Transcriptome-wide analysis of UTRs in non-small cell lung cancer reveals cancer-related genes with SNV-induced changes on RNA secondary structure and miRNA target sites

Radhakrishnan, Sabarinathan; Wenzel, Anne; Novotny, Peter; Tang, Xiaojia; Kalari, Krishna R.; Gorodkin, Jan

Published in:
P L o S One

DOI:
10.1371/journal.pone.0082699

Publication date:
2014

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Transcriptome-Wide Analysis of UTRs in Non-Small Cell Lung Cancer Reveals Cancer-Related Genes with SNV-Induced Changes on RNA Secondary Structure and miRNA Target Sites

Radhakrishnan Sabarinathan1,*, Anne Wenzel1,*, Peter Novotny2, Xiaojia Tang3, Krishna R. Kalari3,4, Jan Gorodkin1,8

1 Center for non-coding RNA in Technology and Health, Section for Animal Genetics, Bioinformatics and Breeding, IOVH, University of Copenhagen, Frederiksberg, Denmark, 2 Bioinformatics Centre, Department of Biology and Biotech Research and Innovation Centre, University of Copenhagen, Copenhagen, Denmark, 3 Division of Biostatistics and Bioinformatics, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America, 4 Department of Cancer Biology, Mayo Clinic Comprehensive Cancer Center, Jacksonville, Florida, United States of America

Abstract

Traditional mutation assessment methods generally focus on predicting disruptive changes in protein-coding regions rather than non-coding regulatory regions like untranslated regions (UTRs) of mRNAs. The UTRs, however, are known to have many sequence and structural motifs that can regulate translational and transcriptional efficiency and stability of mRNAs through interaction with RNA-binding proteins and other non-coding RNAs like microRNAs (miRNAs). In a recent study, transcriptomes of tumor cells harboring mutant and wild-type KRAS (V-Ki-ras2 Kirsten sarcoma viral oncogene homolog) genes in patients with non-small cell lung cancer (NSCLC) have been sequenced to identify single nucleotide variations (SNVs). About 40% of the total SNVs (73,717) identified were mapped to UTRs, but omitted in the previous analysis. To meet this obvious demand for analysis of the UTRs, we designed a comprehensive pipeline to predict the effect of SNVs on two major regulatory elements, secondary structure and miRNA target sites. Out of 29,290 SNVs in 6462 genes, we predict 472 SNVs (in 408 genes) affecting local RNA secondary structure, 490 SNVs (in 447 genes) affecting miRNA target sites and 48 that do both. Together these disruptive SNVs were present in 803 different genes, out of which 188 (23.4%) were previously known to be cancer-associated. Notably, this ratio is significantly higher (one-sided Fisher’s exact test p-value = 0.032) than the ratio (20.8%) of known cancer-associated genes (n = 1347) in our initial data set (n = 6462). Network analysis shows that the genes harboring disruptive SNVs were involved in molecular mechanisms of cancer, and the signaling pathways of LPS-stimulated MAPK, IL-6, iNOS, EIF2 and mTOR. In conclusion, we have found hundreds of SNVs which are highly disruptive with respect to changes in the secondary structure and miRNA target sites within UTRs. These changes hold the potential to alter the expression of known cancer genes or genes linked to cancer-associated pathways.

Introduction

Next-generation genome sequencing is now widely used for the identification of genetic variations in cancer genomes [1,2]. Non-small cell lung cancer (NSCLC) is the most common form of lung cancer and it is often found with activating mutations in the KRAS oncogene which causes the tumor cells to be aggressive and resistant to chemotherapy [3–5]. In a recent study, Kalari et al. [6] performed transcriptome-wide sequencing of NSCLC and identified differentially expressed genes, alternate splicing isoforms and single nucleotide variants (SNV) for tumors with and without KRAS mutations. A network analysis was performed with the genes showing differential expression (374 genes), alternate splicing (259 genes) and SNV-related changes (65 genes) that are differentially present in lung tumor groups with and without KRAS mutations. Integrated pathway analysis identified NFkB, ERK1/2 and AKT pathways as the most significant pathways differentially deregulated in KRAS wild-type as compared with KRAS mutated samples.

A single nucleotide variant (SNV) is a nucleotide change at a single base position that occurs at a low frequency (also referred as a rare variant). SNVs observed in tumor cells are mostly somatic variants and very few are germ-line variants. Genome-wide association studies (GWAS) report that SNVs mostly occur in non-coding regions compared to coding (exonic) regions of RNAs [7]. In the past, however, most studies have been focused on the effect of SNVs in coding regions (known as cSNVs or nsSNVs) [8] rather than the effect of SNVs in the regulatory non-coding DNA.
Predicting Functional Impact of UTR SNVs

The occurrence of genetic variation(s) in UTRs could potentially affect their sequence and/or structural motifs and thus lead to changes in post-transcriptional regulation [17–20]. For example, a SNP in a let-7 miRNA target site (miRTS) in the 3' UTR of KRAS has been identified to affect the binding of let-7 miRNAs. This result in the overexpression of KRAS, leading to the increased risk of NSCLC [21]. In addition, recent studies report that genetic variation can potentially create, change or destroy miRNA targets sites, which results in dysregulation of the target mRNA [22,23]. Notably this has been identified in tumor cells as well [24]. Furthermore, a cancer-driven mutation present in an IRES in human p53 mRNA alters the structure of the IRES element, which inhibits binding of a trans-acting factor essential for translation [25]. The recent number of web servers and data bases developed to deal with variants affecting miRTSs also demonstrates the growing importance of target site variants [22,26–28].

In this study, we predict the possible effects of 29,290 SNVs associated with NSCLC that are located in the UTR regions of mRNAs. The local effect of SNVs on the secondary structure and miRNA target sites, which is described in the following sections. A list of cancer-associated genes was obtained from COSMIC [36] and Qiagen/SABiosciences [37]. This list includes 1347 of the 6462 genes considered in this analysis. In order to find the enrichment of genes carrying disruptive SNVs, which have effect on secondary structure and/or miRTSs, in cancer, we performed a one-sided Fisher’s exact test. This was computed from a 2 x 2 contingency table (n = 6462) with the number of genes carrying/not carrying disruptive SNVs on the one side and the number of cancer-associated/other genes on the other side. Similarly, the enrichment for disruptive SNVs in cancer-associated genes was computed by counting the total number of SNVs (n = 29,290) into disruptive/non-disruptive SNVs on the one hand, and those being and not being present in cancer-associated genes on the other.

A set of experimentally verified examples of SNPs with effects on miRNA target sites has been extracted from the literature (see Table S1). These 19 SNPs (affecting 25 miRNA-mRNA interactions) have been used to test the filtration criteria used in the miRNA part of our pipeline.

\[ d_{\pi_1,\pi_2}(P, P') = \sum_{i=1}^{\pi_1} \sum_{j=1}^{\pi_2} (P_y - P'_y)^2, \]

where \( P_y \) is the probability of bases \( i \) and \( j \) being paired. The second measure uses the position-wise pair probabilities \( \pi \). For a local region \( [u, v] \), the vector \( \pi \) contains the elements \( \pi_i[u, v] = \sum_y P_y \). Then the difference between two vectors \( \pi \) and \( \pi^* \) is measured by...
Predicting Functional Impact of UTR SNVs

For each SNV in the data set, a subsequence of 30 nts on either side of the SNV position was retrieved. Further, all 2042 human mature miRNA sequences from miRBase (v19) [42] were used to scan for possible target sites in wild-type and mutant (with SNV) subsequences. As a first step, TargetScan (version 6.0) [30] was used to identify pairs of SNVs and miRNAs for which the type of seed match differs between wild-type and mutant or is only present in either of them. The different seed types used by TargetScan are 7mer-1a, 7mer-m8, and 8mer-1a (in increasing strength), where ‘1a’ refers to an adenosine in the miRTS 3’ to the seed match (i.e., opposite the first nucleotide of the miRNA) and ‘-m8’ refers to a Watson-Crick-matched nucleotide in position 8. Subsequently, the interaction energy of these pairs was computed using miRanda (version 3.3a) [31]. As a seed match change is already required by the TargetScan filtration, the parameters for miRanda were set to not weigh the seed region too high (‘-scale 2’ instead of default 4) and with relaxed cutoffs (score 45, energy ~5 kcal/mol), in order to capture cases where a poor seed match can be compensated. To classify an interaction as working we later apply a more robust computational method (see below).

