SCIENTIFIC REPORTS



SUBJECT AREAS: TRANSCRIPTION GENE REGULATION GENE EXPRESSION CELL GROWTH

Received 9 December 2011

Accepted 16 January 2012

Published 2 February 2012

Correspondence and requests for materials should be addressed to R.R.Z. (rzhang@med. wayne.edu)

* These authors contributed equally to this work.

RNA-DNA differences are rarer in proto-oncogenes than in tumor suppressor genes

Feng Gao^{1*}, Yan Lin^{1*} & Randy Ren Zhang²

¹Department of Physics, Tianjin University, Tianjin 300072, China, ²Center for Molecular Medicine and Genetics, School of Medicine, Wayne State University, Detroit 48201, USA.

It has long been assumed that DNA sequences and corresponding RNA transcripts are almost identical; a recent discovery, however, revealed widespread RNA-DNA differences (RDDs), which represent a largely unexplored aspect of human genome variation. It has been speculated that RDDs can affect disease susceptibility and manifestations; however, almost nothing is known about how RDDs are related to disease. Here, we show that RDDs are rarer in proto-oncogenes than in tumor suppressor genes; the number of RDDs in coding exons, but not in 3'UTR and 5'UTR, is significantly lower in the former than the latter, and this trend is especially pronounced in non-synonymous RDDs, i.e., those cause amino acid changes. A potential mechanism is that, unlike proto-oncogenes, the requirement of tumor suppressor genes to have both alleles affected to cause tumor 'buffers' these genes to tolerate more RDDs.

Proteins are translated from mRNAs, which are transcribed from genomic DNA. It has long been assumed that DNA sequences and corresponding RNA transcripts are almost identical; a recent discovery, however, revealed widespread differences between them¹. Cheung and coworkers sequenced and compared DNA and RNA sequences from B cells of 27 human individuals, and found more than 10,000 sites that showed RNA-DNA differences (RDD). Many of these RDDs were observed in other tissues, including primary skin cells and the brain, from other unrelated individuals. Mass spectrometry showed that sequences of many proteins corresponded to the RNA variants, rather than genomic DNA, indicating that the RNA forms were translated into proteins¹.

This breakthrough represents a largely unexplored aspect of human genome variation. Traditionally, genetic studies have been focused on DNA sequence polymorphisms, but because of the presence of RDDs, future studies will likely also need to include RNA variants. It has been speculated that RDDs can affect disease susceptibility and manifestations¹; however, almost nothing is known about how RDDs are related to disease.

Here, we have examined whether RDD distributions differ between proto-oncogenes and tumor suppressor genes. A proto-oncogene, which usually encodes proteins that regulate cell growth and differentiation, is a gene that, due to mutations, can become an oncogene to induce tumors². In contrast, a tumor suppressor gene, or anti-oncogene, which usually encodes proteins that repress cell cycle or promote apoptosis, is a gene that protects humans from tumor induction³. Consequently, we found that proto-oncogenes have significantly rarer RDDs than tumor suppressor genes, and this is especially pronounced for RDDs that lead to non-synonymous amino acid changes.

Results

To examine if RDDs have different occurrence between tumor suppressor genes and proto-oncogenes, we compared the RDD distributions between these 2 kinds of genes. Maizels and coworkers compiled a database of tumor suppressor genes and proto-oncogenes by extensively searching the Online Mendelian Inheritance in Man (OMIM) database as a primary source, followed by confirmation of gene classification based on published literatures⁴.

A gene can have RDDs in the 5' untranslated region (UTR), 3' UTR, and coding exons, and those in coding exons can be either synonymous or non-synonymous. The database contained 55 tumor suppressor genes and 95 proto-oncogenes. We calculated the number of RDDs per gene for the 2 classes of genes. Tumor suppressor genes and proto-oncogenes had 27 and 14 RDDs, respectively, in coding exons. The number of RDD per gene (0.491 vs. 0.147) was more than 3 times higher in the former than the latter (Fig. 1). We performed the chi-square tests,

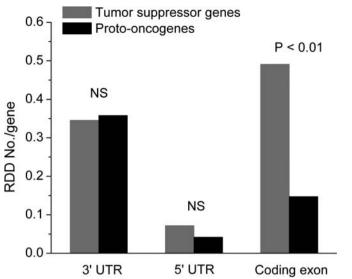


Figure 1 | RNA-DNA differences (RDDs) are rarer in proto-oncogenes than in tumor suppressor genes. The number of coding-region RDDs is significantly lower in proto-oncogenes than in tumor suppressor genes, while RDDs at 3' UTR and 5' UTR show no significant difference.

which showed that the difference was statistically significant (P < 0.01). Interestingly, the number of RDD at the 3' UTR and 5' UTR showed no significant differences (Fig. 1). Because coding regions usually have a more dominant role than 3'UTR and 5' UTR, this result seems to suggest that the RDD number difference in tumor suppressor genes and proto-oncogenes is biologically meaningful.

If RDDs are indeed related to biological functions of the genes, we would expect that the difference between the 2 kinds of genes to become more pronounced in RDDs that lead to non-synonymous amino acid changes, compared to synonymous RDDs. The numbers of non-synonymous RDDs for tumor suppressor genes and proto-oncogenes were 20 and 9, respectively (P < 0.01). The RDD number per gene in tumor suppressor genes was about 4 times of that in proto-oncogenes (0.364 vs. 0.095) (Fig. 2). Therefore, these data is consistent with the notion that RDDs are related to gene functions.

