated 53 SNPs of a total of 288 trait-associated 3′-UTR SNPs as mediating at least one of these mechanisms. Using a qualitative systems biology approach, we demonstrate how our findings can be used to support biological interpretation of GWAS results as well as to provide new experimentally testable hypotheses.

Citation: Arnold M, Ellwanger DC, Hartsperger ML, Pfeufer A, Stümpflen V (2012) *Cis*-Acting Polymorphisms Affect Complex Traits through Modifications of MicroRNA Regulation Pathways. PLoS ONE 7(5): e36694. doi:10.1371/journal.pone.0036694

Editor: Chunyu Liu, University of Illinois at Chicago, United States of America

Received October 26, 2011; **Accepted** April 5, 2012; **Published** May 11, 2012

Funding: This work was supported by the Helmholtz Alliance on Systems Biology (project CoReNe) and the Federal Ministry of Education and Research (BMBF) in its MedSys initiative (project SysMBo, FKZ: 0315494A). It was also supported by grants by the British Foundation for the study of infant deaths FSID, Grant No. 261 (Pfeufer) and the German National Genome Research Network NGFN 01GR0803 and BMBF 01EZ0874 (Pfeufer). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: v.stuempflen@helmholtz-muenchen.de

☯ These authors contributed equally to this work

Introduction

The state of knowledge regarding development, progression and inheritability of human diseases has long since outrun the classical understanding. There are only few disorders where monogenic causes could be determined. Also, the genetic links to disease incident are missing to a substantial part and the complexity of pathways encompassing pathogenic effects can still not be limited upwards. In recent years, more and more evidence is emerging that microRNAs (miRNAs), a class of small non-coding RNAs, play an important role in human traits. Databases collecting information on miRNAs mediating human disease such as miR2Disease [1] or PhenomiR [2] list several hundred miRNAs with proven roles in way above 100 human diseases.

MiRNAs are key posttranscriptional regulators of most known cellular processes and have been associated with cell fate decision, development, and stress response. Additionally, miRNAs have been identified to be usable as biomarkers for human diseases [2–5]. With growing knowledge on their targets, which are believed to make up more than 60% of all protein-coding genes [7], new regulatory and disease-mediating gene networks were discovered [8–11]. Because of the ability of single miRNAs to regulate not only one but up to several hundred genes, they depict a promising drug target for disease pathways involving multiple genes. With the recent advances of the crosslinking immunoprecipitation (CLIP) technology, it has become feasible to experimentally determine miRNA-target interactions and the exact binding sites of the RNA-binding proteins (RBPs) on transcriptome scale [12–15].

As soon as the importance of miRNA functioning for human health was realized, approaches were undertaken with the objective to identify potential interrelations of miRNA dysregulation and genetic variation. However, until now neither the data on

trait-associated polymorphisms nor experimentally verified miRNA targeting information provided a sufficient basis for such analyses. For instance, mutations in the 3′-UTR, the major target of miRNA-mediated regulation, have long been (and are still) neglected for the most part in association studies. Therefore, only few particular cases of polymorphisms affecting miRNA regulation pathways have been identified, yet [16,17]. Up to now such studies are often limited to effects on (mostly predicted) miRNA target sites [17,18]. However, the 3′-UTR harbors several other functional elements which may, if affected, also mediate disruption of miRNA regulation pathways. It has been assumed, for instance, that the loss of a polyadenylation (poly(A)) signal can cause genetic diseases by non-specific degradation of the mRNA [19]. Recent experiments suggest that this effect may be based on a functional correlation of poly(A) signal efficiency and miRNA-mediated repression [20]. Further, the structural accessibility of an RNA region is an important feature for the binding affinity of RNA-induced silencing complex (RISC) target sites [21]. It has been shown that mutations in RNAs have large local as well as global structural effects [22] and that altered target accessibility can reduce miRNA-mediated posttranscriptional repression to a scale comparable to that of mutations disrupting miRNA recognition element (MRE) sequence complementarity [23]. Finally, polymorphisms affecting splice sites can lead to radical sequence changes increasing susceptibility to diseases, an effect which is suspected to be partly due to altered translation efficiency of the affected mRNA [24] - which is characteristic for miRNA functioning.

The success of genome-wide association studies (GWAS) in determining the genetic causes of some common diseases led to an immense increase of the efforts put on screening common alleles for disease involvement. Since then, more and more loci associated with trait susceptibility are detected. GWAS can, however, neither identify causal genes in associated loci nor provide functional mechanisms behind observed association signals. Consequently, many identified GWAS signals are awaiting mechanistic characterization and the rate at which GWAS signals are currently discovered necessitates systematic and scalable functional approaches [25].

When it became clear that only few diseases are due to common risk alleles, sequencing endeavors have been performed to identify putatively rare variants hiding behind the signals detected in GWAS. In some cases these approaches successfully identified rare causative variants predisposing to disease. However, the expected breakthrough failed to appear as in several cases sequencing of risk loci did not provide further knowledge [26–28]. With the situation additionally compounded by the fact that most GWAS-identified alleles are located within non-coding regions, now one focus is laid on the identification of regulatory variants. In this work, we concentrate on the influence of GWAS-identified variants on miRNA-mediated *cis*-acting regulatory effects.

We assess the potential impact of trait-associated SNPs on miRNA regulation pathways using publicly available GWAS data [29,30]. We describe potential posttranscriptional effects of SNPs by systematically investigating mutations within the 3′-UTR of human transcripts for interference with poly(A) signals, 3′-UTR splicing, 3′-UTR secondary structure changes and MREs. Using the example of *rs10923* in the 3′-UTR of *SMC4* we show how our findings can be utilized in the biological interpretation of GWAS results in a systems biological manner.

Results

Trait-associated Variants are Significantly Overrepresented in the 3′-UTR

We compared the amount of trait-associated variants within the predefined five function classes of SNPs (intergenic, intronic, 5′-untranslated region (5′-UTR), coding sequence (CDS), and 3′-UTR) to examine a potential location bias of these markers. Of 18,884 SNPs contained in the extended GWAS-SNP set (GWAS-SNPs and their highly correlating LD partners, $r^2 \geq 0.8$), we found 436 to be located in the 3′-UTR of 326 human genes (OR = 2.331, $P < 10^{-52}$, referred to as 3′-UTR SNPs). This is a higher enrichment than for sentinel SNPs only (OR = 2.059, $P < 10^{-10}$). We calculated the probability to get an 3′-UTR enrichment this strong in a random subset of HapMap-SNPs [31–33] of comparable size which confirmed significance ($P < 1.1 \cdot 10^{-7}$). For further validation of this enrichment we looked at the dependencies between the OR and the minor allele frequency (MAF) as well as different r^2 thresholds.

When we adjusted for r^2 in the extension of the GWAS-SNPs, we found that the distribution of ORs locally stabilizes around a threshold of $r^2 = 0.8$ (**Figure 1A**). This limit thus seems to fit the data better than more rigid thresholds and was therefore chosen in this work. To rule out a false increase of the enrichment due to correlating SNPs in the same 3′-UTRs (SNP-gene ratio ~ 1.34) we binned the complete HapMap-SNPs into blocks with an all-vs.-all $r^2 \geq 0.8$. More than one million HapMap SNPs were binned together in about 371,000 blocks containing more than two SNPs. The remaining SNPs only showed pairwise or no LD at the chosen threshold. When we included all SNPs after binning, the OR for 3′-UTR enrichment was even greater than without binning (OR = 4.27, 95% CI = 3.84–4.74, $p < 10^{-114}$). Considering only the SNPs within the $\sim 371{,}000$ blocks the enrichment still holds significance (OR = 1.828, 95% CI = 1.63–2.04, $p < 10^{-24}$). As about 10% of the 3′-UTR SNPs are not contained in theses blocks, we suggest that this value presents an underestimation of the actual enrichment. The reason for the stronger enrichment after binning is that the SNP count within the LD blocks depends on the location of the SNPs. While intronic and intergenic SNPs are reduced to less than 35% (block-SNP ratio) by binning, SNPs in exonic regions present less extensive LD patterns (reduction only to about 81%).