Finally, a local region predicted with maximum Euclidean distance (d_{max}) or minimum Pearson correlation coefficient (r_{min}) and the corresponding p-value is then reported. We employ both measures independently as both measures hold their respective strengths and weaknesses (see [29] for details). We generated two lists (each with p < 0.1) of candidates, d_{max} and r_{min} and each of them is subjected to a multiple-testing correction using the Benjamini-Hochberg procedure [38], which limits the false discovery rate to be no more than a chosen threshold (typically 10%).

To analyze whether the RNAsnp predicted local region is structurally conserved, we used the annotations of conserved RNA secondary structure predictions from our in-house pipeline [39], which makes use of a range of tools including CMfinder [40] and RNAz [41] programs.

Predicting SNVs’ effect on microRNA target sites

Figure 1. Pipeline for the analysis of effect of SNVs on UTRs of mRNA.

doi:10.1371/journal.pone.0082699.g001
conservative energy threshold of $-11 \text{kcal/mol}$ based on our previous study [43]. For each pair of miRNA and 61mer, only the strongest binding site (lowest $\Delta G$) that differs between WT and SNV sequence is retained. Putative interactions are classified as created, destroyed or altered upon mutation based on miRanda predictions. The create set contains target sites that are induced by the SNV, i.e., have an interaction energy of $-11 \text{kcal/mol}$ or lower in the SNV variant, while no interaction is predicted in the wild-type (either due to score or energy threshold). Similarly, a loss of target site would be recorded in the destroy set, if the interaction is predicted for the wild-type but not with the SNV. Finally, the alter set contains putative interactions that are predicted with a binding energy of at most $-11 \text{kcal/mol}$ for at least one variant. For these, the energy difference observed for the binding of miRNA before and after SNV introduction was measured as their log-ratio $b = \text{ld} (\Delta G_{\text{SNV}} / \Delta G_{\text{WT}})$, for $\Delta G_{\text{WT}} > 0$. The $b$ is 0 if there is no change in energy; negative if the wild type has the stronger interaction (lower energy), positive otherwise. Given the size of the data set, we focus on the (top) candidates whose absolute $b$ value is above the mean ($\mu$) of absolute $b$ values from all pairs classified as alter. The efficiency of this threshold clearly varies with the data, but it will always retain the top candidates with highest relative energy difference. Even though it should not be seen as a fixed cut-off, we applied it to our set of known examples, where 14 out of 23 interactions exceed the value we applied here (see Table S1). Threshold values based on the distribution of MFE changes have been used in a similar way before [24].

In order to reduce false positive predictions, the miRNA target sites predicted for the wild-type (destroy or alter) were cross-checked with experimentally identified microRNA-target interaction maps. These data, derived through Ago CLIP-Seq, was downloaded from starBase [44]. Only SNVs that are located inside stringent Ago CLIP-Seq peak clusters with a biological complexity (BC) of at least two were retained. This filter cannot be used for the interactions from the create set, as CLIP-Seq data is available for the wild-type only.

Finally, the set of miRNAs was filtered for those expressed in the respiratory system (lung and trachea) according to the miRNA body map [45]. The overview of the miRNA analysis is described in Figure 2.

From the PhenomiR [46] database, we retrieved information about miRNAs that have been found to be up- or down regulated in lung cancer. This set comprises 264 individual miRNA stem-loop accessions, 3 of which are ‘dead entries’. The remaining 261 stem-loops give rise to 430 mature miRNA products, which we refer to as lung cancer-associated miRNAs. In this data set, 27 miRNAs are specific to NSCLC [47] type according to miRNA body map [45]. The later data set is referred as NSCLC-associated miRNAs.

Figure 2. Pipeline for the analysis of SNVs’ effect on miRNA target sites in more detail (dashed box from Figure 1). The flow chart shows the different steps of prediction and filtration with the number of individual SNVs, miRNAs, and pairs of these at each stage.

doi:10.1371/journal.pone.0082699.g002
Ingenuity Pathways Analysis

Interactome networks of candidate genes were constructed using Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, www.ingenuity.com; build version: 220217; content version: 16542223). Network generation is based on the 'Global Molecular Network' in IPA, which comprises an extensive, manually curated set of gene-gene relationships based on findings from the scientific literature. The genes of interest (candidates put out by our pipeline and used as input for IPA) that are also present in this global network are the so-called focus genes. Highly-interconnected focus genes are the starting points in network generation. Additional non-focus genes from the Global Molecular Network might be used as linker genes between small networks. Networks are extended until an approximate size of 35 genes, which is considered optimal for visualization and interpretation (for details see [48]). The p-value computed for each network represents the probability to find the same (or higher) number of focus genes in a randomly selected set of genes from the global network. It is computed by a right-tailed Fisher Exact Test with (non-)focus molecules on the one side and molecules (not) in the network on the other side of a 2x2 contingency table. This is transformed into a score which is the negative log of the p-value. Furthermore, IPA was used to identify the top diseases and disorders, molecular and cellular functions, and canonical pathways associated with the genes in our candidate sets (so called 'focus genes'). The p-value for a given (disease, function or pathway) annotation describes the likelihood that the association between the input gene set and the annotation is due to random chance. This is also based on a right-tailed Fisher’s exact test as specified above.

Results

Effect of SNVs on RNA secondary structure

The structural effects of 29,290 UTR SNVs were predicted using RNAsnp (v1.1). Both the Euclidean distance (\(d_{\text{max}}\)) and Pearson Correlation Coefficient (\(r_{\text{min}}\)) measures of RNAsnp (mode 1) were independently employed to predict the effect of SNVs on RNA secondary structure. Table 1 lists 28 high-confidence SNVs with p-value < 0.05 predicted by both \(d_{\text{max}}\) and \(r_{\text{min}}\) measures of RNAsnp.