Each gene can have either one or more than one RDDs. In addition to the RDD numbers per gene, we thus also calculated the number of genes that contain RDDs, and consistent results were obtained. The number of genes that contained RDD was 36 (65.45%) and 24 (25.26%) in tumor suppressor genes and proto-oncogenes, respectively (p < 0.0001). The number of genes that contained non-synonymous RDDs in tumor suppressor genes and proto-oncogenes were 15 (27.27%) and 7 (7.37%) (P = 0.0015), and those for synonymous RDDs were 6 (10.91%) and 5 (5.26%), respectively (Fig. 3). Therefore the number of genes that contained RDD was still about 4 times in tumor suppressor genes than in proto-oncogenes.

To facilitate the research on RDDs, we created a database of RNA-DNA differences (DRDD). The database contains detailed information about RDDs, such as RDD location, involved base changes, involved amino acid changes, if any, and sequences and names of RDD-containing genes. The information is stored and operated by an open-source database management system, MySQL, which allows rapid data retrieval. Users can browse and search for RDD records, and can also Blast the query genes against the database. The database will be periodically updated to incorporate newly discovered RDDs, e.g., those in the reference⁵. DRDD can be accessed from the website: http://tubic.tju.edu.cn/drdd.

Discussion

Some mechanisms, such as transcriptional errors⁶ and RNA editing^{7,8}, are known to explain the exceptions for the complete fidelity

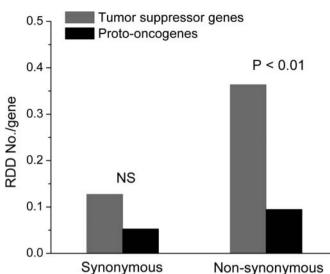


Figure 2 | Significantly lower number of non-synonymous RDDs in proto-oncogenes than in tumor suppressor genes.

from genomic DNA to mRNA. But transcription errors are very uncommon because of proofreading and repair mechanism⁹ and RNA editing mainly only involves A to G transition⁸. The RDDs identified in¹ included a large number (in the order of thousands) RDD events for all the possible 12 categories, that is, A to C/G/T, C to A/G/T, G to A/C/T and T to A/C/G. Therefore, these RDDs suggest unknown mechanisms that increase human genetic variation and diversify the human proteome because many RNA variants are translated into proteins that were identified by mass spectrometry¹.

Our results, for the first time, show that RDDs are much rarer in proto-oncogenes than in tumor suppressor genes. The database was compiled in the year 2006⁴, and therefore, the result will still need to be further validated by other studies, which hopefully include larger and more updated datasets. Nevertheless, because the difference of RDD numbers in tumor suppressor genes and proto-oncogenes is relatively large (4 times of non-synonymous RDDs in the former vs. the latter), the conclusion is not likely to change.

Our results indicate that proto-oncogenes are much intolerable to RDDs than tumor suppressor genes. One possible mechanism is that, unlike proto-oncogenes, tumor suppressor genes usually follow a two-hit hypothesis¹⁰, which suggests that both alleles that code for

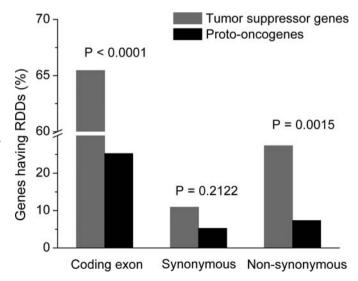


Figure 3 | Significantly lower percentage of RDD-containing genes in proto-oncogenes than in tumor suppressor genes.