By analyzing the MAF, we found the extended GWAS-SNPs to hold a commonly higher MAF than the HapMap-SNPs, regardless of their chromosomal location. However, the comparison of the MAF distribution of the 3′-UTR SNPs to the MAF distribution of other extended GWAS-SNPs revealed a slight trend of 3′-UTR SNPs towards moderate MAF frequencies between 0.1 and 0.4. This trend becomes more explicit when comparing the 3′-UTR SNPs to the combined extended GWAS-SNPs in the other two exonic regions (i.e. 5′-UTR and CDS). In comparison, 3′-UTR SNPs show underrepresentation of the intervals 0.0–0.1 (OR = 0.88), 0.2–0.3 (OR = 0.70) and 0.4–0.5 (OR = 0.78) whereas the other two intervals are significantly ($P < 0.05$) overrepresented ($OR_{0.1–0.2} = 1.40$ and $OR_{0.1–0.2} = 1.59$). To investigate if 3′-UTR SNP enrichment values hold only for specific MAFs, we recalculated the ORs against the HapMap-SNP set in dependency of the MAF. The ORs resulting for the five MAF intervals follow roughly the pattern of over−/underrepresentation observed in the comparison with the other extended GWAS-SNPs but never lose significance or fall below an OR of 2.0 (**Figure 1B**).

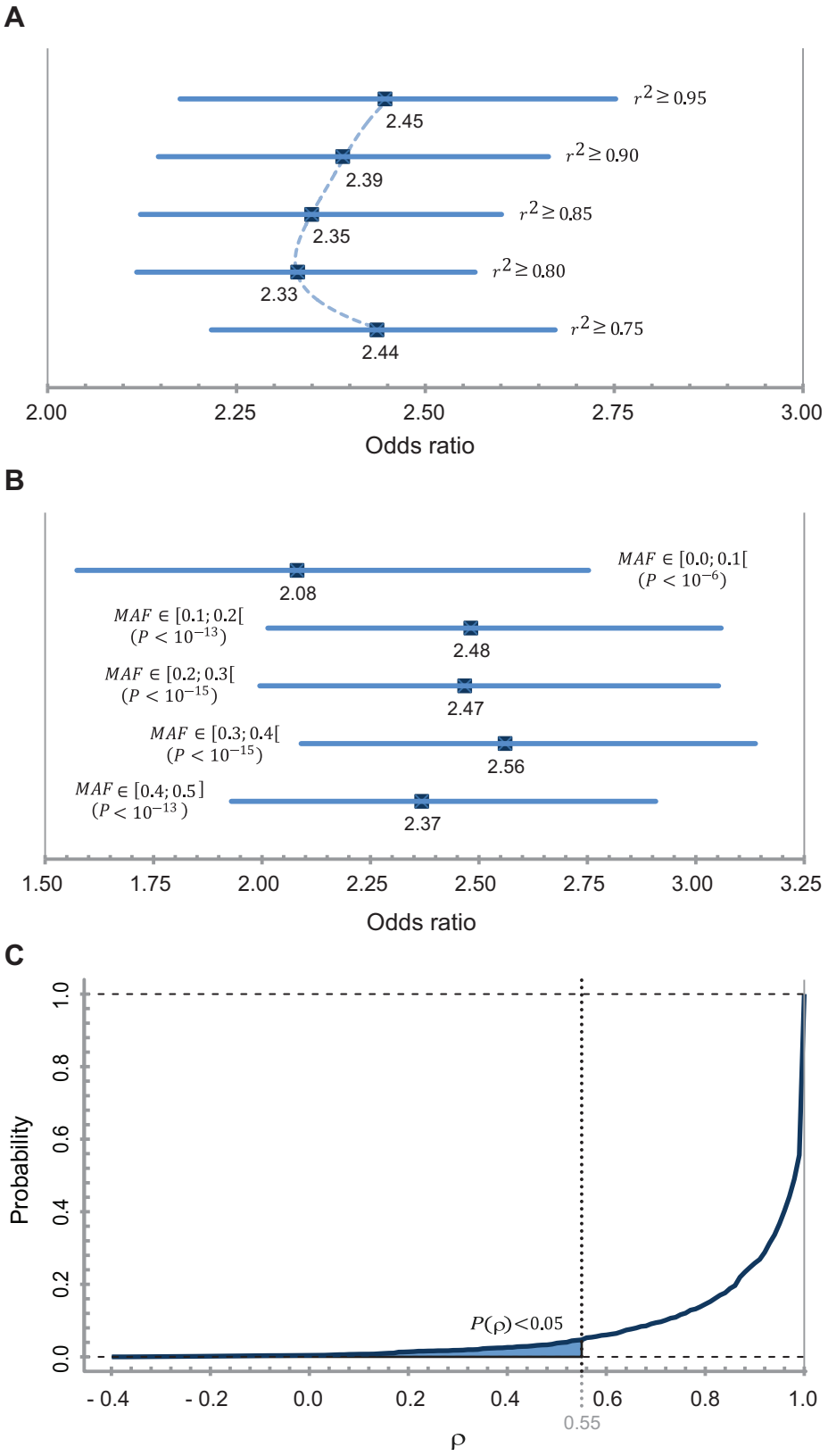

Figure 1. Statistical analysis of 3′-UTR enrichment values and determination of the folding correlation coefficient threshold. A: SNP enrichment in the 3′-UTR in dependency of different LD thresholds. Displayed are the ORs and confidence intervals for five cut-offs. Accumulative 3′-UTR SNP sets were calculated. The fitted distribution (dashed line) points out the stabilization of the OR around a threshold of 0.8. **B:** SNP enrichment in the 3′-UTR in dependency of the minor allele frequency. Displayed are the ORs and confidence intervals for the 5 different MAF bins. SNP counts were compared within the respective bins. **C:** Probability distribution of correlation coefficients (*p*) between wild-type and mutated structures

of RBP-binding regions. Below a cut-off for the correlation coefficient of 0.55 (displayed in gray) the probability to observe a change of RNA secondary structure of this scale by chance amounts to less than 5%.
doi:10.1371/journal.pone.0036694.g001

Enrichment Analysis Indicates Gene Involvement in Lipid Metabolism

To investigate whether the 326 genes affected by 3′-UTR mutations share common characteristics, we performed gene enrichment analyses with respect to disease involvement and functional annotations.

By mapping the traits associated with 3′-UTR variants to MeSH terms we retrieved a total of 49 observed disease classes. The most abundant categories were immune system diseases, mental disorders, digestive system diseases, nervous system diseases, and neoplasms. The distribution of the 3′-UTR SNPs within these disease classes showed no significant overrepresentation compared to the count of studies performed for the single disorders in the GWAS Catalog. When comparing the number of 3′-UTR SNPs per disease to the count of all non-3′-UTR extended GWAS-SNPs, we found only lipid concentrations to be significantly ($P<1.3\cdot10^{-3}$) enriched in the 3′-UTR set.

Gene set enrichment analysis (GSEA) revealed only four significantly enriched ($P<0.05$ after Bonferroni correction) functional annotations in this set: lipid metabolism, axon growth, activation of the immune response/inflammation (9 terms), and regulation of/response to cell signaling (10 terms). Using DAVID [34,35], we also checked for overrepresentation of disease terms not limited to the GWAS Catalog and found three enriched terms ($P<0.05$ after correction): Dyslipidemia (background set: Online Mendelian Inheritance in Man database), neurological diseases, and infections (background set: Genetic Association Database [36]).