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV</th>
<th>RNAsnp (d_{\text{max}}) (p-value)</th>
<th>% overlap with conserved secondary structure(a)</th>
<th>RNAsnp (r_{\text{min}}) (p-value)</th>
<th>% overlap with conserved secondary structure</th>
<th>dbSNP 135</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSR</td>
<td>NM_001195102</td>
<td>3</td>
<td>A2638G</td>
<td>0.0030</td>
<td>89(b)</td>
<td>0.0213</td>
<td>100(b)</td>
<td>rs1138092</td>
</tr>
<tr>
<td>MEF2A</td>
<td>NM_005587</td>
<td>3</td>
<td>A2046U</td>
<td>0.0032</td>
<td>100(b)</td>
<td>0.0238</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>PPM1A</td>
<td>NM_177952</td>
<td>3</td>
<td>G2231A</td>
<td>0.0059</td>
<td>90(a)</td>
<td>0.0118</td>
<td>100(a)</td>
<td></td>
</tr>
<tr>
<td>MAP1K1</td>
<td>NM_139012</td>
<td>3</td>
<td>A2304C</td>
<td>0.0076</td>
<td>100(a,96)</td>
<td>0.0224</td>
<td>91(a,91)</td>
<td></td>
</tr>
<tr>
<td>PHC2</td>
<td>NM_198040</td>
<td>3</td>
<td>A3730C</td>
<td>0.0105</td>
<td>-</td>
<td>0.0137</td>
<td>75(b)</td>
<td></td>
</tr>
<tr>
<td>BECN1</td>
<td>NM_003766</td>
<td>3</td>
<td>U1970C</td>
<td>0.0120</td>
<td>100(b)</td>
<td>0.0487</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>NFKBIE</td>
<td>NM_004556</td>
<td>3</td>
<td>G1659C</td>
<td>0.0144</td>
<td>84(a)</td>
<td>0.0473</td>
<td>100(a)</td>
<td></td>
</tr>
<tr>
<td>MAPK1</td>
<td>NM_002745</td>
<td>3</td>
<td>U2360G</td>
<td>0.0148</td>
<td>100(a)</td>
<td>0.0262</td>
<td>100(a)</td>
<td>rs13058</td>
</tr>
<tr>
<td>DHC2R4</td>
<td>NM_014762</td>
<td>3</td>
<td>A4192C</td>
<td>0.0159</td>
<td>100(b)</td>
<td>0.0494</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>ADAMTS1</td>
<td>NM_006988</td>
<td>3</td>
<td>U4320G</td>
<td>0.0166</td>
<td>94(b)</td>
<td>0.0215</td>
<td>83(b)</td>
<td></td>
</tr>
<tr>
<td>SRF</td>
<td>NM_003131</td>
<td>3</td>
<td>C3504U</td>
<td>0.0173</td>
<td>100(b)</td>
<td>0.0074</td>
<td>100(b)</td>
<td>rs3734681</td>
</tr>
<tr>
<td>CASP2</td>
<td>NM_032982</td>
<td>3</td>
<td>U2139C</td>
<td>0.0189</td>
<td>56(b)</td>
<td>0.0230</td>
<td>62(b)</td>
<td></td>
</tr>
<tr>
<td>LIFNG</td>
<td>NM_001040167</td>
<td>3</td>
<td>C1381G</td>
<td>0.0215</td>
<td>64(b)</td>
<td>0.0333</td>
<td>-</td>
<td>rs4721752</td>
</tr>
<tr>
<td>SH3PXD2A</td>
<td>NM_014631</td>
<td>3</td>
<td>C8560U</td>
<td>0.0223</td>
<td>100(b)</td>
<td>0.0073</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>KITLG</td>
<td>NM_000899</td>
<td>3</td>
<td>U1057G</td>
<td>0.0226</td>
<td>50(b,84)</td>
<td>0.0080</td>
<td>90(b)</td>
<td></td>
</tr>
<tr>
<td>PRKAB1</td>
<td>NM_006253</td>
<td>3</td>
<td>U1875C</td>
<td>0.0237</td>
<td>100(b)</td>
<td>0.0462</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>TFG</td>
<td>NM_001195479</td>
<td>5</td>
<td>G309C</td>
<td>0.0256</td>
<td>-</td>
<td>0.0398</td>
<td>56(b)</td>
<td></td>
</tr>
<tr>
<td>FTH1</td>
<td>NM_002032</td>
<td>3</td>
<td>U819G</td>
<td>0.0262</td>
<td>93(b)</td>
<td>0.0259</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>BCL2L2</td>
<td>NM_001199839</td>
<td>3</td>
<td>C2469A</td>
<td>0.0275</td>
<td>78(b)</td>
<td>0.0294</td>
<td>100(b)</td>
<td>rs3210043</td>
</tr>
<tr>
<td>CDKN1C</td>
<td>NM_000076</td>
<td>3</td>
<td>G1334C</td>
<td>0.0290</td>
<td>87(b)</td>
<td>0.0253</td>
<td>87(b)</td>
<td></td>
</tr>
<tr>
<td>TIA1</td>
<td>NM_0022173</td>
<td>3</td>
<td>U4082A</td>
<td>0.0316</td>
<td>100(b)</td>
<td>0.0363</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>NFKBIE</td>
<td>NM_004556</td>
<td>3</td>
<td>U1644G</td>
<td>0.0333</td>
<td>98(b)</td>
<td>0.0347</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>DAPK3</td>
<td>NM_001348</td>
<td>3</td>
<td>G1662U</td>
<td>0.0334</td>
<td>54(b)</td>
<td>0.0216</td>
<td>94(b)</td>
<td>rs3745982</td>
</tr>
<tr>
<td>NCOA1</td>
<td>NM_003743</td>
<td>3</td>
<td>C4893G</td>
<td>0.0342</td>
<td>100(b)</td>
<td>0.0176</td>
<td>100(b)</td>
<td>rs17737058</td>
</tr>
<tr>
<td>PCBP4</td>
<td>NM_001174100</td>
<td>3</td>
<td>C1790G</td>
<td>0.0413</td>
<td>59(b)</td>
<td>0.0187</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>SH3PXD2A</td>
<td>NM_014631</td>
<td>3</td>
<td>U8562A</td>
<td>0.0451</td>
<td>100(b)</td>
<td>0.0096</td>
<td>100(b)</td>
<td></td>
</tr>
<tr>
<td>ID2</td>
<td>NM_002166</td>
<td>5</td>
<td>C143G</td>
<td>0.0464</td>
<td>87(b)</td>
<td>0.0474</td>
<td>83(b)</td>
<td></td>
</tr>
<tr>
<td>GPX3</td>
<td>NM_002084</td>
<td>3</td>
<td>U1552G</td>
<td>0.0474</td>
<td>73(b)</td>
<td>0.0427</td>
<td>100(b)</td>
<td></td>
</tr>
</tbody>
</table>

\(a\)The conserved RNA secondary structure predicted by CMfinder and RNAz programs (through our in-house pipeline [39]) are highlighted with the symbols \(\#\) and \(\$\), respectively.

doi:10.1371/journal.pone.0082699.t001

Table 1. List of 28 high-confidence SNVs with p-value < 0.05 predicted by both \(d_{\text{max}}\) and \(r_{\text{min}}\) measures of RNAsnp.
local RNA secondary structure. The distribution of p-values calculated for the 29,290 UTR SNVs is shown in Figure S1A. At a significance level of 0.1 (chosen from our previous study [29]), 3237 and 3062 SNVs were predicted, respectively, by $d_{\text{max}}$ and $r_{\text{min}}$ measures. Further, the adjustment for multiple comparisons (using Benjamini–Hochberg procedure [38]) provided 3204 and 1813 SNVs respective to $d_{\text{max}}$ and $r_{\text{min}}$ measures. After fusing these two lists, we got 3561 unique SNVs in 2411 genes.

Further, we calculated the distance between the location of these 3561 SNVs and the predicted local region where the maximum structural change was detected (Figure S1B). It shows that the majority of the SNVs cause structural change in and around the SNV position. In addition, the length distribution of the predicted local region shows that the majority of SNVs have effect on the local region of size 50 to 100 nts, however, certain SNVs ($n = 47$) have effect on a global structure where the size of predicted local region exceeds 300 nts (Figure S1C). Furthermore, we checked whether these disruptive SNVs are enriched in GC or AU rich regions, as sequences with such biased nucleotide content have been shown more sensitive to structural changes caused by mutations [49]. For each SNV we computed the GC content of its flanking regions (as previously using 200 nts up- and downstream), see Materials and Methods. This showed that both the data set SNVs and the disruptive SNVs are highly enriched in the regions with GC content ranging from 40 to 60 percent (see Figure S2), which should therefore make them less sensitive to variations. In addition, we found that there were no significant differences between GC content distributions of disruptive SNVs and the data set SNVs (see Figure S2 with Kolmogorov-Smirnov).

It is known that the UTRs of mRNAs harbor evolutionarily conserved regulatory elements ([50–52], see also reviews [53,54]). Thus, we cross-checked for the overlap between the disrupted local region predicted by RNAsnp and the conserved RNA secondary structures predicted using our in-house pipeline [39] (see Material and Methods sections for details). Interestingly, the local region predicted for 472 SNVs (p-value<0.1) overlap with the predicted conserved RNA secondary structures. These 472 SNVs correspond to 408 genes; out of which 111 SNVs

---

**Figure 3. Results of SNV U1552G predicted to cause significant local secondary structure changes in 3’ UTR of GPX3 mRNA.** The dot plot from RNAsnp web server [67] shows the base pair probabilities corresponds to the local region predicted with significant difference ($d_{\text{max}}$ p-value: 0.0474) between wild-type and mutant. The upper triangle represents the base pair probabilities for the wild-type (green) and the lower triangle for the mutant (red). On the sides, the minimum free energy (MFE) structure of the wild-type and mutants are displayed in planar graphic representation. The SECIS region is highlighted in blue circle and the SNV position is indicated with arrow mark.

doi:10.1371/journal.pone.0082699.g003
correspond to 98 genes that are involved in cancer-associated pathways (see File S1.xlsx).

Based on the p-value, the above 111 SNVs were further classified into two groups: 28 as high-confidence for which both \(d_{\text{max}}\) and \(r_{\text{min}}\) p-value \(< 0.05\) (Table 1), and the other 83 as medium-confidence (either \(d_{\text{max}}\) or \(r_{\text{min}}\) p-value \(< 0.1\)) (see File S2.xlsx). We predict that the SNV-induced structural changes in the UTR regions could potentially affect the stability of the mRNA or disrupt the function of regulatory elements present in the UTRs. For example, the SNV A2304C (Table 1) present in the 3' UTR of MAPK14 mRNA shows a significant structural change (p-value: 0.0076) in the local RNA secondary structure which is structurally conserved according to both CMfinder and RNAz predictions from our in-house pipeline [39]. This structural conservation shows that the region is under evolutionary pressure to maintain the structure which is likely to have some functional importance. The protein encoded by MAPK14 gene is a member of the MAP kinase family, which is known to be involved in many pathways related to cell division, maturation and differentiation (reviewed in [55]). Also, it has been predicted to be one of the key players in the lung cancer interactome [6]. Thus the alteration in the gene expression of MAPK14 at a post-transcriptional level due to the SNV-induced structural change could potentially affect the MAKP14-related signaling pathways.