| Tumor supresor 1 AFC Advertiges ENSC0000132493 deriv (122033) Tin (2 Sin (3 gene 3 BKCAI budding uninhibred by benzimidazoles I homolog ENSC000017243 deriv (38175357 AhoC Rio Sin (3407 gene 3 BKCAI budding uninhibred by benzimidazoles I homolog ENSC000017243 deriv (3817537) AhoC Hib (3 | | No | Gene Abbr. | Gene name | Ensembl ID | RDD coordinate | RDD type | AA change |
|---|---|------------------------------------|---|---|-------------------------------------|-------------------------------------|---------------------|-------------------|
| 3 BCA1 braari cancer 1 ENSG000012048 ch/17, 3842714 G to T Hoi T 4 BUB1 Budding unihibited by benzinidazoles 1 homolog FNSG000012048 ch/17, 3842714 G to T FN to T 5 CDC2 Cyclin-dependent kinnes 1 FNSG0000170312 ch/17, 3842714 G to T FN to T 6 CDKN2A Cyclin-dependent kinnes 1 FNSG0000170312 ch/16, 5217091 To G Ho G Ho G 7 DAK71 Death-associated protein kinnes 1 FNSG00001741 ch/16, 6833379 To G To G Ho G To G 7 DAK71 Death-associated protein kinnes 1 FNSG000001741 ch/16, 6833879 To G To G Ho G | Tumor suppressor | - ~ | APC BIM | Adenomatous polyposis coli Bloom svudrome | ENSG00000134982 ENSG00000197299 | chr5: 112204309 chr15: 89155557 | T to G A to C | S to A R to S |
| 4 BUB1 Budding unihibited by benzimidazoles 1 homolog FNSG0000159679 Chi 7: 3437514 G to 7 % to 6 H to 6 K to 6 H to 6 K to 6 | 2 | ၊က | BRCA 1 | breast cancer 1 | ENSG0000012048 | chr17: 38453188 | A to C | H to P |
| 4 BUBI Budding uninhibited by benzimidazoles 1 homolog ENISC0000169579 chr: hub G hub G <th< td=""><td></td><td></td><td></td><td></td><td></td><td>chr17: 38497514</td><td>GtoT</td><td>S to I</td></th<> | | | | | | chr17: 38497514 | GtoT | S to I |
| 5 CDC2 Cyclindependent kinase 1 ENSG0000170312 chr/s chr/s li11112186 Aib G NibD 7 DARVI Convoza Cyclindependent kinase 1 ENSG0000192730 chr/s 98510940 Tib G LibP 7 DARVI Death-associated protein kinase 1 ENSG000019730 chr/s 98510940 Tib G LibP 8 FANCA Fanconi anemia ENSG0000194109 chr/s 89510940 Tib G LibP 9 MAD21 MAD2 Inotati arrest deficient/like 1 ENSG0000154109 chr/s 88358799 Cto G HibO Sto P 10 MLJ3 Mybeiol/ymphoid or mixed/ineage leukemia ENSG00000104109 chr/s 88358799 Cto G HibO Sto P 11 MMAT1 Man03 Nibm ENSG00000104109 chr/s 110.6 Nib 110.6 Nib 110.6 Nib 11 MMAT1 Man03 Nibm ENSG00000104109 chr/s 110.5 110.7 110.6 11 | | 4 | BUB1 | Budding uninhibited by benzimidazoles 1 homolog | ENSG00000169679 | chr2: 111140447 | A to G | H to R |
| 5 CDC2 Cyclindependent kinase i 1 ENSG000017312 chr): 6217991 Tio A Hie G Tie A 7 DARD CKN2A Cyclindependent kinase i 1 ENSG000013739 chr): 8945638 Tie A 7 DARD Fancori anemia ENSG0000137741 chr): 8945638 Tie G Tie A 8 FANCA Fancori anemia ENSG0000187741 chr): 88436379 Tie G Tie A 9 MAD211 MAD211 MAD211 chr): 88436379 Tie G Tie A 11 MM171 MyaD211 ENSG0000164109 chr): 88335797 Tie G Tie G 11 MM171 MaD2 MaD2 ENSG0000164109 chr): 88335797 Tie G Tie G 11 MM171 MaD2 MaD2 ENSG0000104320 chr): 151426637 Tie G Tie G Tie G 11 MN471 MaD3 MaD3 ENSG00000104320 chr): 151206413 Tie G Uo 11 MN471 MaD3 Tie G Tie G Uo <td></td> <td></td> <td></td> <td></td> <td></td> <td>chr2: 111112186</td> <td>A to G</td> <td>N to D</td> | | | | | | chr2: 111112186 | A to G | N to D |
| 6 CDKN2A Cyclin dependent kinase 1 ENSG000014/289 chrs. 85519940 Tip G Tip Mor 7 DAPK1 Dearth-associated protein kinase 1 ENSG000019/230 chrs. 85519940 Tip G Tip Mor 8 FANCA Fanconi anemia ENSG000019/230 chrs. 85519340 Tip G Tip Mor 9 MAD211 MAD22 milotic arrest deficient-like 1 ENSG000015569 chrs. 88313879 Cip G Tip Mor 10 ML12 MAD21 MAD22 milotic arrest deficient-like 1 ENSG0000015569 chrs. 18121206413 A to G Tip V 11 MMAT1 NMAT1 Menege a tots homolog 1 ENSG0000013420 chrs. 18121206413 A to G Tip V 12 NBN Nith Nith Mor Nith S | | 5 | CDC2 | Cyclin-dependent kinase 1 | ENSG0000170312 | chr10: 62217991 | T to A | H to Q |
| 7 DAPK1 Dearth-associated protein kinase 1 ENSG0000195730 chr.s. 89510940 Tio.C Lib P 8 FANCA Fanceni anemia ENSG0000187741 chr.s. 8946353 Tio.C Lib P 9 MAD21 MAD2 mitoric arrest deficient-like 1 ENSG000014109 chr.s. 112.064113 Tio.C Lib P 10 ML13 Myeloid/ymphoid or mixed-lineage leukemia 3 ENSG00000144109 chr.s. 112.06413 Tio.C Work 11 MWAT1 Memoge a trois homolog 1 ENSG00000144109 chr.s. 16840454 Tio.C Work 12 NBN NMAT1 Memoge a trois homolog 1 ENSG00000144109 chr.s. 1631476867 Tio.C Work 13 MXTR Natural killentron ENSG00000144057 chr.s. 16344826 Tio.C Work 13 MXTR Natural killentron ENSG00000144057 chr.s. 16354313 Tio.C Work 13 MVTR Natural killentron ENSG00000144857 chr.s. 173543313 Tio.C Work 14 ABLI Colo morogenei< | | 9 | CDKN2A | Cyclin-dependent kinase inhibitor 2A | ENSG00000147889 | chr9: 21960979 | T to G | S to A |