Evidence for Impact on miRNA-mediated Regulation

The efficacy of a miRNA to control target mRNA translation relies, among others, on three sequence-based features: correct mRNA processing, presence of a functional MRE, and accessibility of the RISC binding site. To find out to which extent trait-associated SNPs in 3′-UTRs affect miRNA functioning, we examined four mechanisms potentially compromising these features (**Figure 2, Table 1**. This analysis was limited to transcripts featuring both 3′-UTR SNPs and validated RISC binding sites. The according data set contained 288 SNPs on 409 transcripts and 219 genes, respectively. Firstly, we investigated potential effects of SNPs on mRNA processing by interfering with poly(A) signals. We found four SNPs affecting hexamers with a sequence characteristic for poly(A) signals, however, none of these hexamers was located near a validated poly(A) site. A functional effect of those variants on mRNA processing thus seems unlikely. Secondly, we analyzed the impact on mRNA splicing. We identified seven SNPs (~2.4%) predicted to interfere with RNA splice sites (**Figure 2D, Table S1**). Six of those are predicted to create new acceptor sites and one to create a new donor site (**Figure 2D**). SNPs interfering with splice sites located at an exon/intron or intron/exon border as annotated in RefSeq were not found. The probability to observe such an effect by chance is $P=1.78\cdot10^{-2}$ for acceptor sites and $P=1.41\cdot10^{-2}$ for donor sites. In all seven cases, the predicted gain of splice sites results in exon shortening, leading to a noticeable loss (46% on average) of RISC binding sites in the accordant transcripts. Thirdly, we searched for SNPs which may affect the secondary structure of the 3′-UTR proximal to a validated RISC binding site causing an altered accessibility of the region. This resulted in 14 SNPs (~4.9%) predicted to affect the binding affinity of the RISC through changed secondary structure of the 3′-UTR (**Figure 2B, Table S2**). Fourthly, we examined direct effects of SNPs on MREs located in validated RISC binding sites. We found 22 SNPs (~7.6%) disrupting MREs (**Figure 2C, Table S3**), and 28 SNPs (~9.7%) creating new MREs (**Figure 2C, Table S4**). The overlap between the SNP sets creating and disrupting MREs, i.e. SNPs substituting the MRE of one miRNA by a MRE of another miRNA, amounts to 13 variants. Accordingly, a total of 37 unique SNPs (~12.8%) directly affect MREs. The probability of obtaining these amounts of SNPs affecting MREs randomly was $P=1.27\cdot10^{-2}$ (disruption) and $P=8.76\cdot10^{-4}$ (creation). Additionally, we found that only 11% of SNPs enhancing (i.e. extending an already existing seed match) or creating a MRE were conserved across mammals which was a lower fraction than for SNPs causing one of the other effects (folding = 29%, splicing = 29%, MRE disruption = 27%).

SMC4 – from Primary Biliary Cirrhosis to Cancer

The autoimmune disease primary biliary cirrhosis (PBC) is associated with the damaging of the small bile ducts and is mediated by auto-antibodies [37–39]. The autoimmune response caused by those antibodies leads to inflammation followed by aggregation of dead cells. Apoptosis is induced, among other things, by reactive molecules effecting DNA damage [40,41] leading to a build-up of scar tissue (i.e. cirrhosis). The genetic background of the disease was focus of a recent GWAS [42], however, the rationale of the study was restricted to the major histocompatibility complex and interleukins. Causative variants could not be identified so far. One of the SNPs (*rs4679904*) reported in the study is in high LD ($r^2=0.86$) with the 3′-UTR SNP *rs10923* in *SMC4* which is part of the *Condensin I* complex [43]. Interestingly, DNA repair genes such as *PARP1* and *XRCC1* are over-expressed in cirrhotic tissue and are in this context hypothesized to feature pathogenic effects [44]. For full functioning of the *PARP1-XRCC1* complex in single-strand break repair, an association with the *Condensin I* complex is established [45].

The SNP *rs10923* is localized within an experimentally validated RISC binding site [14] and lies directly in the seed complementary region of *hsa-mir-299-5p*, a miRNA that has been shown to be up-regulated in PBC patients [46]. The minor G allele of *rs10923* disrupts the 6*mer*γ seed complementary region (**Table 1**) and thus the ability of the miRNA to bind the transcript (**Figure 3**) [47]. Although we also observe the gain of a new MRE when introducing the minor allele of the SNP (**Table 1**), expression data from the MuTHER study [48] supports the hypothesis of deactivated miRNA-control as it shows significant ($P_{combined}<5.9\cdot10^{-4}$) association of the G allele with increased *SMC4* expression in lymphoblastoid cell lines. We suggest that, by this process, *rs10923* contributes to the phenomenon of DNA repair perturbation in cirrhosis (**Figure 4**) [44]. Beyond that this may indicate a rate-limiting character of *SMC4* in the generation of the *Condensin I-PARP1-XRCC1* complex. Furthermore, PBC patients present elevated risk to develop different types of cancer [49,50] representing a potential explanation for the deranged DNA repair functionality in the disease. In cancer development and progression, up-regulated DNA damage response is associated with mutagenesis and resistance to radio- and chemotherapy [41,51–53].

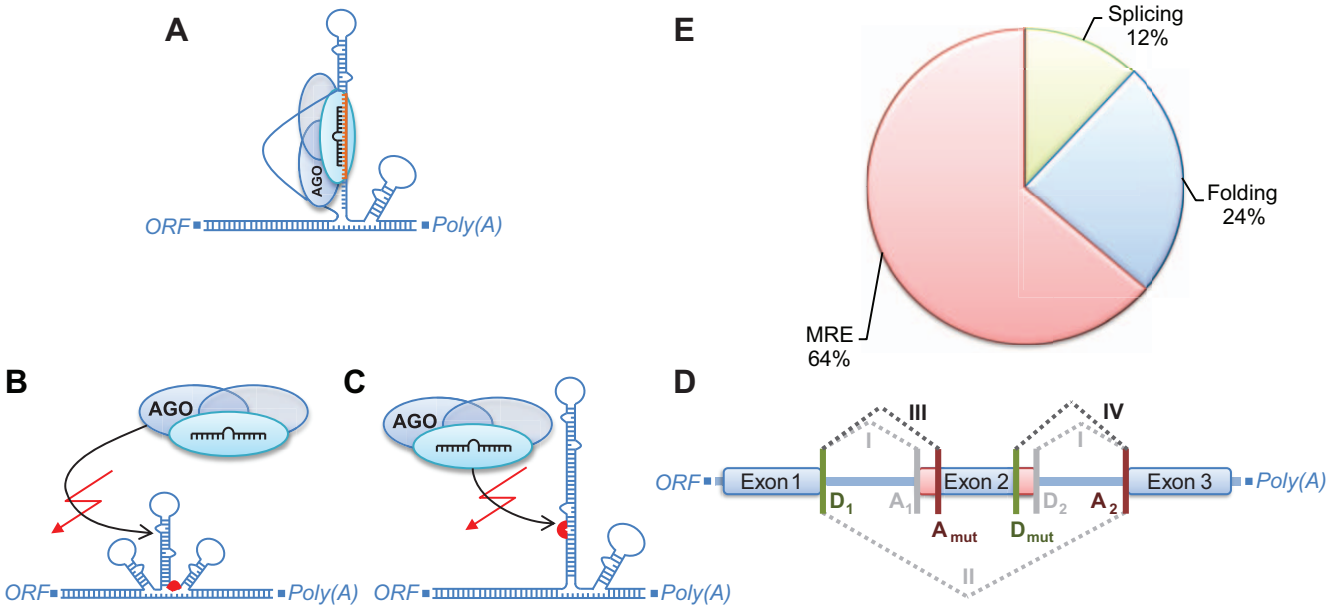

Figure 2. Mechanisms mediated by 3′-UTR SNPs affecting miRNA targeting. A: Regular binding of the RISC to the target mRNA. **B:** Binding of the RISC and, thus, miRNA-mediated silencing is inhibited by a change in RNA secondary structure. **C:** A mutation within the MRE seed site disrupts the ability of a certain miRNA to target a transcript. Here, the opposite effect also applies, i.e. a new MRE seed site is formed by a polymorphism which enables targeting by a miRNA usually not controlling the respective transcript. **D:** Altered splicing by acceptor or donor splice site gain. The existing splice variants (I and II, grayed) are extended by mutationally introduced additional splice variants: (III) A present acceptor site (A_1) is substituted by a new acceptor site (A_{mut}), and (IV) a naturally occurring donor site (D_2) is replaced by a new donor site (D_{mut}). Both effects may lead to a considerable loss of exon sequence (displayed in red) and, thus, RISC binding sites. **E:** The percentages of classified SNPs mediating the single mechanisms. The greatest amount of functionally annotated 3′-UTR SNPs directly affect MRE sequences, followed by SNPs changing the RNA secondary structure and SNPs with an predicted effect on 3′-UTR splicing.
doi:10.1371/journal.pone.0036694.g002

Discussion

Although thousands of SNP-trait associations have been published and the applicability of GWAS is broadly appreciated, the associations identified by GWAS hold predictive power for only a small fraction of disease incidence so far [25,54]. Finding the causative variants predisposing individual disease risk therefore presents the main dilemma of GWAS. As long as comprehensive sequencing data is missing for most associated loci, frameworks are needed providing a first characterization of potential disease-mediating mechanisms.