As another example, the gene GPX3 is responsible for the coding of plasma glutathione peroxidase, an antioxidant enzyme that contains selenocysteine in its active site and catalyzes the reduction of hydrogen peroxide. The amino acid selenocysteine is encoded by the UGA codon, which normally functions as a stop codon. In the GPX3 mRNA, the alternate recognition of a UGA

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV*</th>
<th>RNAsnp (p-value)*</th>
<th>% overlap of predicted local region with conserved RNA secondary structure</th>
<th>dbSNP 135</th>
</tr>
</thead>
<tbody>
<tr>
<td>SH3PD2A</td>
<td>NM_014631</td>
<td>3</td>
<td>C8560U</td>
<td>0.0223</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>SH3PD2A</td>
<td>NM_014631</td>
<td>3</td>
<td>U8562A</td>
<td>0.0451</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>MAPK1</td>
<td>NM_002745</td>
<td>3</td>
<td>G1633A</td>
<td>0.0815</td>
<td>100(^\d)</td>
<td>rs41282607</td>
</tr>
<tr>
<td>MAPK1</td>
<td>NM_002745</td>
<td>3</td>
<td>U2360G</td>
<td>0.0148</td>
<td>100(^\d)</td>
<td>rs13058</td>
</tr>
<tr>
<td>ACOX1</td>
<td>NM_004035</td>
<td>3</td>
<td>U4708G</td>
<td>0.0650</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>ACOX1</td>
<td>NM_004035</td>
<td>3</td>
<td>A6386U</td>
<td>0.0750</td>
<td>59(^*)</td>
<td></td>
</tr>
<tr>
<td>ADAMTS1</td>
<td>NM_006988</td>
<td>3</td>
<td>U4320G</td>
<td>0.0166</td>
<td>94(^\d)</td>
<td></td>
</tr>
<tr>
<td>ADAMTS1</td>
<td>NM_006988</td>
<td>3</td>
<td>U3449C</td>
<td>0.0626</td>
<td>86(^\d)</td>
<td></td>
</tr>
<tr>
<td>CDC42</td>
<td>NM_001039802</td>
<td>5</td>
<td>C159A</td>
<td>0.0665</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>CDC42</td>
<td>NM_001039802</td>
<td>5</td>
<td>G152A</td>
<td>0.0901</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>ID2</td>
<td>NM_002166</td>
<td>5</td>
<td>C143G</td>
<td>0.0464</td>
<td>87(^\d)</td>
<td></td>
</tr>
<tr>
<td>ID2</td>
<td>NM_002166</td>
<td>5</td>
<td>C129G</td>
<td>0.069</td>
<td>87(^\d)</td>
<td></td>
</tr>
<tr>
<td>NFKBIE</td>
<td>NM_004556</td>
<td>3</td>
<td>G1659C</td>
<td>0.0144</td>
<td>84(^\d)</td>
<td></td>
</tr>
<tr>
<td>NFKBIE</td>
<td>NM_004556</td>
<td>3</td>
<td>U1644G</td>
<td>0.0333</td>
<td>98(^\d)</td>
<td></td>
</tr>
<tr>
<td>RASSF1</td>
<td>NM_170714</td>
<td>3</td>
<td>A1907U</td>
<td>0.0629</td>
<td>64(^\d)</td>
<td></td>
</tr>
<tr>
<td>RASSF1</td>
<td>NM_170714</td>
<td>3</td>
<td>G1904A</td>
<td>0.0659</td>
<td>64(^\d)</td>
<td></td>
</tr>
<tr>
<td>RXRB</td>
<td>NM_021976</td>
<td>3</td>
<td>U2066G</td>
<td>0.0268</td>
<td>59(^\d)</td>
<td>rs2744537</td>
</tr>
<tr>
<td>RXRB</td>
<td>NM_021976</td>
<td>3</td>
<td>U2053A</td>
<td>0.0452</td>
<td>73(^\d)</td>
<td>rs5030979</td>
</tr>
<tr>
<td>PCBP4</td>
<td>NM_001174100</td>
<td>3</td>
<td>U1862G</td>
<td>0.0401</td>
<td>93(^\d)</td>
<td></td>
</tr>
<tr>
<td>PCBP4</td>
<td>NM_001174100</td>
<td>3</td>
<td>C1790G</td>
<td>0.0413</td>
<td>59(^\d)</td>
<td></td>
</tr>
<tr>
<td>MTA2</td>
<td>NM_004739</td>
<td>5</td>
<td>A227G</td>
<td>0.0462</td>
<td>73(^\d)</td>
<td></td>
</tr>
<tr>
<td>BECN1</td>
<td>NM_003766</td>
<td>3</td>
<td>U1970C</td>
<td>0.012</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>CT3B</td>
<td>NM_147782</td>
<td>3</td>
<td>A2561G</td>
<td>0.0569</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>HTT</td>
<td>NM_002111</td>
<td>3</td>
<td>C9948G</td>
<td>0.0987</td>
<td>100(^\d)</td>
<td>rs362305</td>
</tr>
<tr>
<td>HTT</td>
<td>NM_002111</td>
<td>3</td>
<td>U9947C</td>
<td>0.0225*</td>
<td>100(^\d)</td>
<td></td>
</tr>
<tr>
<td>BECN1</td>
<td>NM_003766</td>
<td>3</td>
<td>G2053A</td>
<td>0.0329*</td>
<td>98(^\d)</td>
<td>rs11552193</td>
</tr>
<tr>
<td>LMAN2</td>
<td>NM_032737</td>
<td>3</td>
<td>A3713G</td>
<td>0.0554*</td>
<td>59(^\d)</td>
<td></td>
</tr>
<tr>
<td>LMAN2</td>
<td>NM_032737</td>
<td>3</td>
<td>U3662C</td>
<td>0.0638*</td>
<td>57(^\d)</td>
<td></td>
</tr>
<tr>
<td>CT3B</td>
<td>NM_147782</td>
<td>3</td>
<td>A2581G</td>
<td>0.0925*</td>
<td>50(^\d)</td>
<td></td>
</tr>
<tr>
<td>MTA2</td>
<td>NM_004739</td>
<td>5</td>
<td>C267G</td>
<td>0.1035*</td>
<td>53(^\d)</td>
<td></td>
</tr>
</tbody>
</table>

*SNVs that were predicted by both \(d_{\text{max}}\) and \(r_{\text{min}}\) measures are highlighted with \(\dagger\).

The p-value corresponding to the \(r_{\text{min}}\) measure is highlighted with \(*\).

The conserved RNA secondary structure predicted by CMfinder and RNAz program (through our in-house pipeline [39]) are highlighted with the symbols \(\dagger\) and \(^\d\), respectively.

doi:10.1371/journal.pone.0082699.t002
### Table 3. List of genes which have more than one miRNA target site change (create, alter, destroy) in their UTRs.