| B FANCA Fanconi anemia ENSG0000187741 chrl: 6.88336379 TheG TheA P MAD2L1 MAD2 milotic arrest deficient-like 1 ENSG0000164109 chrl: 6.8833679 TheC TheC The No 10 MU13 MyaD1 Meaoge a trois homolog 1 ENSG00000144109 chrl: 11206413 TheC The C Wues 11 MVAT Meaoge a trois homolog 1 ENSG0000014320 chrl: 4.6034826 The C Wues 12 NBN Nibrin Nibrin ENSG0000014320 chrl: 4.6034826 The C Wues 13 NKTR Nutrol killertumor recognition sequence ENSG0000014320 chrl: 4.5034826 The C Wue 13 NKTR Nutrol killertumor recognition sequence ENSG0000014320 chrl: 4.5749053 The C Whe 13 NKTR Nutrol killertumor recognition factor ENSG00000155699 chrl: 4.57490533 The C Uh E 14 REST REL The C MNe ENSG00000155699 chrl: 1.72402122 The C Uh E 15 TSCI The REST The C MNe | | ~ | DAPK1 | Death-associated protein kinase 1 | ENSG00000196730 | chr9: 89510940 | T to C | L to P |
| B FANCA Fanconi anemia ENSG0000187741 chrl 6: 88353799 Cto G Hth 0 9 MAD211 MAD2 mitotic arrest deficient-like 1 ENSG0000164109 chrl 6: 88353799 Tto C Sto P 10 ML3 Mveloid/fymphold or mixed/ineage leukemia 3 ENSG0000055609 chrl 3: 121206413 Ato G Dto G 11 MNAT Nenrge a frois homolog 1 ENSG00000104322 chrl 3: 91027572 Ato G Dto G 12 MS Nibrin EST NetTR Natural killer/unor recognition sequence ENSG000010432 chrl 3: 5336164 Tto C Wue 3 13 NKTR Natural killer/unor recognition sequence ENSG0000014057 chrl 3: 5336164 Tto C Wue 3 14 REST Unberous sclerosis 1 ENSG00000155699 chrl 3: 5732013 Gto T Mue 1 15 TSC1 Uuberous sclerosis 1 ENSG00000155699 chrl 3: 57409253 Gto T Mue 1 16 Tooloons ENSG00000155699 chrl 3: 57409253 Gto T Mue 1 ENSG000000155699 <td< td=""><td></td><td></td><td></td><td></td><td></td><td>chr9: 89486358</td><td>A to G</td><td>T to A</td></td<> | | | | | | chr9: 89486358 | A to G | T to A |
| 9 MAD211 MAD2 milotic arrest deficient/like 1 ENSC00000164109 chr.16: 88404654 ThCC S ho P 10 ML3 MMD21 MAD21 MAD21 MAD21 MAD21 ThCC S ho P 11 ML3 Menage a froit homolog 1 ENSC00000164109 chr.3 51476657 ThC C Wo B 12 NBN Nibrin Menage a froit homolog 1 ENSC00000104320 chr.3 40.6 Un G 13 NKTR Natural killer4umor recognition sequence ENSC00000104320 chr.3 42635164 Tho G Un G 13 NKTR Natural killer4umor recognition sequence ENSC00000104093 chr.3 42636164 Tho G Un G 14 REST TbL Cabl moregane Th SSC00000035699 chr.3 470533 Gh T Che G Un G 15 FLI Cabl moregane Th SSC000000135699 chr.3 470533 Gh T Che G Un G 16 C Cabl moregane ENSC000000135699 chr.3 170.62 Un G Un B Un SSC000001565699 chr.3 170.52 | | 8 | FANCA | Fanconi anemia | ENSG00000187741 | chr16: 88358799 | CtoG | НюQ |
| 9 MAD2L1 MAD2 Mucr | | | | | | chr16: 88404654 | T to C | S to P |
| 10 ML3 Myeloid/ymphoid or mixed/ineage leukemia 3 ENSG0000055609 ch7: 151476867 ToC W or No 11 MMAT1 Menage a trois homolog 1 ENSG0000020426 chr14: 6034822 A in G D in G 12 NBN Nibrin ENSG0000104320 chr14: 6034825 T in G V in G 13 NKTR Natural killertumar recognition sequence ENSG0000014857 chr3: 42536164 T in G V in G 14 REST RE1 siloneing transcription factor ENSG0000015699 chr12: 119766013 G in T M in G 15 TSCI Uberouncogene T in ABL1 ENSG0000015699 chr12: 119765053 G in T M in C 16 FI/I Friend leukemia virus integration 1 ENSG0000015599 chr12: 11976868 A in G H in C 2 ETV6 ETS framslocation virus integration 1 ENSG00000151702 chr11: 128180503 T in C U in P 3 JU/IB Un B proto-oncogene ENSG00000151202 chr11: 128180503 T in C U in P 6 M | | 6 | MAD2L1 | MAD2 mitotic arrest deficient-like 1 | ENSG00000164109 | chr4: 121206413 | A to G | I to V |
| 11 MNAT1 Menage a frois homolog 1 ENSG0000020426 chrl 4: 60344826 A to G D to G 12 NBN Nibrin ENSG0000104320 chrl 4: 50341826 A to G D to G 13 NKTR Natural killer-tumor recognition sequence ENSG00000104320 chrl 3: 42336164 T to A N to G 14 REST RE1-silencing transcription factor ENSG00000155699 chrl 3: 42336164 T to A N to B 15 TSC1 Tuberous sclerosis 1 ENSG00000155699 chrl 3: 7490953 G to T C to F 15 TSC1 Tuberous sclerosis 1 ENSG00000156699 chrl 3: 123409533 G to T C to F 15 TSC1 Tuberous sclerosis 1 ENSG00000156699 chrl 3: 11928688 A to C D to A 2 MAP3K8 Minogenorativated protein kinase kinasekinase 8 ENSG00000171223 chrl 11: 128180503 T to A V to N 3 FU1 Friend leukemia viral oncogene homolog ENSG00000171223 chrl 19: 12765027 A to G D to A 4 JUNB V-mybmyeloblastosis viral oncogene homolog ENSG00000177223 chrl 19: 12765027 </td <td></td> <td>10</td> <td>MLL3</td> <td>Myeloid/lymphoid or mixed-lineage leukemia 3</td> <td>ENSG00000055609</td> <td>chr7: 151476867</td> <td>T to C</td> <td>W to R</td> | | 10 | MLL3 | Myeloid/lymphoid or mixed-lineage leukemia 3 | ENSG00000055609 | chr7: 151476867 | T to C | W to R |
| 12 NBN Nibrin ENSG0000104320 chr8: 91027572 A bG Q b R 13 NKTR Ndurd killer-tumor recognition sequence ENSG0000114857 chr3: 42636164 ToG V bG 13 NKTR Ndurd killer-tumor recognition sequence ENSG0000084093 chr3: 42636164 ToG V bG 14 REST TSCT Tuberous sclerosis 1 ENSG0000084093 chr3: 4263613 G bT M bG 15 TSCT Tuberous sclerosis 1 ENSG00000155699 chr3: 4263613 G bT M bF 15 TSCT Tuberous sclerosis 1 ENSG00000155699 chr9: 132740222 ToC U P 2 ETV6 ETS translocation variant 6 ENSG00000151702 chr11: 128180503 ToC U P 3 FU1 Friend leukemia virus integration 1 ENSG00000171223 chr10: 30776876 ToC U P 3 FU1 Friend leukemia virus integration 1 ENSG00000171223 chr10: 30776876 ToC U P 4 UNB VmBpmyeloblastosis viral oncogene homolog <td></td> <td>[]</td> <td>MNATI</td> <td>Menage a trois homolog 1</td> <td>ENSG00000020426</td> <td>chr14: 60344826</td> <td>A to G</td> <td>D to G</td> | | [] | MNATI | Menage a trois homolog 1 | ENSG00000020426 | chr14: 60344826 | A to G | D to G |