Here, we suggest that associated common or unknown rare variants in the 3′-UTR depict posttranscriptionally regulatory variants which, to differing extent, affect disease development and progression by interfering with miRNA regulatory networks. The differing impact of 3′-UTR SNPs can be explained with two facts: first, human mRNAs are mostly targeted by multiple miRNAs at different target sites which may allow for compensating the loss of a single binding site; and second, the extent to which mRNAs are silenced by miRNAs depends on multiple factors such as tissue-specific miRNA expression, the binding affinity of the RISC and the degradation rate of the mRNA [55,56]. Thus, unlike mutationally driven alterations of the amino acid sequence of a protein or disruption of a transcription factor binding site in promoter or enhancer regions, interferences with miRNA regulation may show tissue-specific effects sizing from low or not distinguishable to high or even causative.

Based on recently available data, our results clearly show that several lines of evidence become obvious that miRNAs play an important role in genetically determined posttranscriptional disease development and progression. There is no evidence for considerable direct interference of miRNA processing: only one SNP (*rs2168518*) was found to be located in the hairpin of *hsa-mir-4513*. But we observed a significant overrepresentation of trait-associated SNPs in the 3′-UTR that strongly suggests a functional coherence between genetic variants and miRNA regulation pathways on the *cis*-regulatory level.

Several studies assessing the disease-mediating power of miRNAs confirmed their involvement in the mediation of diverse traits [3,57,58] suggesting that traits associated with 3′-UTR SNPs belong to heterogeneous disease classes. We observed a similar trend as our results showed a correlation between the number of traits belonging to the distinct disease classes and the number of published studies on the corresponding traits in the GWAS catalog [30]. This indicates that with an increasing number of studies on other traits the count of disease classes associated with 3′-UTR SNPs will increase, too. Moreover, an uniform number of studies on the traits should lead to an equal distribution of the represented disease classes. Of a total of 49 disease classes assigned to the 3′-UTR SNPs, only lipid concentrations showed a significant enrichment within the phenotypes investigated by GWAS.

The functional enrichment analysis revealed that genes having a 3′-UTR SNP are involved in a wide range of cellular processes. The most intriguing finding is the replication of the enrichment of lipid traits by GSEA which, additionally, detected an enrichment of the lipid metabolism on the level of gene function. The enrichment of terms regarding the regulation of the immune response and of inflammatory processes, on the contrary, reflects the phenotype bias of GWAS towards autoimmune and inflammatory diseases [59]. The other overrepresented terms are connected to multiple downstream effects indicating a major participation of the genes in the initiation and regulation of cellular pathways.

Table 1. SNPs affecting functional elements with *cis*-regulatory effects on miRNA regulation.

SNP	Gene	Effects	Traits
rs1121	PDXDC1	MRE creation	Height
rs4564	DLD	MRE disruption	Ulcerative Colitis
rs6706	TRIP6	MRE disruption	Resting Heart Rate
rs7089	TMUB2	MRE disruption; MRE creation	Bone Density
rs7097	POLR1D	MRE creation	Large B-Cell Lymphoma
rs7118	ZFP90	MRE disruption; MRE creation	Ulcerative Colitis
rs7119	HMG20A	MRE disruption	Type 2 Diabetes
rs7371	GNAI3	Acceptor gain	Major Depressive Disorder
rs7444	UBE2L3	Folding	Crohn's Disease; Systemic Lupus Erythematosus
rs8523	ELOVL2	MRE disruption; MRE creation	Phospholipid levels
rs9253	MEAF6	MRE disruption	Hematological Phenotypes
rs9927	PYGB	MRE creation	Liver Enzyme Levels
rs10923	SMC4	MRE disruption; MRE creation	PBC
rs11700	E2F4	MRE creation	Coronary Heart Disease
rs12439	CLIC4	MRE disruption; MRE creation	Height
rs12916	HMGCR	MRE creation	Cholesterol levels; Metabolic Traits
rs12956	RYBP	Folding	Height
rs13099	TMED10	Folding	Height
rs42038	CDK6	Folding; Acceptor gain	Height
rs42039	CDK6	MRE creation	Rheumatoid Arthritis
rs232775	MYSM1	MRE creation	Diabetic Retinopathy
rs699779	NOTCH2	Acceptor gain; MRE disruption	Type 2 Diabetes
rs823136	RAB7L1	MRE creation	Parkinson's Disease
rs835575	NOTCH2	Folding; MRE disruption; MRE creation	Type 2 Diabetes
rs835576	NOTCH2	MRE disruption; MRE creation	Type 2 Diabetes
rs1045100	ATG16L1	MRE disruption; MRE creation	Crohn's Disease
rs1045407	ZNF678	Folding; MRE creation	Height
rs1046917	FN3KRP	Folding	Glycated Hemoglobin Levels
rs1047440	CEP120	MRE disruption; MRE creation	Body Mass Index
rs1058588	VAMP8	MRE disruption	Prostate Cancer
rs1379659	SLIT2	MRE disruption	Echocardiographic Traits
rs2032933	RMI2	MRE creation	Celiac Disease
rs2071518	NOV	MRE creation	Blood Pressure
rs2077579	DDX6	Folding	PBC
rs2229302	HOXB2	MRE disruption; MRE creation	Primary Tooth Development
rs2244967	VSTM4	Acceptor gain	Serum Uric Acid
rs2282301	RIT1	Folding	Conduct Disorder
rs2293578	SLC39A13	MRE creation	Body Mass Index
rs2564921	RFT1	Folding	Height
rs3816661	CD276	MRE disruption	Liver Enzyme Levels
rs3821301	TANC1	Folding	Sudden Cardiac Arrest
rs4770433	SACS	MRE disruption	Protein Quantitative Trait Loci
rs4819388	ICOSLG	Folding; MRE creation	Celiac Disease
rs4973768	SLC4A7	Donor gain	Breast Cancer
rs6722332	WDR12	Acceptor gain	Coronary Heart Disease; Myocardial Infarction
rs7350928	KIAA1267	MRE disruption; MRE creation	Parkinson's Disease
rs7528419	CELSR2	Acceptor gain	Cholesterol levels; Metabolic traits; Cardiovascular Disease; Myocardial Infarction; Response to Statins
rs8176751	ABO	MRE creation	Hematological Phenotypes

Table 1. Cont.

SNP	Gene	Effects	Traits
rs10892082	PAFAH1B2	Folding	Protein QTLs; Triglyceride Levels
rs11067231	MMAB	MRE creation	Cholesterol levels
rs11542478	FAM110C	Folding	Information Processing Speed
rs11713355	SLC6A6	MRE disruption; MRE creation	Cognitive Performance
rs17574361	KIAA1267	MRE disruption; MRE creation	Parkinson's Disease

The first column gives the rs-number of the SNPs, in the second column the HGNC symbol of the affected genes are listed and the third column describes the functional mechanisms which could be assigned to the SNPs. The last column contains all traits associated with the respective SNP.
doi:10.1371/journal.pone.0036694.t001

Despite the fact that we could not find an interrelation of polymorphisms and poly(A) signals, our extensive analysis of the potential impact of GWAS-identified SNPs on functional elements in the 3′-UTR revealed several mechanisms whereby variants may affect miRNA-mediated regulation. The smallest fraction of potentially functional 3′-UTR SNPs affects 3′-UTR splicing. These SNPs are predicted to mediate miRNA target site loss, mostly through the gain of acceptor splice sites (donor splice site gain only occurred once) resulting in shortened 3′-UTRs. That altered splicing is the rarest found mechanism can be explained by the huge impact it mediates on miRNA targeting which manifests in the high fraction of target site loss (46% on average) for the affected transcripts. The second mechanism is the SNP-mediated alteration of RNA secondary structure of a RISC binding region. The impact of RNA folding on the binding affinity of RBPs has already been described [21,23]. However, the extent to which this phenomenon translates into miRNAs mediating human disease development is unknown. With our results we provide a first data basis on potentially pathogenic RNA structural changes which may serve as a starting point to investigate this matter further. The third and most abundant mechanism we could identify is the direct alteration of MRE sequences. We find not only that GWAS-identified markers in the 3′-UTR show a significant enrichment within MREs, but also identify a novel scenario of how miRNA dysregulation may take effect: the substitution of the recognition element of one miRNA by one of another miRNA. While a disruption (or creation, respectively) of a MRE enables a rather straightforward rationale, that is the tissue-specific repression (or enhancement, respectively) of miRNA regulation, this scenario makes interpretation rather complex. Such a substitution may imply concurrent but simultaneously diverging effects in different tissues, depending on the respective expression patterns of the two miRNAs, possibly leading to systemic disturbances of several cell types. The overlap between the two sets of SNPs which disrupt and create MREs amounts to 13 polymorphisms and constitutes more than one third of the set of variants affecting MREs - which is a surprisingly high number. We believe that the transcripts affected by a SNP mediating this effect may present quite interesting targets for further studies. Moreover, this highest fraction (39%) of effective 3′-UTR SNPs shows a low conservation indicating that the creation of a MRE may be an abundant process of functional SNPs.