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV</th>
<th>dbSNP 135</th>
<th>miRNA(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTR3</td>
<td>NM_005721</td>
<td>3</td>
<td>G2078A</td>
<td>rs6642</td>
<td>mir-662</td>
</tr>
<tr>
<td>ACTR3</td>
<td>NM_005721</td>
<td>5</td>
<td>U232G</td>
<td></td>
<td>mir-18a-3p</td>
</tr>
<tr>
<td>AMD1</td>
<td>NM_001033059</td>
<td>5</td>
<td>A262G</td>
<td></td>
<td>mir-1236-3p</td>
</tr>
<tr>
<td>AMD1</td>
<td>NM_001033059</td>
<td>5</td>
<td>A263G</td>
<td></td>
<td>mir-1236-3p</td>
</tr>
<tr>
<td>ARL5B</td>
<td>NM_178815</td>
<td>3</td>
<td>A947U</td>
<td></td>
<td>mir-409-3p</td>
</tr>
<tr>
<td>ARL5B</td>
<td>NM_178815</td>
<td>3</td>
<td>G2609U</td>
<td>rs12096599</td>
<td>mir-362-3p</td>
</tr>
<tr>
<td>ARL5B</td>
<td>NM_178815</td>
<td>3</td>
<td>U946C</td>
<td></td>
<td>mir-124-5p, mir-599</td>
</tr>
<tr>
<td>BCL2L13</td>
<td>NM_001270731</td>
<td>3</td>
<td>U1850G</td>
<td>rs725768</td>
<td>mir-361-3p</td>
</tr>
<tr>
<td>BCL7A</td>
<td>NM_001024808</td>
<td>3</td>
<td>G1801C</td>
<td></td>
<td>mir-650</td>
</tr>
<tr>
<td>CALM1</td>
<td>NM_006888</td>
<td>3</td>
<td>A1872G</td>
<td>rs63576962</td>
<td>mir-211-5p</td>
</tr>
<tr>
<td>CBX1</td>
<td>NM_001127226</td>
<td>3</td>
<td>A1839U</td>
<td>rs6847</td>
<td>mir-548b-5p, mir-548c-5p, mir-548d-5p</td>
</tr>
<tr>
<td>CBX5</td>
<td>NM_012117</td>
<td>3</td>
<td>A2592G</td>
<td></td>
<td>mir-887</td>
</tr>
<tr>
<td>CCND2</td>
<td>NM_001759</td>
<td>3</td>
<td>C11158U</td>
<td></td>
<td>mir-654-5p</td>
</tr>
<tr>
<td>IGFBP5</td>
<td>NM_000599</td>
<td>3</td>
<td>G2493C</td>
<td></td>
<td>mir-675-5p</td>
</tr>
<tr>
<td>IGFBP5</td>
<td>NM_000599</td>
<td>3</td>
<td>G3898U</td>
<td>rs13403592</td>
<td>mir-29a-3p, mir-29b-3p, mir-29c-3p</td>
</tr>
<tr>
<td>JAK1</td>
<td>NM_002227</td>
<td>3</td>
<td>U5007A</td>
<td></td>
<td>mir-106a-5p, mir-106b-5p, mir-17-5p, mir-20a-5p, mir-20b-5p, mir-519a-3p, mir-519b-3p, mir-519c-3p, mir-519d, mir-520g, mir-520h, mir-526b-3p, mir-93-5p</td>
</tr>
<tr>
<td>KLF10</td>
<td>NM_005655</td>
<td>3</td>
<td>C2615A</td>
<td>rs6935</td>
<td>mir-337-3p, mir-614</td>
</tr>
<tr>
<td>KRAS</td>
<td>NM_033360</td>
<td>3</td>
<td>U1049G</td>
<td>rs712</td>
<td>mir-151a-5p, mir-877-5p</td>
</tr>
<tr>
<td>KREMEN1</td>
<td>NM_001039570</td>
<td>3</td>
<td>A5345C</td>
<td></td>
<td>mir-519a-3p, mir-520b, mir-520c-3p, mir-636</td>
</tr>
<tr>
<td>LMNB2</td>
<td>NM_032737</td>
<td>3</td>
<td>C2928G</td>
<td></td>
<td>mir-423-5p</td>
</tr>
<tr>
<td>LMNB2</td>
<td>NM_032737</td>
<td>3</td>
<td>C2929A</td>
<td></td>
<td>mir-423-5p</td>
</tr>
<tr>
<td>NCK2</td>
<td>NM_003581</td>
<td>3</td>
<td>G1974U</td>
<td></td>
<td>mir-137, mir-488-3p</td>
</tr>
<tr>
<td>NDUFB7</td>
<td>NM_004146</td>
<td>5</td>
<td>G39C</td>
<td>rs45628939</td>
<td>mir-192-5p, mir-215</td>
</tr>
<tr>
<td>NR1D2</td>
<td>NM_005126</td>
<td>3</td>
<td>A2782G</td>
<td></td>
<td>mir-338-5p</td>
</tr>
<tr>
<td>NR1D2</td>
<td>NM_005126</td>
<td>3</td>
<td>U2791C</td>
<td></td>
<td>mir-504</td>
</tr>
<tr>
<td>P4HA1</td>
<td>NM_0017962</td>
<td>5</td>
<td>C174G</td>
<td></td>
<td>mir-412</td>
</tr>
<tr>
<td>P4HA1</td>
<td>NM_00142595</td>
<td>5</td>
<td>C174G</td>
<td></td>
<td>mir-412</td>
</tr>
<tr>
<td>PANX1</td>
<td>NM_015368</td>
<td>3</td>
<td>G2082A</td>
<td>rs1046805</td>
<td>mir-10a-5p, mir-10b-5p</td>
</tr>
<tr>
<td>PBX1</td>
<td>NM_001204961</td>
<td>3</td>
<td>A4035G</td>
<td></td>
<td>mir-188-5p</td>
</tr>
<tr>
<td>PBX1</td>
<td>NM_001204961</td>
<td>3</td>
<td>G2877U</td>
<td></td>
<td>mir-187-5p, mir-222-3p</td>
</tr>
<tr>
<td>PBX1</td>
<td>NM_001204963</td>
<td>3</td>
<td>A3249G</td>
<td>rs12723035</td>
<td>mir-326, mir-330-5p</td>
</tr>
<tr>
<td>PDPK1</td>
<td>NM_26213</td>
<td>3</td>
<td>U2034G</td>
<td></td>
<td>mir-504, mir-518a-3p, mir-518b, mir-518d-3p, mir-518f-3p</td>
</tr>
<tr>
<td>PRKAB2</td>
<td>NM_003399</td>
<td>3</td>
<td>U4199G</td>
<td></td>
<td>mir-150-5p, mir-532-3p</td>
</tr>
<tr>
<td>PTBP1</td>
<td>NM_0002827</td>
<td>3</td>
<td>G125A</td>
<td>rs118042879</td>
<td>mir-141-5p, mir-942</td>
</tr>
<tr>
<td>RAP1A</td>
<td>NM_02884</td>
<td>3</td>
<td>C1100A</td>
<td>rs6573</td>
<td>mir-135a-3p, mir-196a-5p, mir-196b-5p</td>
</tr>
<tr>
<td>SDC4</td>
<td>NM_02999</td>
<td>3</td>
<td>U1874G</td>
<td></td>
<td>mir-361-3p</td>
</tr>
<tr>
<td>SDC4</td>
<td>NM_02999</td>
<td>3</td>
<td>U1878G</td>
<td></td>
<td>mir-548-3p</td>
</tr>
<tr>
<td>SESN2</td>
<td>NM_30459</td>
<td>3</td>
<td>A2864C</td>
<td>rs10494394</td>
<td>mir-182-5p, mir-92a-1-5p, mir-96-5p</td>
</tr>
<tr>
<td>SLC39A6</td>
<td>NM_012349</td>
<td>3</td>
<td>C3543U</td>
<td></td>
<td>mir-144-3p</td>
</tr>
<tr>
<td>SLC39A6</td>
<td>NM_012349</td>
<td>3</td>
<td>G3543U</td>
<td></td>
<td>mir-101-3p, mir-139-5p, mir-144-3p</td>
</tr>
<tr>
<td>SMAD5</td>
<td>NM_005903</td>
<td>3</td>
<td>C2825A</td>
<td></td>
<td>mir-124-3p, mir-500a-3p, mir-501-3p, mir-502-3p, mir-506-3p</td>
</tr>
</tbody>
</table>
codon as a selenocysteine codon is mediated by the cis-acting regulatory element, selenocysteine insertion sequence (SECIS), present in the 3' UTR and other trans-acting co-factors [56]. The SNV U1552G (Table 1) located in the 3'9 UTR of GPX3 mRNA was predicted to cause significant structural effect (p-value: 0.0474) in the local region which contains the SECIS regulatory element. Figure 3 shows the base pair probabilities corresponding to the local region (NM_002084:1544 to 1692) of wild-type and mutant mRNA. It can be seen that the wild-type has higher base pair probabilities to form the stable stem-loop structure of SECIS (highlighted with a circle in Figure 3), whereas in the mutant form it is disrupted due to the SNV, which is located outside the SECIS region. Previous study has shown that the characteristic stem-loop structure of SECIS is essential for the efficiency of UGA recoding in vivo and in vitro [56]. Based on this, we speculate that the SNV