| 13 NKTR Natural killer-tumor recognition sequence ENSG00000114857 chr8: 91045867 Tpo G V ho G 13 REST RE1-silencing transcription factor ENSG00000148599 chr3: 42636164 Tpo A N ho K 15 TSC1 Tuberous scleng transcription factor ENSG0000015699 chr3: 42636103 G ho T C ho F 15 TSC1 Tuberous scleng transcription factor ENSG0000015699 chr2: 132740222 T ho A N ho L 2 ETV6 ETV6 ETV6 ENSG00000151702 chr12: 11928688 A ho C D ho A 3 FUI Friend leukemia virus integration 1 ENSG00000171223 chr11: 128180503 T ho C U ho R 6 MYB V-mybmyeloblastosis viral oncogene 1 ENSG0000017228 chr11: 128180503 T ho C U ho R 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000172028 chr13: 11723868 T ho C U ho R 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000172028 chr19: 307587262 A ho G U ho R 7 RALA V-mybmyeloblastosis viral oncogene homolog < | | 12 | NBN | Nibrin | ENSG00000104320 | chr8: 91027572 | A to G | QtoR |
| 13 NKTR Natural killer-tumor recognition sequence ENSG00000114857 chr3: 42636164 T to A N to K 14 REST RE1-silencing transcription factor ENSG0000084093 chr3: 42636164 T to A N to K 15 TSC1 Tuberous selerosis 1 ENSG0000084093 chr3: 134766013 G to T C to F 15 TSC1 Tuberous selerosis 1 ENSG00000155699 chr3: 11928688 A to C L to P 2 ETV6 ETV6 ETV6 ENSG0000015702 chr11: 128180503 T to C L to P 3 FU1 Friend levelemic virus integration 1 ENSG00000171223 chr10: 30776876 T to C V to N 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000171223 chr10: 30776876 T to A C to S 7 RALA V-mybmyeloblastosis viral oncogene homolog ENSG00000171223 chr10: 30776876 T to A C to S 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000171233 chr10: 30776876 T to A C to S 7 RALA V-mybmyeloblastosis viral oncogene homolog ENSG00000177233 | | | | | | chr8: 91045867 | T to G | V to G |
| 14 REST RE1-silencing transcription factor ENSG0000084093 chr4: 5749053 G to T C to F 15 TSC1 Tuberous sclerosis 1 ENSG0000165699 chr9: 132746013 G to T C to P 15 TSC1 Tuberous sclerosis 1 ENSG00000156699 chr9: 1327462013 G to T C to P 2 ETV6 ETS translocation variant 6 ENSG00000139083 chr11: 128180503 T to C U to P 3 FUI Final elvemia virus integration 1 ENSG00000112223 chr19: 12265027 A to C D to A 5 MAP3K8 Mitogen-activated protein kinase kinase kinase 8 ENSG00000117223 chr10: 30776876 T to A Y to N 6 MYB V-roybmyeloblastosis viral oncogene homolog ENSG0000016451 chr5: 33557262 A to G D to A 7 RALA V-rol simian leukemia viral oncogene homolog ENSG0000016451 chr5: 33567625 A to G D to A 7 RALA V-rol simian leukemia viral oncogene homolog ENSG0000016451 chr5: 33567625 A to G D to A 7 RALA V-rol simian leukemia viral oncogene homolog A ENSG00 | | 13 | NKTR | Natural killer-tumor recognition sequence | ENSG00000114857 | chr3: 42636164 | T to A | N to K |
| 15 TSC1 Tuberous sclerosis 1 ENSG0000165699 chr9: 134766013 G to T M to I Proto-oncogene 1 ABL1 Cabl oncogene 1 ENSG0000097007 chr9: 132740222 T to C L to P 2 ETV6 ETS translocation variant 6 ENSG0000139083 chr12: 11928688 A to C U to P 3 FU1 Friend leukemia virus integration 1 ENSG00000151702 chr11: 128180503 T to C W to R 4 JUNB Jun B proto-oncogene 1 ENSG0000171223 chr11: 128180503 T to C W to R 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000171223 chr10: 30776876 T to C V to S 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG0000018513 chr10: 30776876 T to A C to S 7 RALA V-ral simian leukemia viral oncogene homolog ENSG000000451 chr7: 39696651 A to G EV6 Et oG 7 RALA V-ral simian leukemia viral oncogene homolog ENSG000000451 chr7: 39702910 C to S A to G Et oG 7 RALA V-ral simian leukemia viral | | 14 | REST | RE1-silencing transcription factor | ENSG0000084093 | chr4: 57490953 | GtoT | ChoF |
| Proto-oncogene 1 ABL1 C-abl oncogene 1 ENSG000097007 chr9: 132740222 T to C U to P 2 ETV6 ETS translocation variant 6 ENSG0000151702 chr11: 128180503 T to C W to R 3 FUI1 Friend leukemia virus integration 1 ENSG0000151702 chr11: 128180503 T to C W to R 4 UNB Jun B proto-oncogene1 ENSG0000171223 chr19: 12265027 A to G H to R 5 MAP3K8 Mitogen-activated protein kinase kinasekinase 8 ENSG000017968 chr10: 30776876 T to A V to N 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG0000018513 chr10: 30776876 T to A C to S 7 RALA V-ral simian leukemia viral oncogene homolog ENSG0000018513 chr10: 30776876 T to A C to S 7 RALA V-ral simian leukemia viral oncogene homolog ENSG00000006451 chr7: 395096510 C to S C to S 7 RALA Varia Simian leukemia viral oncogene homolog ENSG00000018513 chr7: 39509510 C to S C to S 7 RALA Varia Simian leukemia viral oncoge | | 15 | TSCI | Tuberous sclerosis 1 | ENSG00000165699 | chr9: 134766013 | GtoT | M to I |
