

Integrative Analysis of mRNA Expression and Half-Life Data Reveals *Trans*-Acting Genetic Variants Associated with Increased Expression of Stable Transcripts

Thong T. Nguyen¹, Cathal Seoighe^{1,2*}

¹ School of Mathematics, Statistics & Applied Mathematics, National University of Ireland, Galway, Ireland, ² Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Anzio Road, Observatory, Cape Town, South Africa

Abstract

Genetic variation in gene expression makes an important contribution to phenotypic variation and susceptibility to disease. Recently, a subset of *cis*-acting expression quantitative loci (eQTLs) has been found to result from polymorphisms that affect RNA stability. Here we carried out a search for *trans*-acting variants that influence RNA stability. We first demonstrate that differences in the activity of *trans*-acting factors that stabilize RNA can be detected by comparing the expression levels of long-lived (stable) and short-lived (unstable) transcripts in high-throughput gene expression experiments. Using gene expression microarray data generated from eight HapMap3 populations, we calculated the relative expression ranks of long-lived transcripts versus short-lived transcripts in each sample. Treating this as a quantitative trait, we applied genome-wide association and identified a single nucleotide polymorphism (SNP), rs6137010, on chromosome 20p13 with which it is strongly associated in two Asian populations ($p = 4 \times 10^{-10}$ in CHB – Han Chinese from Beijing; $p = 1 \times 10^{-4}$ in JPT – Japanese from Tokyo). This SNP is a *cis*-eQTL for *SNRPB* in CHB and JPT but not in the other six HapMap3 populations. *SNRPB* is a core component of the spliceosome, and has previously been shown to affect the expression of many RNA processing factors. We propose that a *cis*-eQTL of *SNRPB* may be directly responsible for inter-individual variation in relative expression of long-lived versus short-lived transcript in Asian populations. In support of this hypothesis, knockdown of *SNRPB* results in a significant reduction in the relative expression of long-lived versus short-lived transcripts. Samples with higher relative expression of long-lived transcripts also had higher relative expression of coding compared to non-coding RNA and of RNA from housekeeping compared to non-housekeeping genes, due to the lower decay rates of coding RNAs, particularly those that perform housekeeping functions, compared to non-coding RNAs.

Citation: Nguyen TT, Seoighe C (2013) Integrative Analysis of mRNA Expression and Half-Life Data Reveals *Trans*-Acting Genetic Variants Associated with Increased Expression of Stable Transcripts. PLoS ONE 8(11): e79627. doi:10.1371/journal.pone.0079627

Editor: Georg Stoecklin, German Cancer Research Center, Germany

Received: July 17, 2013; **Accepted:** October 3, 2013; **Published:** November 18, 2013

Copyright: © 2013 Nguyen, Seoighe. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by Science Foundation Ireland (07/SK/M1211b) and the Irish Research Council for Science Engineering and Technology. 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: Cathal.Seoighe@nuigalway.ie

Introduction

RNA stability plays a major role in gene expression regulation in virtually all organisms, from bacteria to mammals [1,2,3]. Indeed, steady-state gene expression levels represent the equilibrium of two opposing biological processes: RNA transcription and RNA decay. Changes in gene expression levels can result from alteration in either of these processes [1,4]. Recent studies have investigated RNA stability using high-throughput techniques in diverse organisms, from yeast [5,6] to *Arabidopsis* [7], mouse [8,9,10], and human [10,11,12,13,14], and for both coding and non-coding RNAs [9,15]. Several of these studies have reported strong correlations between RNA stability and steady-state gene expression levels. In addition, RNA stability has been shown to be related to physiological function [8,12]. For example, genes encoding proteins involved in housekeeping functions tend to have stable mRNAs [10,15]. The modulation of RNA stability can, in turn, have a major impact on cellular processes, including proliferation, differentiation, and adaptation to environmental stimuli [1,2,3]. Dysregulation of RNA stability has been linked to

several human diseases, such as chronic inflammation [16], cardiovascular disease and cancer [17,18,19].

The regulation of RNA stability is achieved through interactions between *trans*-acting RNA-binding proteins and *cis*-acting elements within RNAs [20,21]. Among RNA-binding proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs) are key factors that regulate major steps of gene expression, including pre-mRNA processing, RNA stability, and translation [22,23,24]. For example, *HNRNPA2B1*, a member of the hnRNP family, was found to stabilize a large number of target transcripts carrying a conserved structural RNA element in the 3' untranslated regions [13]. Knockdown of *HNRNPA2B1* resulted in a remarkable increase in the relative decay rate of the target transcripts and, consequently, a significant decrease in their expression levels [13]. The contribution of RNA decay to gene expression levels was also investigated in a recent study where a subset of *cis*-acting expression quantitative loci (*cis*-eQTLs) was found to be a consequence of variation in decay rates [25]. A moderate number of genetic variants were found to significantly associate with inter-individual variation in both gene expression and RNA decay, for which variation in RNA decay could explain the association with

gene expression level [25]. Despite increased appreciation of the role of RNA stabilization in determining gene expression levels there has been no investigation of *trans*-acting genetic variants that affect the stabilization of RNA.