### Table 3. Cont.

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV</th>
<th>dbSNP 135</th>
<th>miRNA(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMNDC1</td>
<td>NM_005871</td>
<td>3</td>
<td>G1228A</td>
<td>rs1050755</td>
<td>miR-329, miR-362-3p</td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td></td>
<td>miR-30a-3p, miR-30d-3p, miR-30e-3p, miR-452-3p</td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td></td>
<td>miR-30a-3p, miR-30d-3p, miR-30e-3p, miR-595</td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>U2474A</td>
<td></td>
<td>miR-30a-3p, miR-30e-3p</td>
</tr>
<tr>
<td>TNFRSF19</td>
<td>NM_00120458</td>
<td>3</td>
<td>A2148U</td>
<td></td>
<td>miR-766-3p</td>
</tr>
<tr>
<td>TNFRSF19</td>
<td>NM_00120458</td>
<td>3</td>
<td>A2452U</td>
<td>rs79570196</td>
<td>miR-26a-5p</td>
</tr>
<tr>
<td>TOMM20</td>
<td>NM_014765</td>
<td>3</td>
<td>A3198C</td>
<td></td>
<td>miR-149-3p</td>
</tr>
<tr>
<td>TOMM20</td>
<td>NM_014765</td>
<td>3</td>
<td>U3378G</td>
<td></td>
<td>miR-129-1-3p, miR-129-2-3p</td>
</tr>
<tr>
<td>TOMM20</td>
<td>NM_014765</td>
<td>3</td>
<td>U3379G</td>
<td></td>
<td>miR-150-5p, miR-532-3p</td>
</tr>
</tbody>
</table>

The miRNA IDs are boldface if the interaction is predominant in the wild-type (destroy or alter with \( lr < 0.276 \)) and italics if the interaction is specific to the mutant (create or alter with \( lr > 0.276 \)); the hsa- prefix is omitted for brevity.

doi:10.1371/journal.pone.0082699.t003

### Table 4. List of miRNAs with more than two targets in the filtered data set.

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV</th>
<th>miRNA</th>
<th>( \Delta G_{WT} )</th>
<th>( \Delta G_{SNV} )</th>
<th>dbSNP 135</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALM1</td>
<td>NM_006888</td>
<td>3</td>
<td>C2472G</td>
<td>hsa-miR-29a-3p</td>
<td>N/A</td>
<td>−16.70</td>
<td></td>
</tr>
<tr>
<td>CKLF</td>
<td>NM_016326</td>
<td>5</td>
<td>U72G</td>
<td>hsa-miR-29a-3p</td>
<td>N/A</td>
<td>−20.70</td>
<td></td>
</tr>
<tr>
<td>IGFBP5</td>
<td>NM_000599</td>
<td>3</td>
<td>G3898U</td>
<td>hsa-miR-29a-3p</td>
<td>N/A</td>
<td>−13.83</td>
<td>rs13403592</td>
</tr>
<tr>
<td>CALM1</td>
<td>NM_006888</td>
<td>3</td>
<td>C2472G</td>
<td>hsa-miR-29b-3p</td>
<td>N/A</td>
<td>−18.12</td>
<td></td>
</tr>
<tr>
<td>CKLF</td>
<td>NM_016326</td>
<td>5</td>
<td>U72G</td>
<td>hsa-miR-29b-3p</td>
<td>N/A</td>
<td>−18.76</td>
<td></td>
</tr>
<tr>
<td>IGFBP5</td>
<td>NM_000599</td>
<td>3</td>
<td>G3898U</td>
<td>hsa-miR-29b-3p</td>
<td>N/A</td>
<td>−11.93</td>
<td>rs13403592</td>
</tr>
<tr>
<td>CALM1</td>
<td>NM_006888</td>
<td>3</td>
<td>C2472G</td>
<td>hsa-miR-29c-3p</td>
<td>N/A</td>
<td>−15.13</td>
<td></td>
</tr>
<tr>
<td>CKLF</td>
<td>NM_016326</td>
<td>5</td>
<td>U72G</td>
<td>hsa-miR-29c-3p</td>
<td>N/A</td>
<td>−18.83</td>
<td></td>
</tr>
<tr>
<td>IGFBP5</td>
<td>NM_000599</td>
<td>3</td>
<td>G3898U</td>
<td>hsa-miR-29c-3p</td>
<td>N/A</td>
<td>−12.71</td>
<td>rs13403592</td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td>hsa-miR-30a-3p</td>
<td>−12.66</td>
<td>−8.97</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td>hsa-miR-30a-3p</td>
<td>−12.66</td>
<td>−6.32</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>U2474A</td>
<td>hsa-miR-30a-3p</td>
<td>−12.66</td>
<td>−6.86</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td>hsa-miR-30e-3p</td>
<td>−12.53</td>
<td>−8.02</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td>hsa-miR-30e-3p</td>
<td>−12.53</td>
<td>−6.19</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>U2474A</td>
<td>hsa-miR-30e-3p</td>
<td>−12.53</td>
<td>−6.73</td>
<td></td>
</tr>
<tr>
<td>BCL2L13</td>
<td>NM_001270731</td>
<td>3</td>
<td>U1850G</td>
<td>hsa-miR-361-3p</td>
<td>N/A</td>
<td>−17.85</td>
<td>rs725768</td>
</tr>
<tr>
<td>SDC4</td>
<td>NM_002999</td>
<td>3</td>
<td>U1874G</td>
<td>hsa-miR-361-3p</td>
<td>−15.88</td>
<td>−20.18</td>
<td></td>
</tr>
<tr>
<td>SOX4</td>
<td>NM_003107</td>
<td>3</td>
<td>G4753A</td>
<td>hsa-miR-361-3p</td>
<td>−20.76</td>
<td>−14.48</td>
<td>rs11556729</td>
</tr>
<tr>
<td>BCL2L13</td>
<td>NM_001270731</td>
<td>3</td>
<td>U2269A</td>
<td>hsa-miR-519b-3p</td>
<td>N/A</td>
<td>−11.07</td>
<td>rs74932682</td>
</tr>
<tr>
<td>JAK1</td>
<td>NM_002227</td>
<td>3</td>
<td>U5007A</td>
<td>hsa-miR-519b-3p</td>
<td>−16.69</td>
<td>−12.31</td>
<td></td>
</tr>
<tr>
<td>OSMR</td>
<td>NM_003999</td>
<td>3</td>
<td>C4534U</td>
<td>hsa-miR-519b-3p</td>
<td>N/A</td>
<td>−13.92</td>
<td></td>
</tr>
<tr>
<td>FAM46C</td>
<td>NM_017709</td>
<td>3</td>
<td>A1459G</td>
<td>hsa-miR-614</td>
<td>−17.89</td>
<td>−22.28</td>
<td>rs2066411</td>
</tr>
<tr>
<td>KLF10</td>
<td>NM_005655</td>
<td>3</td>
<td>C2615A</td>
<td>hsa-miR-614</td>
<td>−22.30</td>
<td>−17.54</td>
<td>rs6935</td>
</tr>
<tr>
<td>RHEB</td>
<td>NM_005614</td>
<td>3</td>
<td>A1229G</td>
<td>hsa-miR-614</td>
<td>N/A</td>
<td>−15.19</td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0082699.t004
U1552G induced structural change in the SECIS element may affect the efficiency of UGA recoding. Further, considering both the set of high-confidence and medium-confidence SNVs, we found that the genes (n = 15) listed in Table 2 harbor more than one disruptive SNV in the predicted conserved structural region of mRNA. For example, the gene ID2 encodes for DNA-binding protein inhibitor ID-2, which is a critical factor for cell proliferation and differentiation in normal vertebrate development. Overexpression of the ID-2 protein is frequently observed in various human tumors, including NSCLC [57]. In the mRNA sequence of ID2 gene, two SNVs located in the 5’ UTR region were independently predicted to cause significant local structural change in the conserved region. Previous studies have shown that SNP or mutation induced structural changes in the 5’ UTRs can lead to uncontrolled translation or overexpression of the respective proteins [58,59]. We predict that the two SNVs that cause significant change in the structurally conserved region could affect the translation efficiency of ID2 mRNA.

Out of the 472 disruptive SNVs obtained from the secondary structure analysis (before intersecting with the cancer-associated genes), 199 overlap with SNPs from dbSNP (build 135). Of these 199 SNVs, 17 are in linkage disequilibrium (LD) with other SNPs that are located proximal (+/-200 nt) to the SNP position. These 17 pairs were tested with RNAsp to check whether the SNP in LD with (disruptive) SNP is a structure-stabilizing haplotype [60]. Of these 17 pairs, five were predicted to cause no significant structural changes, which could be possible structure-stabilizing haplotypes; whereas the other 12 pairs have shown significant structural changes (see File S3.xls).