| 2 ETV6 ETS translocation variant 6 ENSG0000139083 chr12: 11928688 A to C D to A 3 FU1 Friend leukemia virus integration 1 ENSG0000151702 chr11: 128180503 T to C W to R 4 JUNB Jun B proto-oncogene1 ENSG0000171223 chr19: 12765027 A to G W to R 5 MAP3K8 Mitogen-activated protein kinase kinasekinase 8 ENSG0000171223 chr10: 307787136 T to A Y to N 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG0000018513 chr10: 30776876 T to A Y to N 7 RALA V-ral simian leukemia viral oncogene homolog ENSG0000016451 chr7: 39569651 A to G N to D 7 RALA V-ral simian leukemia viral oncogene homolog ENSG0000006451 chr7: 39569651 A to G E to G 0chtrit tumor supressorgene incluée AIM AIR BRCA2 CAPR, CDR1, CDR4, CDR1, MLH1, MLH3, MSH2, MSH6, NF1, NF2, PMS1, PMS4, SMAD4, SMARCA3, SM | Proto-oncogene | - | ABL1 | C-abl oncogene 1 | ENSG00000097007 | chr9: 132740222 | T to C | L to P |
| 3 FU1 Friend leukemia virus integration 1 ENSG0000151702 chr11: 128180503 T to C W to R 4 JUNB Jun B proto-oncogene1 ENSG0000171223 chr19: 12765027 A to G H to R 5 MAP3K8 Mitogen-activated protein kinase kinasekinase 8 ENSG0000171223 chr10: 30787136 T to A Y to N 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000118513 chr10: 30758766 T to A C to S 7 RALA V-ral simian leukemia viral oncogene homolog ENSG0000006451 chr7: 395696551 A to G N to D 7 RALA V-ral simian leukemia viral oncogene homolog ENSG00000006451 chr7: 39702910 C to A A to G | • | 2 | ETV6 | ETS translocation variant 6 | ENSG00000139083 | chr12: 11928688 | A to C | D to A |
| 4 JUNB Jun B proto-oncogene 1 ENSG0000171223 chr19: 12765027 A to G H to R 5 MAP3K8 Mitogen-activated protein kinase kinasekinase 8 ENSG0000107968 chr10: 30787136 T to A Y to N 6 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000118513 chr10: 30776876 T to A C to S 7 RALA V-ral simian leukemia viral oncogene homolog ENSG0000006451 chr5: 33696651 A to G N to D 7 RALA V-ral simian leukemia viral oncogene homolog A ENSG0000006451 chr7: 33696651 A to G E to G 20her humor supressorgenes include AIM, ATR, BRCA2, CAPR, CDR1, CDR4, CDRV1A, CDRV2B, CHER2, DCC, FBXW7, FHT, MCC, MDC1, MLH1, MLH3, MSH2, MSH6, NF1, NF2, PMS1, PRM5, RM6, SMAD4, SMARCA3, SMARCB1, TA A to G E to G | | ო | FUT | Friend leukemia virus integration 1 | ENSG00000151702 | chr11: 128180503 | T to C | W to R |
| 5 MAP3K8 Mitogen-activated protein kinase kinasekinase 8 ENSG0000107968 chr10: 30787136 T to A Y to N cho S MYB V-mybmyeloblastosis viral oncogene homolog ENSG0000118513 chr6: 135557262 A to G N to D 7 RALA V-ral simian leukemia viral oncogene homolog A ENSG0000006451 chr7: 39696651 A to G E to G chr7: 39696651 A to G E to G chr7: 30702910 C to A A to E to G chr2: two chrost cases include ATM, ATR BRCA2, CASP3, CD81, CD84, CDKN1A, CDKN2B, CHER2, DCC, FBXW7, FHT, MCC, MDC1, MLH3, MSH2, MSH6, NF1, NF2, PMS1, RBM5, RBM5, SMAD4, SMARCA3, SMARCB1, T | | 4 | JUNB | Jun B proto-oncogene l | ENSG00000171223 | chr19: 12765027 | A to G | H to R |
| MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000118513 MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000118513 Chró: 135557262 A to G N to D Chró: 39696651 A to G E to G Chró: 39702910 C to A A to E Cohron suppressor genes include ATM, ATR, BRCA2, CAPH, CDK4, CDKN1A, CDKN2B, CHEK2, DCC, FBXW7, FHT, MCC, MDC1, MLH1, MLH3, MSH2, MSH6, NF1, NF2, PMS1, RBM5, RBM5, SMAD4, SMARCA3, SMARCB1, 71 | | 5 | MAP3K8 | Mitogen-activated protein kinase kinasekinase 8 | ENSG00000107968 | chr10: 30787136 | T to A | Y to N |
| MYB V-mybmyeloblastosis viral oncogene homolog ENSG00000118513 chr6: 135557262 A to G N to D RALA V-ral simian leukemia viral oncogene homolog A ENSG0000006451 chr7: 39696651 A to G E to G RALA V-ral simian leukemia viral oncogene homolog A ENSG0000006451 chr7: 39702910 C to A A to G Cho A A to E Chr3: 39702910 C to A A to E | | | | | | chr10: 30776876 | T to A | C to S |
| 7 RALA V-ral simian leukemia viral oncogene homolog A ENSG0000006451 chr7: 39696651 A to G E to G chr7: 39702910 C to A A to E ^{ander humor suppressor genes include ATM, ATR, BRCA2, CASP3, CD82, CDH1, CDK4, CDKN28, CHEK2, DCC, FBXW7, FHT, MCC, MDC1, MLH3, MSH2, MSH2, NF1, PMS2, PTCH, PTEN, RBN5, SMAD4, SMARCA3, SMARCB1, TU and the transference and the ATM, ATR, BRCA2, CASP3, CD82, CDH1, CDK4, CDKN1A, CDKN28, CHEK2, DCC, FBXW7, FHT, MCC, MDC1, MLH3, MSH2, MSH2, NF1, PMS2, PTCH, PTEN, RBN5, SMAD4, SMARCA3, SMARCB1, TU and the transference and the ATM, ATR, BRCA2, CASP3, CD82, CDH1, CDK4, CDKN1A, CDKN28, CHEK2, DCC, FBXW7, FHT, MCC, MDC1, MLH3, MSH2, MSH2, NF1, PMS2, PTCH, PTEN, RB1, RBN5, SMAD4, SMARCA3, SMARCB1, TU and the transference and tran} | | 9 | MYB | V-mybmyeloblastosis viral oncogene homolog | ENSG00000118513 | chró: 135557262 | A to G | N to D |
| Chr.7: 39702910 C to A A to E Chr. 2010 C to A A to E Control of the control of t | | ~ | RALA | V-ral simian leukemia viral oncogene homolog A | ENSG0000006451 | chr7: 39696651 | A to G | EtoG |
| Other timor suppressor genes include ATM, ATR, BRCA2, CAPR3, CDR4, CDKN1A, CDKN2B, CHEK2, DCC, FBXW7, FHIT, MCC, MDC1, MLH3, MSH2, MSH2, MSH2, PMS1, PMS2, PTCH, PTEN, RB1, RBM5, SMAD4, SMARCB1, TC PURCE TIMOR TECS T | | | | | | chr7: 39702910 | C to A | A to E |
| | ^o Other tumor suppressor gen THBS1, TIMP3, TP53, TP53Bf | es include ATM, 22, TREX1, TUSC | ATR, BRCA2, CASP3, CL 3, VHL, WRN and W71. | 382, CDH1 , CDK4, CDKN1A, CDKN2B, CHEK2, DCC, FBXW7, FHIT, MCC, MDC | CI, MLH1, MLH3, MSH2, MSH6, NF1, NF | 2, PMS1, PMS2, PTCH, PTEN, RB1, RBM | 15, SMAD4, SMARCA3, | , SMARCB1, TGFBR2 |