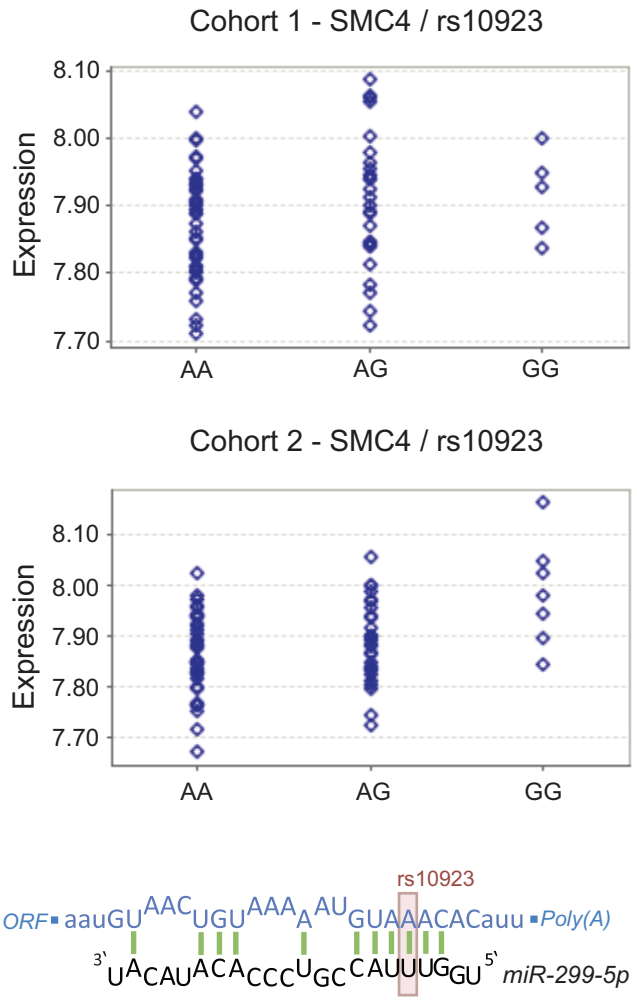

Figure 3. Impact of the SNP *rs10923* on miRNA-mediated repression of *SMC4*. Shown is the mRNA:miRNA duplex for the reference allele of *rs10923* (lower part). The minor allele of the SNP (position adumbrated by the light red box) disrupts the seed complementary region. In the upper part of the figure, the expression pattern of *SMC4* in lymphoblastoid cells is illustrated. The minor G allele of the polymorphism is significantly ($P_{combined} < 5.9 \cdot 10^{-4}$) linked to an increased abundance of *SMC4* transcript. For the illustration of expression values Genevar output was adapted [89].
doi:10.1371/journal.pone.0036694.g003

In general, our findings indicate that miRNAs play an important role in genetic variants causing trait development. Therefore, novel aspects not only for the interrelations in pathogenic disturbance of cellular processes, but also for the coherence of different traits can be addressed. For instance, differential tissue-specific expression patterns of miRNAs in combination with genetic variants may shed light on the still unknown functions driving the same cellular pathways to feature different effects in diseased and healthy individuals. Thus, closing the gap between trait-associated mutations and impaired miRNA-

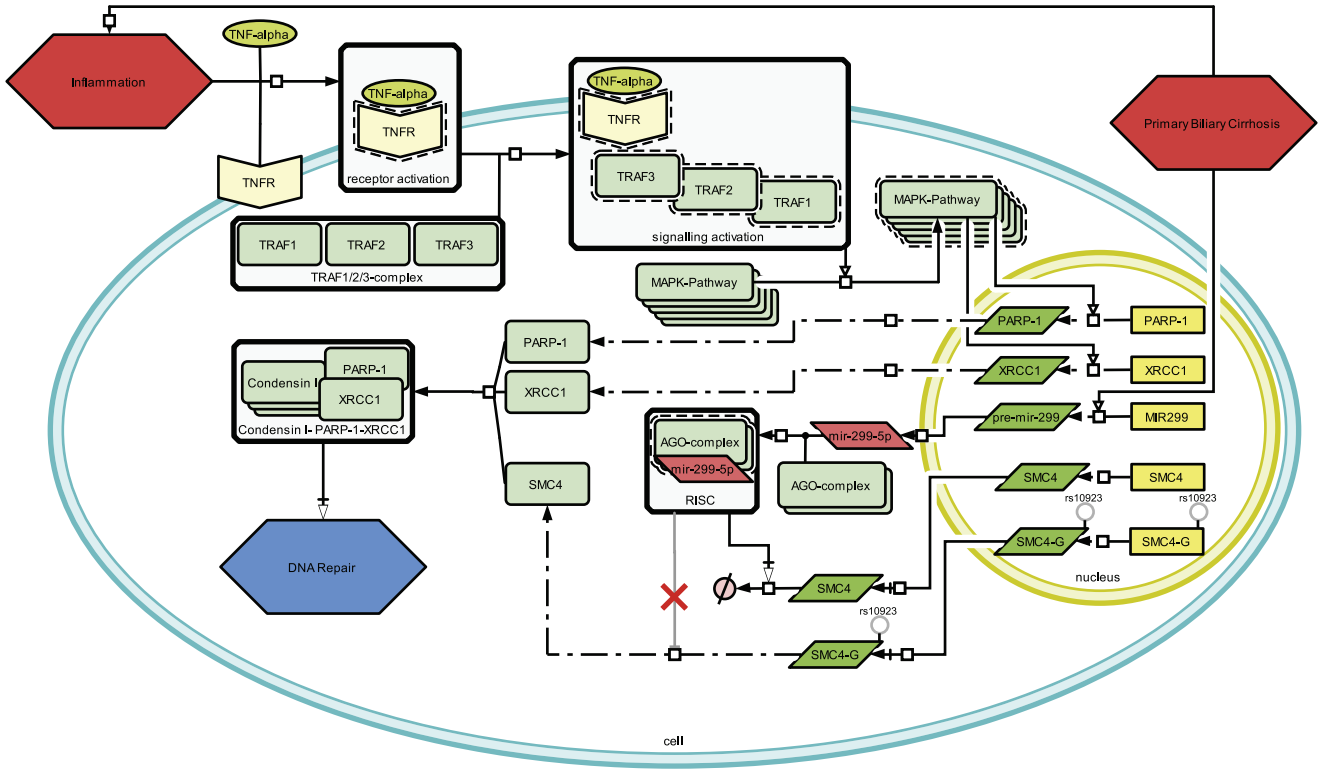

Figure 4. Impact model of mutated SMC4 in primary biliary cirrhosis. Inflammation follows the autoimmune response leading to the activation of the *MAPK*-pathway via signal molecules as e.g. *TNF-alpha*. Transcription factors activated as downstream effect of *MAPK* activation lead to over-expression of DNA repair genes. The in PBC over-expressed *hsa-mir-299-5p* is hypothesized to target *SMC4* at the seed complementary region where *rs10923* is located. With the major allele, *SMC4* is silenced, whereas the mutated *SMC4*-G cannot be bound by *hsa-mir-299-5p* and therefore is translated without interference. This results in the more frequent association of the *Condensin I-PARP1-XRCC1* complex contributing to disturbed DNA repair in cirrhosis tissue.
doi:10.1371/journal.pone.0036694.g004

mediated gene regulation may lead the way to a new understanding of the interconnection of those functional entities.