Effect of SNVs on microRNA target sites

Screening all human mature miRNAs against all identified SNVs with flanking sequence yields 2,598,810,180 possible combinations. The initial TargetScan step is a conservative filter and reduces the set of SNV-miRNA pairs to 0.2% of this. We then apply miRanda as a second target prediction method, followed by a set of filters. The distribution of $l_r$ values in the alter set is shown in Figure S3, only cases with relative changes larger than the described cut off are considered (see Materials and Methods). Figure 2 shows the different steps with individual counts of putative interaction sites at each stage. This gives us 490 SNVs in 447 genes predicted to affect 707 interactions with 344 miRNAs (see File S4.xls). After intersection with known cancer-associated genes (final step in the pipeline, Figure 1), we find 124 SNVs and 148 miRNAs to be involved in 186 interactions that differ with the mutation. These SNVs that induce putative miRTS changes can be further classified into those enhancing interaction with the mutant (80) or wild-type (52) variant.

Table 3 lists all genes that contain more than one miRTS predicted to be changed between wild-type and SNV. This includes examples where the same SNP changes the target site for different mature miRNAs from the same family, but also examples where different SNVs within the gene cause a gain or loss of a miRTS. Similarly, all miRNAs with more than two changed target sites are presented in Table 4. It lists members of the miR-29 family which have previously been reported to act as tumor suppressors as well as oncogenes (see [61] for a review).

Of the 148 miRNAs (responsible for 186 putative interactions) in our final candidate set, 89 are lung cancer-associated miRNAs (in 117 interactions) (indicated in File S4.xls). Table 5 lists all 14 putative target sites in our final candidate set that include NSCLC-associated miRNAs. Notably, the list includes four miRNAs with more than one predicted target changed. For miR-184 one target site is created while another one is weakened upon introduction of the mutation. Moreover, miR-30a, d, and e are predicted to target the 3’ UTR of SUZ12 gene. However, the predicted interactions are likely to be functional in the wild-type and lost in the mutant due to SNV-induced changes at the seed region. SUZ12 has previously been shown to be directly targeted by miR-200b and inhibition of this miRNA increases the formation of cancer stem cells (CSCs) [62], which contribute to tumor aggressiveness. The loss of miR-30 regulation by (one of) the three adjacent SNVs in the seed of the target site could have a similar effect in NSCLC.

Furthermore, for 40 SNVs the predicted miRTSs were found to be located inside the local region where a significant secondary structural change was predicted by RNAsp. Of these, 15 SNVs were located in the cancer-associated genes (see Table 6). Based on the previous studies [63,64], we speculate that the SNV-induced miRTS change along with the secondary structural changes in and

Table 5. List of target predictions of NSCLC-associated miRNAs derived from the microRNA body map [45].

<table>
<thead>
<tr>
<th>Gene</th>
<th>mRNA</th>
<th>UTR</th>
<th>SNV</th>
<th>miRNA</th>
<th>$\Delta$GWT</th>
<th>$\Delta$GSNV</th>
<th>dbSNP 135</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIF4EBP2</td>
<td>NM_004096</td>
<td>3</td>
<td>C5092G</td>
<td>hsa-miR-15b-5p</td>
<td>-16.10</td>
<td>-10.65</td>
<td></td>
</tr>
<tr>
<td>EIF4EBP2</td>
<td>NM_004096</td>
<td>3</td>
<td>C5092G</td>
<td>hsa-miR-16-5p</td>
<td>-18.20</td>
<td>-13.97</td>
<td></td>
</tr>
<tr>
<td>EIF4EBP2</td>
<td>NM_004096</td>
<td>3</td>
<td>C5092G</td>
<td>hsa-miR-195-5p</td>
<td>-17.63</td>
<td>-12.23</td>
<td></td>
</tr>
<tr>
<td>KIF3B</td>
<td>NM_004798</td>
<td>3</td>
<td>G5433A</td>
<td>hsa-miR-184</td>
<td>-21.40</td>
<td>-14.40</td>
<td>rs41289846</td>
</tr>
<tr>
<td>MED16</td>
<td>NM_005481</td>
<td>5</td>
<td>A129U</td>
<td>hsa-miR-184</td>
<td>N/A</td>
<td>-20.65</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td>hsa-miR-30a-3p</td>
<td>-12.66</td>
<td>-8.97</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td>hsa-miR-30d-3p</td>
<td>-11.68</td>
<td>-8.65</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>C2473G</td>
<td>hsa-miR-30e-3p</td>
<td>-12.53</td>
<td>-8.02</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td>hsa-miR-30a-3p</td>
<td>-12.66</td>
<td>-6.32</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td>hsa-miR-30d-3p</td>
<td>-11.68</td>
<td>-5.57</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>G2475U</td>
<td>hsa-miR-30e-3p</td>
<td>-12.53</td>
<td>-6.19</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>U2474A</td>
<td>hsa-miR-30a-3p</td>
<td>-12.66</td>
<td>-6.86</td>
<td></td>
</tr>
<tr>
<td>SUZ12</td>
<td>NM_015355</td>
<td>3</td>
<td>U2474A</td>
<td>hsa-miR-30e-3p</td>
<td>-12.53</td>
<td>-6.73</td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0082699.t005
around the miRTS can potentially affect the binding of predicted miRNA.

### Functional analysis of genes predicted with SNVs’ effect on UTRs

To illustrate how the candidate SNVs obtained from our pipeline can be further analyzed for potential functionality and co-operability, we investigated the resulting sets from miRNA and RNAsnp analyses individually as well as their combination, each before and after intersection with cancer-related genes. More precisely, the following six gene sets have been tested by Ingenuity Pathways Analysis (see Table 7): 490 SNVs corresponding to 447 genes from miRNA analysis (miRNA in all genes), 124 SNVs corresponding to 104 genes that overlap with our cancer gene set (miRNA in cancer-related genes), 472 SNVs associated with 408 genes from RNAsnp analysis (RNAsnp in all genes), 111 SNVs corresponding to 89 genes that intersect with our cancer gene set (RNAsnp in cancer-related genes), a unique gene list obtained after combination of 447 genes from miRNA analysis and 408 genes from RNAsnp analysis (miRNA and RNAsnp overlap in all genes), and a unique cancer gene list obtained from 104 genes from miRNA and 89 genes from RNAsnp analysis respectively (miRNA and RNAsnp overlap in cancer-related genes).

Based on significant p-values obtained from each of our analyses, we have listed the top 3 networks, diseases and disorders, molecular and cellular functions, and canonical pathways in Table 7. Our results indicate that the top networks identified from our six gene set analyses are highly enriched with cell death and survival as well as cellular growth and proliferation (see Table 7). Figure 4 shows the network of those two cases which were predicted using the genes from the combination of miRNA and RNAsnp analyses, whereas the networks from other gene sets are shown in Figure S4. The networks shown in Figure 4 contain several genes from Table 2 and 3 which were predicted to have more than one disruptive SNV, and also genes from Table 6 for which predictions were obtained from 104 genes from miRNA and 89 genes from RNAsnp analysis respectively (miRNA and RNAsnp overlap in cancer-related genes).

Based on significant p-values obtained from each of our analyses, we have listed the top 3 networks, diseases and disorders,
Table 7. Summary of pathway analysis results using Ingenuity pathway analysis software.