a particular gene must be affected to cause tumor. If only one allele is affected, the other copy of the gene can still function to protect the cell. In contrast, for oncogenes, mutations in one allele can lead to tumor. Therefore, the requirement for affecting both alleles in tumor suppressor genes seems to have a 'buffer' effect for these genes to tolerate more RDDs, compared to proto-oncogenes.

In summary, RDDs, a newly discovered phenomenon, represent a largely unexplored area of human genome variation. Although it has been speculated that RDDs are involved in disease susceptibility and manifestations, no evidence is found to relate RDDs to disease. We here show that RDDs are rarer in proto-oncogenes than in tumor suppressor genes; the number of RDDs in coding exons, but not in 3'UTR and 5'UTR, is significantly lower in the former than the latter, and this trend is especially pronounced in RDDs that cause nonsynonymous amino acid changes. This result suggests that protooncogenes are more intolerable to RDDs than tumor suppressor genes. A potential mechanism is that, unlike proto-oncogenes, the requirement of tumor suppressor genes to have both allele affected to cause tumor 'buffers' these genes to tolerate more RDDs.

During proofreading, we noticed a recent publication¹¹ which suggested that rather than RNA editing events, these RDDs can be the result of accurate transcription from paralogous genes, making the issue of wide-spread human RDDs highly controversial. Therefore, the prevalence of human RDDs reported in the reference¹ needs to be further confirmed by more studies in more tissues and with more disease conditions. The observation made by Schrider et al.¹¹ appears to explain the majority of the RDDs observed in the reference¹. In that case, an alternative explanation for the RDD difference between proto-oncogenes and tumor suppressor genes is that rather than RNA editing mechanisms, it may in fact reflect the different distribution of paralogous genes between the 2 gene classes. This possibility, however, needs to be addressed by future studies.