Here we investigate factors that affect RNA stability in *trans*. We first show that perturbation of RNA stabilization factors that affect multiple genes can be inferred from gene expression data. Given a dataset of RNA decay rates and expression levels, we define the RNA stability score (RS-score), based on the expression of long-lived transcripts relative to short-lived transcripts. Knocking down *HNRNPA2B1*, which has been shown to be involved in stabilization of a large proportion of RNAs [13], leads to a significant reduction in the RS-score. Using gene expression microarray data generated from eight HapMap3 populations [26], we identified a SNP, rs6137010, on chromosome 20p13 that is strongly associated with the RS-score in Asian populations. This SNP is a *cis*-eQTL of *SNRPB*, a gene that encodes a core component of the spliceosome and has been shown to modulate the expression of many RNA processing factors [27]. The C allele of rs6137010 is associated both with higher expression of *SNRPB* and higher RS-score. Knockdown of *SNRPB* results in a significant decrease in the RS-score, suggesting that the *cis*-eQTL for *SNRPB* is responsible for the observed genetic variation in RS-score in Asian populations.

Results and Discussion

Perturbation of RNA stabilization is detectable from gene expression data

We hypothesized that changes in the activity of *trans*-acting factors that are involved in stabilizing multiple RNAs could be detectable by analyzing gene expression profiles. To test this hypothesis we obtained gene expression data from a published study in which the heterogeneous ribonucleoprotein, *HNRNPA2B1*, was knocked down [13]. In the original study this gene was shown to play a role in the stabilization of RNAs containing an abundant structural motif and RNAs containing this motif were downregulated in the knockdown samples compared to controls [13]. However, even in the absence of knowledge of the specific *trans*-acting factor and target RNAs involved it is possible to infer the effects of the knockdown on RNA stability. This is because stable, long-lived transcripts are enriched among the genes that are targeted by *HNRNPA2B1* [13].

We divided genes into two groups by using RNA decay rate data from Goodarzi *et al.* [13]. The first group contains genes expressing long-lived RNAs (decay rate lower than the mean across genes) and the second group contains genes expressing short-lived RNAs (decay rate higher than the mean). We then defined the RS-score for a sample as the difference in the expression rank between these two groups of genes in the sample (see Methods for more details). A higher RS-score implies relatively higher expression levels of long-lived or stable RNAs. A similar idea has previously been used to infer the impact of miRNA regulation on target genes using gene expression data [28]. The regulatory effect score (RE-score) of a miRNA was defined as the difference in the mean expression rank between targets of the miRNA and non-targets. A higher RE-score indicates lower expression levels of target genes and, thereby, a stronger effect of the corresponding miRNA. Analogously, a higher RS-score implies that the long-lived RNAs that are more likely to be subject to stabilization by *trans*-acting factors are relatively more highly expressed in a sample.

The RS-score of the *HNRNPA2B1* knockdown was significantly lower than RS-score of the control in three independent replicates ($p = 3.7 \times 10^{-3}$; paired *t* test) (Figure 1). This is consistent with

expectations because *HNRNPA2B1* is one of the heterogeneous nuclear ribonucleoproteins that influence pre-mRNA processing and other aspects of RNA metabolism and transport. More importantly, *HNRNPA2B1* is involved in stabilizing a large number of genes, particularly genes expressing long-lived RNAs, by binding to a structural RNA motif of target genes [13]. *HNRNPA2B1* knockdown caused a significant reduction in the expression levels of long-lived RNAs (Figure S1), resulting in lower RS-scores in the knockdown samples. These observations suggest that gene expression levels can be used to infer the effects of *trans*-acting factors that are involved in stabilizing large numbers of genes.

The genetics of *trans*-acting factors that affect RNA stability

We obtained gene expression data generated from lymphoblastoid cell lines of 726 individuals in eight HapMap3 populations [26] (Table S1). Using the half-life data from HeLa cells [15], we calculated the RS-score for each of these individuals (see Methods). Interestingly, the RS-score was well correlated with the expression level of *HNRNPA2B1* in most of the populations (Table S2), with the strongest correlation in CHB (Spearman $\rho = 0.48$; $p = 8.4 \times 10^{-6}$). Because the experimental knock down of *HNRNPA2B1* results in a reduction in the RS-score, we hypothesized that *cis*-eQTLs affecting the expression level of *HNRNPA2B1* should also be associated with RS-score. This is the case for four *cis*-eQTLs of this gene in two of the HapMap3 populations (Table S3).