<table>
<thead>
<tr>
<th>miRNA in all genes</th>
<th>miRNA in cancer related genes</th>
<th>RNAsnp in all genes</th>
<th>RNAsnp in cancer related genes</th>
<th>miRNA and RNAsnp overlap in all genes</th>
<th>miRNA and RNAsnp overlap in cancer related genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top 3 networks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top 3 diseases and disorders</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Infectious Disease (1.75E-4–4.85E-2)</td>
<td>1. Cancer (4.36E-6–4.90E-2)</td>
<td>1. Cancer (2.03E-4–4.41E-2)</td>
<td>1. Cancer (9.42E-8–1.27E-2)</td>
<td>1. Infectious Disease (1.04E-6–4.31E-2)</td>
<td>1. Cancer (8.23E-10–1.22E-2)</td>
</tr>
<tr>
<td>2. Cancer (1.42E-3–4.71E-2)</td>
<td>2. Hematological Disease (1.16E-4–4.05E-2)</td>
<td>2. Endocrine System Disorders (4.95E-4–2.38E-2)</td>
<td>2. Hematological Disease (1.46E-6–4.31E-2)</td>
<td>2. Cancer (1.93E-4–4.31E-2)</td>
<td>2. Hematological Disease (1.48E-8–1.22E-2)</td>
</tr>
<tr>
<td>3. Hepatic System Disease (1.42E-3–2.38E-2)</td>
<td>3. Endocrine System Disorders (1.36E-4–3.34E-2)</td>
<td>3. Reproductive System Disease (4.95E-4–4.08E-2)</td>
<td>3. Gastrointestinal Disease (1.08E-4–1.27E-2)</td>
<td>3. Hepatic System Disease (3.74E-7–4.31E-2)</td>
<td>3. Infectious Disease (4.22E-6–8.87E-3)</td>
</tr>
<tr>
<td><strong>Top 3 molecular and cellular functions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top 3 canonical pathways</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. EIF2 Signaling (1.61E-5)</td>
<td>1. Molecular Mechanisms of Cancer (4.14E-7)</td>
<td>1. LPS-stimulated MAPK Signaling (3.37E-5)</td>
<td>1. LPS-stimulated MAPK Signaling (8.4E-10)</td>
<td>1. iNOS Signaling (4.32E-4)</td>
<td>1. Molecular Mechanisms of Cancer (3.41E-12)</td>
</tr>
<tr>
<td>2. mTOR Signaling (4.31E-4)</td>
<td>2. Insulin Receptor Signaling (2.47E-6)</td>
<td>2. iNOS Signaling (3.36E-4)</td>
<td>2. Molecular Mechanisms of Cancer (2.65E-8)</td>
<td>2. EIF2 Signaling (4.52E-4)</td>
<td>2. Glucocorticoid Receptor Signaling (2.88E-9)</td>
</tr>
<tr>
<td>3. Insulin Receptor Signaling (9.85E-4)</td>
<td>3. IGF-1 Signaling (4.17E-6)</td>
<td>3. Germ Cell-Sertoli Cell Junction Signaling (6.75E-4)</td>
<td>3. IL-6 Signaling (5.27E-8)</td>
<td>3. mTOR Signaling (8.85E-4)</td>
<td>3. LPS-stimulated MAPK Signaling (1.9E-8)</td>
</tr>
</tbody>
</table>

The numbers at the end of each cell represent the p-values, but for the top networks it is the p-score (−log₁₀ p-value).

doi:10.1371/journal.pone.0082699.t007

Discussion

With the help of whole-genome sequencing technology, the complete genome of a cancer cell can be sequenced efficiently to identify somatic single nucleotide variants (SNVs) [65]. To date, more than 50 different cancer types and/or sub types have been sequenced [2]. The lung cancer genome was first sequenced in 2010 [66], which reports that the somatic variants were present in both coding and non-coding (UTR and other non-coding RNAs) transcribed regions, which constitute 0.6% and 0.8% respectively of the total somatic mutations identified [22,910]. In a recent study, transcriptome-wide sequencing of non-small cell lung cancer (NSCLC) type with wild-type and mutant KRAS revealed 73,717 SNVs that consisted of both germ-line and somatic variants. Of these SNVs, 29,290 were located in the UTRs of 73,717 SNVs that consisted of both germ-line and somatic variants. Of these SNVs, 29,290 were located in the UTRs of 6462 genes.

We have developed a comprehensive computational pipeline to predict the effects of SNVs located in the UTRs that can potentially affect the post-transcriptional regulation, through
SNV-induced secondary structure changes in the UTRs or changes in miRTs within UTRs. Using this pipeline, we predicted 472 out of 29,290 UTR SNVs to have significant effect on the local RNA secondary structure of UTRs (corresponding to 408 genes). Additionally, 490 out of 29,290 UTR SNVs were predicted to cause changes in a miRNA target site within the UTRs of 447 genes. Of these 490 SNVs, 124 were present in 104 genes that were previously known to be cancer-associated. For these 104 genes, 148 miRNAs were predicted to bind either in the wild-type or mutant. We found 89 out of these 148 miRNAs overlap with lung cancer-associated miRNAs, while eight miRNAs are associated specifically to NSCLC.

Taken together, all these disruptive SNVs, which were predicted to affect secondary structure or miRNA target sites, were present in 803 different genes; out of which 188 (23.4%) were previously known to be cancer-associated. Notably, this ratio is significantly higher (p-value 0.032, one-sided Fisher’s exact test) than the ratio (20.8%) of known cancer-associated genes (n = 1347) in our initial data set of 6462 genes. Similar enrichment (p-value 0.040, one-sided Fisher’s exact test) was observed when comparing the ratio of disruptive SNVs in cancer-associated genes compared to all other genes versus the ratio of the data set SNVs in cancer-associated genes compared to all other genes. However, while comparing the ratios separately on the results obtained from RNA secondary structure and miRTS analysis, we did not find any significant difference (data not shown).

Further, the IPA networks analysis (that addresses the biological relationships between genes/gene products) shows that the physical interaction of genes predicted with SNV effect might be involved in cell death and survival as well as cellular growth and proliferation. However, further analysis of these networks with respect to the topology (e.g., edge counts, neighborhood connectivity, in and out degree) is required. The functional analysis using IPA shows that the genes from our pipeline were involved in canonical pathways such as molecular mechanisms of cancer, IL-6 signaling, LPS-stimulated MAPK signaling pathways, iNOS Signaling, EIF2 signaling and mTOR signaling pathways. Given the large data set of 29,290 SNVs and the generally high false positive rate of established miRNA target prediction methods, we chose stringent filters in the miRNA analysis. The requirement of a TargetScan seed change, used to reduce the initial set of pairs, is present in 60% of our benchmark data (Table S1). The gene nodes were colored to differentiate the known (orange) and unknown (green) cancer-associated genes, and the color outside the node indicates whether the gene comes from miRNA (yellow) or RNAsnp (blue) or both.

In summary, we hypothesize that the SNVs predicted to cause significant changes in the secondary structure of UTRs or miRNA target sites within UTRs can have the potential to alter the expression of genes linked to cancer-associated pathways, and thereby contribute to the development of cancer. Although we do not provide experimental validation to support these predictions, we have highlighted the significant causative SNVs, which will be helpful for further detailed investigation. As for example, the SNV U1552G that affects the structure of the cis-acting regulatory element, selenocysteine insertion sequence (SECIS) (Figure 4), which is associated with the translational control of GPX3 mRNA. The computational pipeline presented here can be adopted for UTR SNV data from other cancer genome and transcriptome studies.

It is worth considering that the SNVs outside the protein-coding regions can have functional impacts causing altered expression of a gene. This may help identification of new cancer driver mutations.

Future directions include protein binding site predictions on both structured and unstructured parts of the UTRs. Merging with the growing amount of experimental data concerning RNA binding proteins, e.g., CLIP-seq, more general types of data than those related to miRNA targets should provide complementary information. For example, additional ranking of predicted binding site structure disruption. Further, if such data is extracted from disease tissue it should provide yet another complementary layer of data pointing to specific candidates.

Supporting Information

Figure S1 Overview of RNAsnp predictions. (PDF)

Figure S2 Distribution of GC-content in regions around (disruptive) SNVs. (PDF)

Figure S3 Histogram of lr values in miRNA analysis alter set. (PDF)

Figure S4 Top three IPA networks for the six different gene sets as described in Table 7. (PDF)

File S1 All candidates from RNAsnp analysis. Excel table listing 472 SNVs in 408 genes. (XLSX)

File S2 Medium-confidence candidates from RNAsnp analysis. Subset of File S1, lists 83 SNVs in cancer-associated genes with either d_{max} or r_{min} p-value < 0.1 (but not both < 0.05). (XLSX)

File S3 RNAsnp predicted effect of SNPs in LD with disruptive SNVs. (XLSX)
miRNAs, with indication of cancer-association genes and listing 490 SNVs in 447 genes predicted to affect target sites of 344 genes.

Table S1 Filtration steps in the miRNA pipeline tested on known examples.

References


Acknowledgments

We thank Christian Anthon for support of the computational infrastructure of this project and Larry Croft for proof-reading the manuscript.

Author Contributions

Conceived and designed the experiments: RS AW KRK JG. Performed the experiments: RS AW PN XT KRK JG. Analyzed the data: RS AW PN XT KRK JG. Contributed reagents/materials/analysis tools: JG KRK. Wrote the paper: RS AW KRK JG.


64. S1097-2765(10)00623-4 [pii];10.1016/j.molecde.2010.08.013 [doi].