We found 326 candidate genes, and on the example of *SMC4* we show that the incorporation of miRNA related information can be used to construct models of potential disease-driving mechanisms in a systems biological fashion. In this case, we present a very simple one-gene/one-miRNA scenario. However, up to now *SMC4* and PBC have never been connected, as well as an association of PBC and cancer only exists in clinical ratios. But our suggestion of a highly active, *SMC4*-dependent DNA repair machinery is not only a likely cellular response mechanism to apoptosis signals in cirrhotic tissue. It also provides the link to cell transformation and cancerogenesis. Therefore, in the context of miRNA biology plausible coherences can be established for common variants with unknown effect. This also points out that, beyond statistical evaluations, case-specific detailed analyses may greatly outrange global approaches in the biological interpretation of GWAS results.

Our findings highlight numerous starting points for biomedical research. To advance the potential of conclusion drawing in GWAS results interpretation, it may be invaluable to overcome the GWAS bias towards the coding sequence and extend approaches based on 3′-UTR mutations. Conversely, more comprehensive data on miRNA binding sites might enhance our understanding of miRNA regulation functioning. It may also be of interest to investigate the most suspicious cases described in this work further. Moreover, determination of miRNA regulation via RISC binding to other mRNA locations than the 3′-UTR might shed light on the effect of coding-synonymous trait-associated alleles. The combination, i.e. extensive measurements of miRNA-mediated gene regulation in patients with traits for which a plausible model of miRNA involvement can be created in the context of associated mutations in the 3′-UTR, may further provide new perspectives of disease progression.

But the implications of our findings go beyond GWAS and miRNAs. Considering the advances made in the exploration of the still poorly understood elements of the genome, the impact of the presented results are pointed out. Just recently, the approaches of the modENCODE project [60,61] led to the validation of the hypothesized hierarchical structure of physical regulatory networks in eukaryotes which are based on a sophisticated interplay of miRNAs and transcription factors. Thus, the important role of this class of non-coding RNAs in the regulatory machinery of the cell is brought out on a large-scale level. If those findings can be transferred from the studied model organisms to human, the analysis of impairment of the transcription factor-miRNA network balance by mutationally altered target site functioning may lead to a completely new definition of genetically predisposed diseases on a RNA-mediated, regulatory basis. Also, miRNAs are only one class of non-coding RNAs which have been proven to feature regulatory power. Interference by e.g. piwi-associated RNAs, small interfering RNAs or large intergenic RNAs which all are incorporated in protein-containing complexes targeting specific genes (especially, their 3′-UTR) will have to be assessed in this context to gather further insights.

Materials and Methods

Acquisition of the SNP Data Sets

We downloaded the Catalog of Published Genome-Wide Association Studies [29], which includes information about 5,101 unique SNP-trait associations with a p-value of $P \leq 1.0 \cdot 10^{-5}$ [30]. We refer to this set as GWAS-SNPs. The GWAS-SNPs were extended by highly correlating SNPs with strong LD ($r^2 \geq 0.8$) in the HapMap3 CEU panel (Utah residents with northern and western European ancestry from the CEPH collection) [31–33]. This set, further referred to as extended GWAS-SNPs, contains 18,884 variants retrieved by using SNAP [62].

As background distribution for localization enrichment we used the 2.7 million SNPs from the CEU panel of the joint HapMap Phases I, II and III (release 27, referred to as HapMap-SNPs) for which genotype information was available [31–33]. All SNPs were mapped to official identifiers [63] using SNAP [62] and their genome build NCBI36/hg18 coordinates were retrieved from the UCSC (University of California Santa Cruz) Table Browser [64]. For a background set representative for the HapMap-SNPs with comparable properties as the extended GWAS-SNPs, we randomly selected 5,101 SNPs from the HapMap-SNP set and extended this set, analogous as for the GWAS-SNPs, with SNPs in strong LD using $r^2 \geq 0.8$. This process was repeated 1000 times. For statistical evaluation of the localization enrichment and the effects on functional elements against this background, we computed the respective statistics for all 1000 sets, fitted a distribution to the resulting values and retrieved the probability to observe the statistic obtained for the extended GWAS-SNPs. As the HapMap-SNPs are a superset of all other SNP sets, i.e. the 1000 random sets and the GWAS-SNPs as well as the extended GWAS-SNPs, we always performed SNP-based annotations for the whole HapMap-SNP set. Thus, we simultaneously retrieved the properties of all other SNP sets.

We mapped all SNPs on genomic locations of protein-coding genes and miRNA genes. The genome annotations were obtained from the UCSC Table Browser [64] based on the NCBI Reference Sequence annotation (genome build NCBI36/hg18) [65] and from miRBase release 18 (genome assembly GRCh37/hg19) [66–68]. We used the UCSC liftOver tool [69] to convert the genomic miRNA hairpin coordinates to the NCBI36/hg18 assembly. The chromosomal function of the SNPs was categorized into five classes: intergenic, intronic, 5′-UTR, CDS, and 3′-UTR. To assess differences regarding the MAF in and between the SNP sets we computed the distribution in bins of ten percent range for all SNPs. MAF data was used as given for the HapMap-SNPs. Adjustment for r^2 values in the localization enrichment analysis was performed accumulatively in 5% steps. For binning of HapMap-SNPs into LD blocks, we searched for all SNP sets with an all-vs.-all $r^2 \geq 0.8$. Each SNP was uniquely assigned to a LD block. The localization of the blocks was defined as the subset of the five classes occurring in the annotation of the SNPs contained in the respective bin.

We used the algorithm PhastCons from the PHAST package [70] to calculate the maximum likelihood of a SNP to be conserved across 17 vertebrates. We required a score greater than 0.57 to classify a site as conserved in mammals [71].

Annotation of RISC-target Regulatory Relationships

To analyze the single nucleotide mutation effect on miRNA targeting, we used the high-throughput transcriptome-wide CLIP-Seq interaction maps describing sites of the RBPs *Argonaute* and *TNRC6* in human *HEK293* cells [14] as provided by the starBase database [72]. The available chromosomal coordinates of the CLIP-Seq clusters were converted to the NCBI36/hg18 genome build and mapped to protein-coding genes according to the NCBI Reference Sequence annotation [65]. The final set contained 139,254 locations of RBP binding regions on 24,442 transcripts. 48% of sites were located within a 3′-UTR.

Examination of Polyadenylation Signals

We obtained chromosomal positions of poly(A) signals from the PolyA DB for mRNA polyadenylation sites [73,74]. As the position of poly(A) sites is described to be located 10–30 nucleotides (nt) downstream of the poly(A) signals [73,75], SNPs within this range site were determined. We then extracted 11 nt long mRNA sequences centered around these 3′-UTR SNPs and examined the sequences for the most abundant poly(A) signal variations according to [75]. We classified a SNP as effecting a poly(A) signal if either creation of a new poly(A) signal sequence or disruption of an existing signal occurs. Classification of a SNP as not effecting a poly(A) signal was carried out if there was no existing signal in the 11 nt sequence for both the wild-type allele and the mutated allele or if the respective allele does not only disrupt a signal but simultaneously creates another poly(A) signal (“synonymous mutation”).

Determination of Splice Sites

We applied the NNSplice algorithm from the Berkeley Drosophila Genome project [76] to identify changes in splice sites. As input we used a genomic DNA sequence window of 60 nt centered at the SNP position of the wild-type and the mutated type. Predicted splice sites with a likelihood greater than 0.5 were retained neglecting cases with marginal changes [77]. All types of splice site change were considered: loss/gain of splice site and increase/decrease of likelihood. The distance of any splice site change to an exon junction site as defined by the NCBI Reference gene annotation (genome build NCBI36) [65] was computed. We filtered lost acceptor sites or sites exhibiting an increase/decrease in their likelihood if they were located between 100 nt upstream and 10 nt downstream of a reference intron/exon border. Lost donor sites or sites with an increased/decreased likelihood were retained if they were located between 10 nt up- and downstream of a reference exon/intron border. A gain of a completely new splice site was always kept [77].