To search more generally for genetic variants associated with the RS-score we used a genome-wide association study (GWAS) approach, treating the RS-score as a quantitative trait. We carried out additive tests of association between single nucleotide polymorphisms (SNPs) genotyped as part of the HapMap3 project

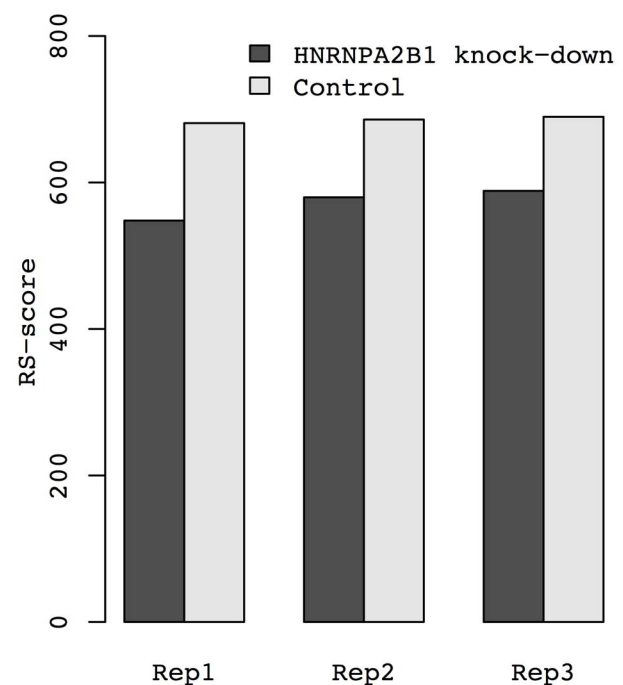


Figure 1. *HNRNPA2B1* knock-down results in reduced RS-score. RS-score was calculated for *HNRNPA2B1* knockdown samples and control samples separately in three independent replicates (Rep1, Rep2, and Rep3). doi:10.1371/journal.pone.0079627.g001

and the RS-score in each population separately (see the Methods section for more details). We found one strong association between a SNP, rs6137010, on chromosome 20p13 and RS-score in the CHB population ($p = 4.4 \times 10^{-10}$; Figure 2). Interestingly, this association is replicated in the other Asian population – JPT ($p = 1.2 \times 10^{-4}$). We used a label permutation procedure to check the robustness of this result to failures in modelling assumptions (see Methods). The association between rs6137010 and RS-score in CHB was stronger than the best associations in each of 1,000 label permutations. Furthermore, the Bonferroni-adjusted p-value of this association is very significant (Bonferroni $p = 5.9 \times 10^{-3}$). Therefore, the association between rs6137010 and RS-score in CHB is robust, genome-wide significant, and replicated in a second population (JPT).

To increase the statistical power of the association tests, we combined individuals from different populations. Because different populations have different ancestries combining individuals from these populations can lead to spurious associations, resulting from structure in the combined population. To tackle this problem, we applied a principal components analysis (PCA) approach [29] (see Methods for more details) to model ancestry differences among all 726 individuals. In a scatter plot of the first and second principal components (Figure S2) three broad clusters are evident, consisting of the African populations, the Asian populations and CEU, MEX, GIH. Given these clusters, we considered four ways of combining populations: CHB+JPT (Asian populations), YRI+MEX+LWK (African populations), CEU+GIH+MEX, and finally all 8 populations (ALL). For each combination, we performed a principal components analysis and included the first five principal components as covariates in the GWAS regression models (see Methods). The SNP rs6137010 was strongly associated with the RS-score in CHB+JPT ($p = 2.0 \times 10^{-12}$; Figure S3). This association is also the best among 1000 permutations and is genome-wide significant (Bonferroni $p = 2.7 \times 10^{-5}$). In total, 6 genetic markers showed genome-wide significant association (Bonferroni $p < 0.05$) but the association with rs6137010 in CHB+JPT was the strongest (Table 1). The P-P plots showed that the p-value of the association with the RS-score at rs6137010

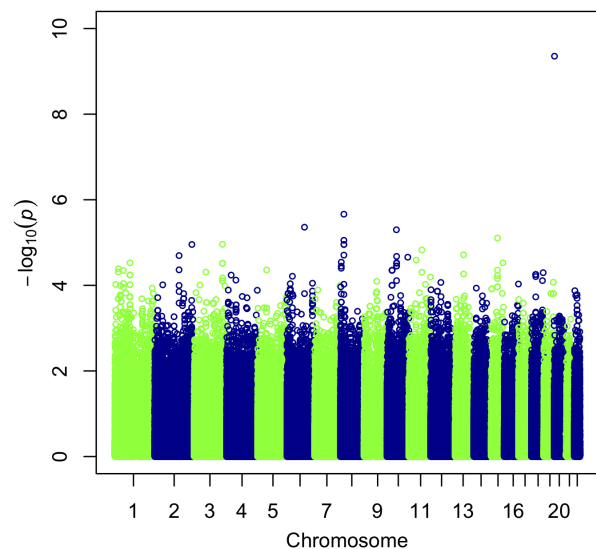


Figure 2. Manhattan plot for GWA with RS-score in CHB. The plot shows $-\log_{10}$ of P-values from tests of association between individual SNP markers and the RS-score. Successive chromosomes are shown in different colors. doi:10.1371/journal.pone.0079627.g002

is very different to other loci in the Asian populations (Figure 3). We found no evidence of population stratification in the GWAS tests of the Asian populations as their genomic inflation factors are less than 1.05 (Table S4). However, unsurprisingly there was evidence of population stratification in three combined populations: CEU+GIH+MEX, YRI+LWK+MKK and ALL (Table S4; Figure S4).