Methods

The 10,210 RDDs, which reside in 4,741 known genes in the human genome, were provided in reference¹. A database of tumor suppressor genes and proto-oncogenes was used to compare RDD difference between the 2 classes of genes. The database, which was based on extensively searching the Online Mendelian Inheritance in Man (OMIM) database⁴, contained 55 tumor suppressor genes and 95 proto-oncogenes. HGNC symbols, which have been assigned by the HUGO Gene Nomenclature Committee (HGNC) as unique gene symbols and names, were used to link the 4,741 known genes containing RDDs with the tumor suppressor genes and

proto-oncogenes. Based on the above information, a program was written in the language of C++ to search for the tumor suppressor genes and proto-oncogenes containing RDDs. Detailed information about the tumor suppressor genes and proto-oncogenes containing non-synonymous RDDs is shown in Table 1. Either chi-square or Fisher exact tests were used to compare the number of RDDs between the 2 classes of genes, and P values less than 0.01 were considered statistically significant.

- 1. Li, M. et al. Widespread RNA and DNA sequence differences in the human transcriptome. *Science* **333**, 53–8 (2011).
- 2. Croce, C. M. Oncogenes and cancer. N Engl J Med 358, 502-11 (2008).
- 3. Bishop, J. M. The molecular genetics of cancer. Science 235, 305-11 (1987).
- 4. Eddy, J. & Maizels, N. Gene function correlates with potential for G4 DNA
- formation in the human genome. *Nucleic Acids Res* 34, 3887–96 (2006).
 Ju, Y. S. *et al.* Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* 43, 745–52 (2011).
- Sydow, J. F. & Cramer, P. RNA polymerase fidelity and transcriptional proofreading. *Curr Opin Struct Biol* 19, 732–9 (2009).
- Powell, L. M. et al. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell 50, 831-40 (1987).
- Chen, S. H. *et al.* Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 238, 363–6 (1987).
- Thomas, M. J., Platas, A. A. & Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* 93, 627–37 (1998).
- Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68, 820–3 (1971).
- 11. Schrider, D. R., Gout, J. F. & Hahn, M. W. Very few RNA and DNA sequence differences in the human transcriptome. *PLoS One* **6**, e25842 (2011).

Acknowledgements

The present work was supported in part by a fund (176412) from Wayne State University to RRZ, and by the National Natural Science Foundation of China (Grant Nos. 90408028, 30800642 and 10747150). Funding for open access charge: NNSF 31171238.

Author contributions

Conceived and designed the experiments: RRZ. Performed the experiments: FG and YL. Analyzed the data: RRZ, FG and YL. Wrote the paper: RRZ.

Additional information

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

How to cite this article: Gao, F., Lin Y & Zhang R.R. RNA-DNA differences are rarer in proto-oncogenes than in tumor suppressor genes. *Sci. Rep.* **2**, 245; DOI:10.1038/srep00245 (2012).