Analysis of RNA Structural Properties

To account for structural changes caused by SNPs we used the RNAfold algorithm from the Vienna RNA Package version 1.8.5 [78]. We considered the ensemble of possible RNA conformations by calculation of the partition function and the base pairing probability matrix of the wild-type and the mutated 3′-UTR sequences [79]. The row sums were computed to define a pairing score for each nucleotide. We extracted a 41 nt long score vector centered at an RBP:RNA interaction site for the wild-type and the mutated structure. The linear correlation between both structures was measured by the Pearson product-moment correlation coefficient [80]. Since a transcript can hold several RBP-binding sites we selected for each SNP the smallest correlation coefficient per transcript. To evaluate the significance of a change in the RNA structural ensemble we calculated the minimal correlation coefficient for all SNPs of all 1000 random samples. Based on this distribution we determined a correlation coefficient of 0.55 having a probability of less than 5% for a type I error (**Figure 1C**). All SNPs inducing a minimal correlation coefficient of less than 0.55 between wild-type and mutated structure were filtered.

Identification of Altered MREs

To find all possible MREs we searched for sites complementary to a canonical miRNA seed sequence [47] within the wild-type and mutated 3′-UTRs. The seeking for short sequence matches may yield a plethora of putative target sites with a high false positive rate. The CLIP-seq methods have been shown to significantly reduce the fraction of false positive MREs [14,15]. Thus, we classified MREs as functional if they were located within a distance of 21 nt to the center of a RBP interaction site [14]. To additionally reduce the rate of false positives we required at least one miRNA sequence read as reported in the accordant Clip-Seq experiment. Further, the enrichment of MREs of each miRNA within RISC-binding regions was calculated. To identify MREs disrupted by a SNP we filtered miRNAs of which MREs were significantly overrepresented within RISC-binding regions ($LOR>0$, $P_{\chi^2}<0.05$). The determination of MREs created by SNPs was performed analogous ($LOR>0$, $P_{\chi^2}<0.05$). We retrieved 258 miRNAs the targeting of which could be disturbed and 324 miRNAs the formation of mRNA:miRNA hybrids of which could be enhanced.

Functional Annotation of Genes Containing 3′-UTR SNPs

We evaluated the enrichment of functional annotations of the 326 genes containing 3′-UTR SNPs using DAVID [34,35] and GSEA [81,82]. Additionally, the genes were annotated according to their associated traits investigated in the corresponding GWAS. To this end, all traits were mapped to official disease terms as contained in the MeSH (Medical Subject Headings) ontology. As disease class we used the upmost level in the hierarchy tree. Trait enrichment analysis limited to the associated traits as contained in the GWAS Catalog was performed using a χ^2 test statistic [80]. For multiple testing correction in Gene Ontology [83] term and disease class enrichment analysis we employed the Bonferroni correction with an overall significance level of $\alpha=0.05$.

Construction of Qualitative Systems Biological Models

For retrieval of phenotype-specific microRNA expression alterations we used PhenomiR [2] and miR2Disease [1]. The interaction data for constructing the network were obtained by manual text-mining and using the public databases IntAct [84], CORUM [85] and KEGG [86–88]. Expression data was taken from the MuTHER study [48], made accessible through the Java-interface Genevar [89]. As in the MuTHER study the association of expression values to an allele is given for two separated twin cohorts, we used a combined p-value from both sets to calculate significance. We used Fisher's combined p-value which was shown to be applicable to expression data [90]. The significance level was adjusted using the rough false discovery rate [91].

Supporting Information

Table S1 SNPs predicted to affect 3′-UTR splicing. The first column lists the SNP rs-numbers, in the second column the respective transcripts are given and the third column contains the genomic locus of the SNP including strand information of the transcripts. The fourth column shows the conservation of the SNP in binary code, i.e. 1 means conserved, 0 means not conserved. The fifth column gives the type of the gained splice site, the score column contains the likelihood of NNSplice and the last column provides the percentage of lost RBP binding sites for the accordant transcripts.
(XLS)

Table S2 SNPs predicted to affect 3′-UTR secondary structure. The first column lists the SNP rs-numbers, in the second column the respective transcripts are given and the third column contains the genomic locus of the SNP including strand information of the transcripts. The fourth column shows the conservation of the SNP in binary code, i.e. 1 means conserved, 0 means not conserved. The fifth column lists the correlation coefficient of the wild-type structure to the mutated structure of the respective transcripts.
(XLS)

Table S3 SNPs disrupting existing MREs. The first column lists the SNP rs-numbers, in the second column the respective transcripts are given and the third column contains the genomic locus of the SNP including strand information of the transcripts. The fourth column shows the conservation of the SNP in binary code, i.e. 1 means conserved, 0 means not conserved. The fifth column lists the miRNAs the MREs of which are affected.
(XLS)

Table S4 SNPs creating new MREs. The first column lists the SNP rs-numbers, in the second column the respective transcripts are given and the third column contains the genomic locus of the SNP including strand information of the transcripts. The fourth column shows the conservation of the SNP in binary code, i.e. 1 means conserved, 0 means not conserved. The fifth column lists the miRNAs for which MREs are created.
(XLS)

Author Contributions

Conceived and designed the experiments: MA DCE MLH VS AP. Performed the experiments: MA DCE. Analyzed the data: MA DCE. Contributed reagents/materials/analysis tools: MA DCE. Wrote the paper: MA DCE AP.

References

1. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, et al. (2009) mir2disease: a manually curated database for microrna deregulation in human disease. Nucleic Acids Res 37: D98–104.
2. Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, et al. (2010) Phenomir: a knowledgebase for microrna expression in diseases and biological processes. Genome Biol 11: R6.
3. Gupta SK, Bang C, Thum T (2010) Circulating micrornas as biomarkers and potential paracrine mediators of cardiovascular disease. Circ Cardiovasc Genet 3: 484–488.
4. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, et al. (2005) Microrna expression profiles classify human cancers. Nature 435: 834–838.
5. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, et al. (2008) Micrornas accurately identify cancer tissue origin. Nat Biotechnol 26: 462–469.
6. Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, et al. (2007) Distinctive patterns of microrna expression in primary muscular disorders. Proc Natl Acad Sci U S A 104: 17016–17021.
7. Friedman RC, Farh KKH, Burge CB, Bartel DP (2009) Most mammalian mrnas are conserved targets of micrornas. Genome Res 19: 92–105.
8. Hebert SS, Strooper BD (2009) Alterations of the microrna network cause neurodegenerative disease. Trends Neurosci 32: 199–206.
9. Johnston RJ, Chang S, Etchberger JF, Ortiz CO, Hobert O (2005) Micrornas acting in a doublenegative feedback loop to control a neuronal cell fate decision. Proc Natl Acad Sci U S A 102: 12449–12454.
10. Re A, Cora D, Taverna D, Caselle M (2009) Genome-wide survey of microrna-transcription factor feed-forward regulatory circuits in human. Mol Biosyst 5: 854–867.
11. Wilbert ML, Yeo GW (2010) Genome-wide approaches in the study of microrna biology. Wiley Interdiscip Rev Syst Biol Med.
12. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, et al. (2003) Clip identifies nova-regulated rna networks in the brain. Science 302: 1212–1215.
13. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, et al. (2008) Hits-clip yields genome-wide insights into brain alternative rna processing. Nature 456: 464–469.
14. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. (2010) Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. Cell 141: 129–141.

15. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute hits-clip decodes microrna-mrna interaction maps. Nature 460: 479–486.
16. de la Chapelle A (2009) Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. Oncogene 28: 3345–3348.
17. Meola N, Gennarino VA, Banfi S (2009) microrna and genetic diseases. Pathogenetics 2: 7.
18. Richardson K, Lai CQ, Parnell LD, Lee YC, Ordovas JM (2011) A genome-wide survey for snps altering microrna seed sites identifies functional candidates in gwas. BMC Genomics 12: 504.
19. Bennett CL, Brunkow ME, Ramsdell F, O'Briant KC, Zhu Q, et al. (2001) A rare polyadenylation signal mutation of the foxp3 gene (aauaaa–¿aaugaa) leads to the ipex syndrome. Immunogenetics 53: 435–439.
20. Walters RW, Bradrick SS, Gromeier M (2010) Poly(a)-binding protein modulates mrna susceptibility to cap-dependent mirna-mediated repression. RNA 16: 239–250.
21. Li X, Quon G, Lipshitz HD, Morris Q (2010) Predicting in vivo binding sites of rna-binding proteins using mrna secondary structure. RNA 16: 1096–1107.
22. Waldispühl J, Clote P (2007) Computing the partition function and sampling for saturated secondary structures of rna, with respect to the turner energy model. J Comput Biol 14: 190–215.
23. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microrna target recognition. Nat Genet 39: 1278–1284.
24. Yang JO, Kim WY, Bhak J (2009) sssnptarget: genome-wide splice-site single nucleotide polymorphism database. Hum Mutat 30: E1010–E1020.
25. Visscher PM, Montgomery GW (2009) Genome-wide association studies and human disease: from trickle to ood. JAMA 302: 2028–2029.
26. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8: e1000294.
27. Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, et al. (2007) Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the il2ra region in type 1 diabetes. Nat Genet 39: 1074–1082.
28. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17: R156–R165.
29. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA (2011) A catalog of published genomewide association studies. Available at: wwwgenomegov/gwastudies Accessed December 18th, 2011.
30. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367.
31. Consortium IH (2003) The international hapmap project. Nature 426: 789–796.
32. Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million snps. Nature 449: 851–861.
33. Consortium IH, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.
34. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc 4: 44–57.
35. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37: 1–13.
36. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431–432.
37. Heathcote J (2000) Update on primary biliary cirrhosis. Can J Gastroenterol 14: 43–48.
38. Mackay IR (2007) Autoimmune diseases of the liver, autoimmune hepatitis and primary biliary cirrhosis: Unfinished business. Hepatol Res 37 Suppl 3: S357–S364.
39. Selmi C, Zuin M, Gershwin ME (2008) The unfinished business of primary biliary cirrhosis. J Hepatol 49: 451–460.
40. Sawa T, Ohshima H (2006) Nitrative dna damage in inammation and its possible role in carcinogenesis. Nitric Oxide 14: 91–100.
41. Kryston TB, Georgiev AB, Pissis P, Georgakilas AG (2011) Role of oxidative stress and dna damage in human carcinogenesis. Mutat Res.
42. Hirschfield GM, Liu X, Xu C, Lu Y, Xie G, et al. (2009) Primary biliary cirrhosis associated with hla, il12a, and il12rb2 variants. N Engl J Med 360: 2544–2555.
43. Onn I, Aono N, Hirano M, Hirano T (2007) Reconstitution and subunit geometry of human condensing complexes. EMBO J 26: 1024–1034.
44. Zindy P, Andrieux L, Bonnier D, Musso O, Langouet S, et al. (2005) Upregulation of dna repair genes in active cirrhosis associated with hepatocellular carcinoma. FEBS Lett 579: 95–99.
45. Heale JT, Ball AR, Schmiesing JA, Kim JS, Kong X, et al. (2006) Condensin i interacts with the parp-1-xrcc1 complex and functions in dna single-strand break repair. Mol Cell 21: 837–848.
46. Padgett KA, Lan RY, Leung PC, Lleo A, Dawson K, et al. (2009) Primary biliary cirrhosis is associated with altered hepatic microrna expression. J Autoimmun 32: 246–253.
47. Ellwanger DC, Büttner FA, Mewes HW, Stümpen V (2011) The sufficient minimal set of mirna seed types. Bioinformatics 27: 1346–1350.
48. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The architecture of gene regulatory variation across multiple human tissues: the muther study. PLoS Genet 7: e1002003.
49. Turissini SB, Kaplan MM (1997) Hepatocellular carcinoma in primary biliary cirrhosis. Am J Gastroenterol 92: 676–678.
50. Howel D, Metcalf JV, Gray J, Newman WL, Jones DE, et al. (1999) Cancer risk in primary biliary cirrhosis: a study in northern england. Gut 45: 756–760.
51. Sage E, Harrison L (2010) Clustered dna lesion repair in eukaryotes: Relevance to mutagenesis and cell survival. Mutat Res.
52. Li SX, Sjolund A, Harris L, Sweasy JB (2010) Dna repair and personalized breast cancer therapy. Environ Mol Mutagen 51: 897–908.
53. Kirschner K, Melton DW (2010) Multiple roles of the ercc1-xpf endonuclease in dna repair and resistance to anticancer drugs. Anticancer Res 30: 3223–3232.
54. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.
55. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, et al. (2008) The impact of micrornas on protein output. Nature 455: 64–71.
56. Guo H, Ingolia NT, Weissman JS, Bartel DP (2010) Mammalian micrornas predominantly act to decrease target mrna levels. Nature 466: 835–840.
57. Friedman JM, Jones PA (2009) Micrornas: critical mediators of differentiation, development and disease. Swiss Med Wkly 139: 466–472.
58. Sheedy FJ, O'Neill LAJ (2008) Adding fuel to fire: micrornas as a new class of mediators of inammation. Ann Rheum Dis 67 Suppl 3: iii50–iii55.
59. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356–369.
60. modE N C O D E Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, et al. (2010) Identification of functional elements and regulatory circuits by drosophila modencode. Science 330: 1787–1797.
61. Gerstein MB, Lu ZJ, Nostrand ELV, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the caenorhabditis elegans genome by the modencode project. Science 330: 1775–1787.
62. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) Snap: a web-based tool for identification and annotation of proxy snps using hapmap. Bioinformatics 24: 2938–2939.
63. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbsnp: the ncbi database of genetic variation. Nucleic Acids Res 29: 308–311.
64. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The ucsc table browser data retrieval tool. Nucleic Acids Res 32: D493–D496.
65. Pruitt KD, Tatusova T, Maglott DR (2007) Ncbi reference sequences (refseq): a curated nonredundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–D65.
66. Griffiths-Jones S (2004) The microrna registry. Nucleic Acids Res 32: D109–D111.
67. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) mirbase: microrna sequences, targets and gene nomenclature. Nucleic Acids Res 34: D140–D144.
68. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) mirbase: tools for microrna genomics. Nucleic Acids Res 36: D154–D158.
69. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2010) The ucsc genome browser database: update 2010. Nucleic Acids Res 38: D613–D619.
70. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034–1050.
71. Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microrna.org resource: targets and expression. Nucleic Acids Res 36: D149–D153.
72. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, et al. (2011) starbase: a database for exploring microrna-mrna interaction maps from argonaute clip-seq and degradome-seq data. Nucleic Acids Res 39: D202–D209.
73. Zhang H, Hu J, Recce M, Tian B (2005) Polya db: a database for mammalian mrna polyadenylation. Nucleic Acids Res 33: D116–D120.
74. Lee JY, Yeh I, Park JY, Tian B (2007) Polya db 2: mrna polyadenylation sites in vertebrate genes. Nucleic Acids Res 35: D165–D168.
75. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res 10: 1001–1010.
76. Reese MG, Eeckman FH, Kulp D, Haussler D (1997) Improved splice site detection in genie. J Comput Biol 4: 311–323.
77. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010) Mutationtaster evaluates diseasecausing potential of sequence alterations. Nat Methods 7: 575–576.
78. Hofacker IL (2003) Vienna rna secondary structure server. Nucleic Acids Res 31: 3429–3431.
79. Halvorsen M, Martin JS, Broadaway S, Laederach A (2010) Disease-associated mutations that alter the rna structural ensemble. PLoS Genet 6: e1001074.
80. R Development Core Team (2011) R: A Language and Environment for Statistica Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3–900051–07–0. Accessed 2011 Dec 18.
81. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
82. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. (2003) Pgc-1alpharesponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273.
83. Consortium GO (2010) The gene ontology in 2010: extensions and refinements. Nucleic Acids Res 38: D331–D335.

84. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The intact molecular interaction database in 2010. Nucleic Acids Res 38: D525–D531.
85. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, et al. (2010) Corum: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res 38: D497–D501.
86. Kanehisa M, Goto S (2000) Kegg: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.
87. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in kegg. Nucleic Acids Res 34: D354–D357.
88. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–D360.
89. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, et al. (2010) Genevar: a database and java application for the analysis and visualization of snp-gene associations in eqtl studies. Bioinformatics 26: 2474–2476.
90. Hess A, Iyer H (2007) Fisher's combined p-value for detecting differentially expressed genes using affymetrix expression arrays. BMC Genomics 8: 96.
91. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29: 1165–1188.