To check the effect of the choice of half-life data on this result, we compared RS-scores calculated using half-life data from HeLa cells and RS-scores calculated using B-cell half-life data [14] and found that they were highly correlated in all populations (Spearman $\rho = 0.73 \pm 0.15$). It has previously been reported that RNAs involved in housekeeping functions tend to have long half-life [10,15]. As an alternative to using half-life data, which has the caveat that it may be cell type dependent, we calculated the RS-score by grouping genes based on whether they are housekeeping or not, using data from Chang *et al.* [30]. We found that the RS-score calculated by grouping the genes in this way was highly correlated with the RS-score based on the half-life in HeLa cells in all populations (Spearman $\rho = 0.70 \pm 0.09$). Moreover, the RS-score (based on the housekeeping data) was significantly associated with rs6137010 in the combined CHB+JPT population ($p = 7.1 \times 10^{-13}$; Bonferroni $p = 9.5 \times 10^{-6}$). We also calculated an equivalent score by considering protein-coding versus non-coding genes. Non-coding genes have been found to have shorter half-life than protein-coding genes [9,15]. This score was also highly correlated with the RS-score calculated from the half-life data and, again, significantly associated with rs6137010 in CHB+JPT ($p = 6.2 \times 10^{-10}$; Bonferroni $p = 8.3 \times 10^{-3}$). These two results are of interest, beyond providing an alternative way to group genes that is not dependent on RNA half-life data that may differ between cell types. They suggest that the proportion of the RNA pool corresponding to non-coding and tissue-specific genes is associated with rs6137010 in Asian populations.

Searching for causal SNPs and causal genes

To search for causal SNPs that may explain the GWAS results we mapped each SNP that shows genome-wide significant association with the RS-score to a gene if the SNP is either within the gene or is a *cis*-eQTL (*cis*-expression Quantitative Trait Locus) of the gene using *cis*-eQTL data from Stranger *et al.* [26] (Table 1). We found that rs6137010, the SNP with the strongest GWAS signal, mapped close to the *SNRPB* gene, which is involved in RNA processing. *SNRPB* encodes part of the core small nuclear ribonucleoprotein particles (snRNPs) that are major components of the spliceosome complex. Although it is 352 kb downstream, rs6137010 is significantly associated with the expression level of *SNRPB* in both CHB ($\rho = 0.50$; $p = 2.3 \times 10^{-6}$) and JPT ($\rho = 0.32$; $p = 3.7 \times 10^{-3}$), but not significantly associated with *SNRPB* expression in any of the other populations studied. The association between rs6137010 and *SNRPB* is strongest among all genes within 1 Mb-window centered on the SNP. Furthermore, the SNP is within an enhancer region as evidenced from whole-genome chromatin state segmentation data [31] available through the UCSC genome browser [32]. These results show that rs6137010 is a *cis*-eQTL of *SNRPB* in Asian populations. Changes in the expression level of *SNRPB* have been reported to affect alternative splicing and abundance of a large number of RNA processing factors [27]. rs6137010 has two alleles, T and C, with C the minor allele in Asian populations but the major allele in the other HapMap3 populations. Asian individuals carrying the C allele at this SNP had higher expression levels of *SNRPB* (Figure 4A) and higher RS-scores (Figure 4B). This suggests that the association between rs6137010 and inter-individual variation

Table 1. Markers associated with the RS-score at Bonferroni $p < 0.05$.

SNP	Location	Function	Associated gene	Population	P-value	Bonferroni
rs6137010	20:2038118	<i>cis</i> -eQTL	SNRPB	CHB+JPT	2.0×10^{-12}	2.7×10^{-5}
rs6137010	20:2038118	Intron	STK35	CHB+JPT	2.0×10^{-12}	2.7×10^{-5}
rs6137010	20:2038118	<i>cis</i> -eQTL	SIRPA	ALL	5.4×10^{-11}	9.0×10^{-4}
rs6137010	20:2038118	intron	STK35	ALL	5.4×10^{-11}	9.0×10^{-4}
rs11136253	8:145179783	<i>cis</i> -eQTL	ZNF707	ALL	1.5×10^{-10}	2.5×10^{-3}
rs11136253	8:145179783	coding-synon	OPLAH	ALL	1.5×10^{-10}	2.5×10^{-3}
rs6137010	20:2038118	<i>cis</i> -eQTL	SNRPB	CHB	4.4×10^{-10}	5.9×10^{-3}
rs6137010	20:2038118	Intron	STK35	CHB	4.4×10^{-10}	5.9×10^{-3}
rs4466324	7:85113458	unknown	None	ALL	6.0×10^{-10}	1.1×10^{-2}
rs17127419	11:122878168	unknown	HSPA8	ALL	9.8×10^{-10}	1.6×10^{-2}
rs12034707	1:178400832	<i>cis</i> -eQTL	TOR1AIP1	ALL	1.8×10^{-9}	3.0×10^{-2}
rs12034707	1:178400832	intron	QSOX1	ALL	1.8×10^{-9}	3.0×10^{-2}
rs10997765	10:69066422	intron	CTNNA3	ALL	1.9×10^{-9}	3.3×10^{-2}

doi:10.1371/journal.pone.0079627.t001

in RNA stability could be mediated by changes in *SNRPB* expression levels.

To identify genes across the human genome whose expression levels are significantly associated with rs6137010, we carried out *trans*-eQTL mapping for this SNP by fitting Spearman rank correlation models and considering only associations with $FDR < 0.1$. FDRs were calculated using the Benjamini and Hochberg procedure [33] as implemented in R [34]. We found 6,396 and 2,585 genes associated with rs6137010 in CHB and JPT, respectively. Among these, 3,194 (in CHB) and 429 (in JPT) genes were positively correlated with the minor allele count of rs6137010. Among the genes that were associated with rs6137010, 25.2% were putative targets for AU-rich element decay, compared to 17.6% of other genes ($p = 0.01$, Fisher exact test). We did not find any genes significantly associated with the SNP in other populations using the same FDR threshold. We carried out Gene Ontology (GO) analyses using DAVID [35] for the positively

correlated genes and, interestingly, found that they were enriched for the GO term ribonucleoprotein complex in both CHB ($p = 1.9 \times 10^{-25}$; Table S5) and JPT ($p = 3.7 \times 10^{-5}$). The ribonucleoprotein complex is known to be involved in many steps of RNA processing such as pre-mRNA splicing and RNA transportation and stabilization. Both *HNRNPA2B1* and *SNRPB* mentioned above belong to the ribonucleoprotein complex. These results indicate that rs6137010 is a *trans*-eQTL cluster that is disproportionately associated with the expression levels of ribonucleoprotein complex genes.

We next turned to investigating further the possible role of *SNRPB* in mediating the association of rs6137010 with the RS-score. We obtained gene expression microarray data generated from HeLa cells in which *SNRPB* was knocked down and compared to controls [27]. Using the HeLa half-life data [15] we calculated and compared RS-scores between the two conditions and found a significant reduction of the RS-score in

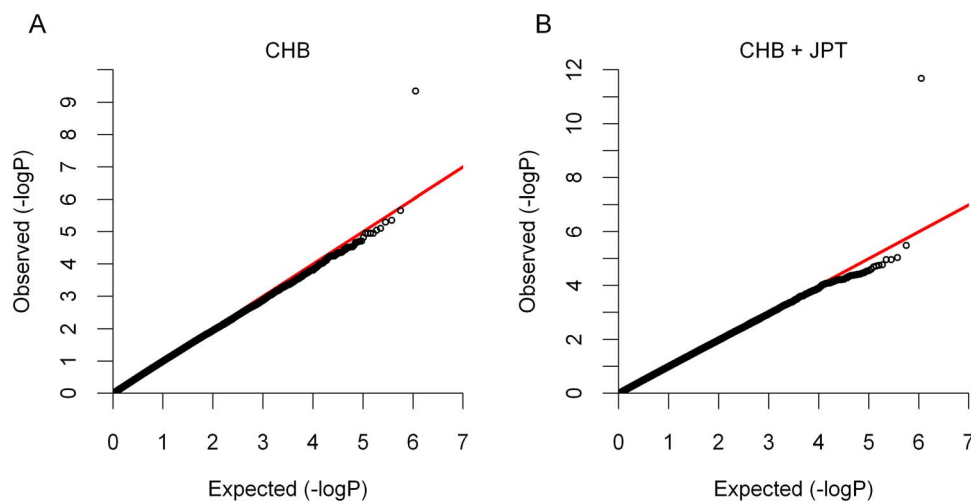


Figure 3. P-P plots of the association with RS-score in **A) CHB** and **B) CHB+JPT**. This figure compares the observed distribution of the $-\log_{10}$ P-values to the expected distribution, given that the P-values come from a uniform distribution in the interval zero to one (as expected under the null hypothesis). The Y-axis shows quantiles of the observed distribution and the X-axis shows the corresponding quantiles under the uniform distribution. The red line is used to compare the expected and observed values.

doi:10.1371/journal.pone.0079627.g003

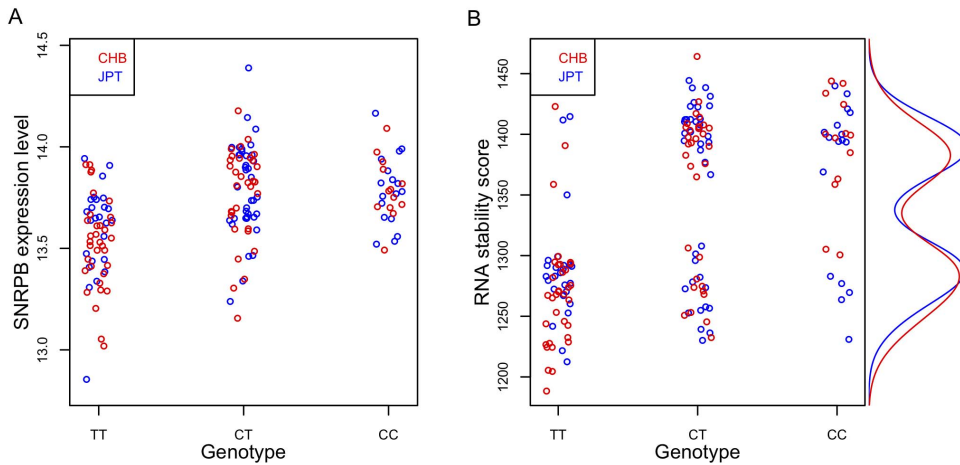


Figure 4. Stripcharts of *SNRPB* expression levels and the RS-score against the genotype of rs6137010 in CHB and JPT. A) *SNRPB* expression levels are significantly different among the three genotypes TT, CT and CC ($p=1.2\times 10^{-5}$ in CHB and $p=1.9\times 10^{-3}$ in JPT from one-way ANOVA). B) RS-scores are significantly different among the three genotypes ($p=4.3\times 10^{-10}$ in CHB and $p=7.0\times 10^{-5}$ in JPT from one-way ANOVA). The bimodal distributions of the RS-score in CHB and JPT are displayed in red and blue lines, respectively. doi:10.1371/journal.pone.0079627.g004

SNRPB knockdown ($p=1.2\times 10^{-6}$ from a two-tailed *t* test; Figure 5). This is consistent with expectations because depletion of *SNRPB* reduces the levels of many RNA processing genes [27], potentially affecting the stability of RNA across the transcriptome. Furthermore, the genes that were differentially expressed upon *SNRPB* knockdown were enriched for genes that showed the strongest association (FDR<0.01) with rs6137010 in CHB ($p=0.002$ from two-tailed Fisher's exact test). These results suggest that rs6137010, by modulating the expression of *SNRPB*, may be directly responsible for inter-individual variation in the RS-score in CHB. Interestingly, the distribution of the RS-score was bimodal in both CHB and JPT (Figure 4B), consistent with the existence of an associated locus with a large effect size. It is tempting to speculate that an ungenotyped causal SNP in strong linkage disequilibrium with rs6137010 may stratify the samples between the two modes of the distribution. Higher resolution genotype data will be necessary to test this hypothesis.

The RS-score of the knockdown of another splicing factor, *SRSF1*, is also significantly lower than of the control ($p=3.4\times 10^{-4}$ from a two-tailed *t* test), but significantly higher than of the *SNRPB* knockdown ($p=1.1\times 10^{-9}$ from a two-tailed *t* test) (Figure 5). This indicates that knocking down *SNRPB* has stronger effect on the RS-score than knocking down *SRSF1*. This is not surprising because *SNRPB* has been found to have a stronger impact than *SRSF1* on the inclusion levels of alternative exons that are enriched for genes encoding RNA processing [27]. *SNRPB*, which plays a central role in modulating expression levels of many RNA processing factors [27], might therefore have the strongest influence in the RS-score among RNA processing factors. Previous studies discovered the involvement of several splicing factors in RNA stability [36,37]. Thus, the core splicing factor *SNRPB* may have an important role in RNA stability as well.

Conclusions

Genetic variants that affect RNA stability in *cis* have been shown to contribute to inter-individual variation in gene expression [25]. Here we demonstrate that the effects of knocking down the expression of *HNRNPA2B1* that stabilizes a large number of RNAs can be detected from gene expression data. In particular, the expression of genes expressing transcripts with a long half-life is

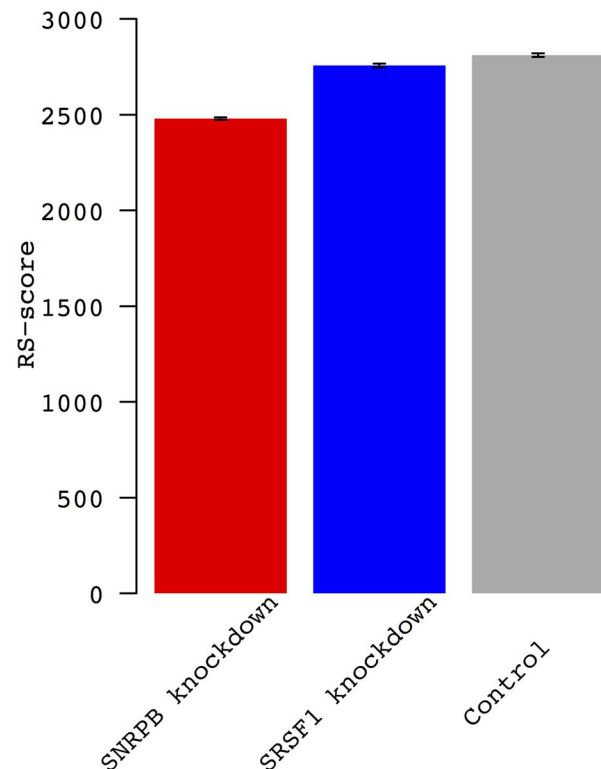


Figure 5. RS-scores calculated from three samples - *SNRPB* knockdown, *SRSF1* knockdown and control. The control corresponds to the sample transfected with nontargeting siRNA. Error bars represent two standard errors. doi:10.1371/journal.pone.0079627.g005

reduced relative to genes with short half-life transcripts. We defined the RS-score to summarize the relative expression of long-lived compared to short-lived transcripts. Treating the RS-score as a quantitative trait, we performed genome-wide association and identified a locus on chromosome 20p13 that is strongly associated

with the RS-score in two Asian populations. This locus is a *cis*-eQTL for *SNRPB*, a core component of the spliceosome that has previously been shown to affect the expression of many RNA processing factors [27]. We propose that the *cis*-eQTL of *SNRPB* may be directly responsible for the association of the RS-score with this locus. Consistent with this model, knockdown of *SNRPB* results in a significant reduction in the RS-score.

Methods

Data

Processed gene expression data generated using the Illumina whole genome expression array from 726 lymphoblastoid cell lines (LCLs) in eight HapMap3 populations (CEU, CHB, GIH, JPT, LWK, MEX, MKK, and YRI) by [26] were downloaded from ArrayExpress [38]. Single nucleotide polymorphisms (SNPs) for the same 726 individuals were obtained from HapMap3 (release 2) [39]. SNPs with minor allele frequency (MAF) $\leq 1\%$ in a population were excluded. This resulted in between 1.1 million and 1.3 million SNPs per population. Half-life data for 11,052 mRNAs and 1,418 ncRNAs in HeLa cells, and for 8,344 genes in B-cells were obtained from Tani *et al.* [15] and Friedel *et al.* [14], respectively.

RNA stability score

We defined the RNA stability score (RS-score), as a measure of the relative expression levels of long-lived and short-lived transcripts in a sample. We first classified all genes as either expressing long or short lived RNAs, by setting a threshold on an available RNA half-life or decay rate data set. Specifically, for the HeLa half-life data [15], we chose the same threshold used by the authors to determine whether a gene expresses long-lived (half-life ≥ 4 hours) or short-lived (half-life < 4 hours) RNA. For the RNA decay rate data [13], a gene was considered as expressing long-lived RNA if its decay rate was greater than the average across genes (corresponding to a relative decay rate greater than 0) and as short-lived if its decay rate was less than average (corresponding to values less than 0). We then ranked all genes in the sample by their expression levels (a higher expression level corresponds to higher rank value). Finally, the RS-score is defined as the difference in the mean rank of genes expressing long-lived RNAs and genes expressing short-lived RNAs. Therefore, higher RS-scores correspond to higher relative expression of genes with longer half-life, consistent with more efficient stabilization of RNA.

Genome-wide association test

Assuming an additive mode of inheritance, we performed linear regression analysis to assess association of RS-score with SNP genotypes, using PLINK v1.07 [40]. We included gender as a covariate in the linear model to correct for any sex bias. To combine samples from different populations, we carried out a principal component analysis (PCA) as implemented in the Eigensoft 4.2 [29,41]. To correct for population stratification in genome-wide association tests, we included the first five principal components in addition to gender as covariates in the linear models.

Permutation testing

Applying a permutation testing procedure by Hirschhorn and Daly [42], in each GWAS test, we carried out 1000 permutations. In each permutation, we randomly shuffled the phenotype values, re-ran the GWAS and recorded the best (lowest) p-value from each run. Finally, we counted how many of these 1000 lowest p-values are less than or equal to the original p-value being evaluated. The

permutation p is defined as this number divided by 1000 (i.e. the proportion of the 1000 lowest p-values that are less than or equal to the original p-value).

Analysis of RNA-seq data from *SNRPB* knockdown samples

We downloaded RNA-seq data generated by Saltzman *et al.* [27] from samples in which *SNRPB* or *SRSF1* were knocked down as well as control samples. The data consisted of three samples for each knock down and three control samples. We mapped the RNA-seq reads to the human genome, build hg19, using Tophat 1.4.1 (with default parameters) [43] and estimated expression levels of RefSeq genes using Cufflinks 1.3.0 (with default parameters) [44]. Using the HeLa half-life data [15], we calculated the RS-score for each of the three samples.

Supporting Information

Figure S1 Gene expression levels in HNRNPA2B1 knockdown relative to control are shown separately for genes expressing short-lived (golden) and long-lived (dark green) RNAs in three independent replicates (Rep1, Rep2, and Rep3). P-values are from Wilcoxon rank sum tests that were used to compare expression levels between these two groups of genes.

(TIFF)

Figure S2 First principal component (PC1) versus second principal component (PC2) for all 726 individuals from 8 populations.

(TIFF)

Figure S3 Manhattan plots for GWA with RS-score in different populations and combined populations. Each Manhattan plot shows the distribution of $-\log_{10}$ of the P-values from tests of association between individual SNP markers and the RS-score.

(TIFF)

Figure S4 P-P plots of the association with RS-score. The expected (X-axis) shows $-\log_{10}$ of random values, drawn from the uniform distribution. The observed (Y-axis) shows $-\log_{10}$ of the P-values from tests of association between individual SNP markers and the RS-score. The red line is used to compare the expected and observed values.

(TIFF)

Table S1 Summary of samples in the eight Hapmap3 populations.

(DOCX)

Table S2 Spearman correlation between HNRNPA2B1 and the RS-score.

(DOCX)

Table S3 Association between *cis*-eQTL of HNRNPA2B1 and the RS-score.

(DOCX)

Table S4 Genomic inflation factors (λ) in different populations.

(DOCX)

Table S5 Top GO terms for genes positively correlated with rs6137010 in CHB.

(DOCX)

Acknowledgments

We are grateful to anonymous reviewers, whose comments helped to improve the manuscript.

References

- Ross J (1995) mRNA stability in mammalian cells. *Microbiol Rev* 59: 423–450.
- Guhaniyogi J, Brewer G (2001) Regulation of mRNA stability in mammalian cells. *Gene* 265: 11–23.
- Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, et al. (2005) Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *Bmc Genomics* 6: 75.
- Lam LT, Pickeral OK, Peng AC, Rosenwald A, Hurt EM, et al. (2001) Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol* 2: RESEARCH0041.
- Bregman A, Avraham-Kelbert M, Barkai O, Duek L, Guterman A, et al. (2011) Promoter elements regulate cytoplasmic mRNA decay. *Cell* 147: 1473–1483.
- Trcek T, Larson DR, Moldón A, Query CC, Singer RH (2011) Single-molecule mRNA decay measurements reveal promoter-regulated mRNA stability in yeast. *Cell* 147: 1484–1497.
- Narsai R, Howell KA, Millar AH, O'Toole N, Small I, et al. (2007) Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. *Plant Cell* 19: 3418–3436.
- Sharova LV, Sharov AA, Nedorezov T, Piao Y, Shaik N, et al. (2009) Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Research* 16: 45–58.
- Clark MB, Johnston RL, Inostroza-Ponta M, Fox AH, Fortini E, et al. (2012) Genome-wide analysis of long noncoding RNA stability. *Genome Res* 22: 885–898.
- Schwahnhauser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global quantification of mammalian gene expression control. *Nature* 473: 337–342.
- Raghavan A, Ogilvie RL, Reilly C, Abelson ML, Raghavan S, et al. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res* 30: 5529–5538.
- Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, et al. (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* 13: 1863–1872.
- Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, et al. (2012) Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature* 485: 264–268.
- Friedel CC, Dolken L, Ruzsics Z, Koszinowski UH, Zimmer R (2009) Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Research* 37.
- Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, et al. (2012) Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res* 22: 947–956.
- Kontoyiannis D, Pasparakis M, Pizarro TT, Cominelli F, Kollias G (1999) Impaired on/off regulation of TNF biosynthesis in mice lacking TNF AU-rich. *Immunity* 10: 387–398.
- Misquitta CM, Iyer VR, Werstnik ES, Grover AK (2001) The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in. *Mol Cell Biochem* 224: 53–67.
- Eberhardt W, Doller A, Akool el S, Pfeilschifter J (2007) Modulation of mRNA stability as a novel therapeutic approach. *Pharmacol Ther* 114: 56–73.
- Schoenberg DR, Maquat LE (2012) Regulation of cytoplasmic mRNA decay. *Nat Rev Genet* 13: 246–259.
- Keene JD, Tenenbaum SA (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Mol Cell* 9: 1161–1167.
- Kishore S, Lubner S, Zavolan M (2010) Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief Funct Genomics* 9: 391–404.
- Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG (1993) hnRNP proteins and the biogenesis of mRNA. *Annu Rev Biochem* 62: 289–321.
- Chaudhury A, Chander P, Howe PH (2010) Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on. *Rna* 16: 1449–1462.

Author Contributions

Conceived and designed the experiments: CS. Performed the experiments: TTN. Analyzed the data: TTN. Contributed reagents/materials/analysis tools: TTN. Wrote the paper: TTN CS.

- Thiele BJ, Doller A, Kahne T, Pregla R, Hetzer R, et al. (2004) RNA-binding proteins heterogeneous nuclear ribonucleoprotein A1, E1, and K are involved in post-transcriptional control of collagen I and III synthesis. *Circ Res* 95: 1058–1066.
- Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, et al. (2012) The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. *PLoS genetics* 8: e1003000.
- Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, et al. (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* 8: e1002639.
- Saltzman AL, Pan Q, Blencowe BJ (2011) Regulation of alternative splicing by the core spliceosomal machinery. *Genes Dev* 25: 373–384.
- Cheng C, Fu XP, Alves P, Gerstein M (2009) mRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer. *Genome Biology* 10.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* 38: 904–909.
- Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, et al. (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One* 6: e22859.
- Ernst J, Kheradpour P, Mikkelsen TS, Shores N, Ward LD, et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43–49.
- Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* 41: D64–69.
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*: 289–300.
- R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing ISBN 3-900051-07-0.
- 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.
- Gupta SK, Carmi S, Waldman Ben-Asher H, Tkacz ID, Naboishchikov I, et al. (2013) Basal splicing factors regulate the stability of mature mRNAs in trypanosomes. *J Biol Chem* 288: 4991–5006.
- Lemaire R, Prasad J, Kashima T, Gustafson J, Manley JL, et al. (2002) Stability of a PKCI-1-related mRNA is controlled by the splicing factor ASF/SF2: a novel function for SR proteins. *Genes & Development* 16: 594–607.
- Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research* 35: D747–D750.
- Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
- Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2: e190.
- Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6: 95–108.
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511–U174.