SCIENTIFIC REPORTS

OPEN

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 Zhang[2]

[1]Department of Physics, Tianjin University, Tianjin 300072, China, [2]Center 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 them[1]. 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 proteins[1].

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 manifestations[1]; 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 tumors[2]. 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 induction[3]. 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 literatures[4].

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,

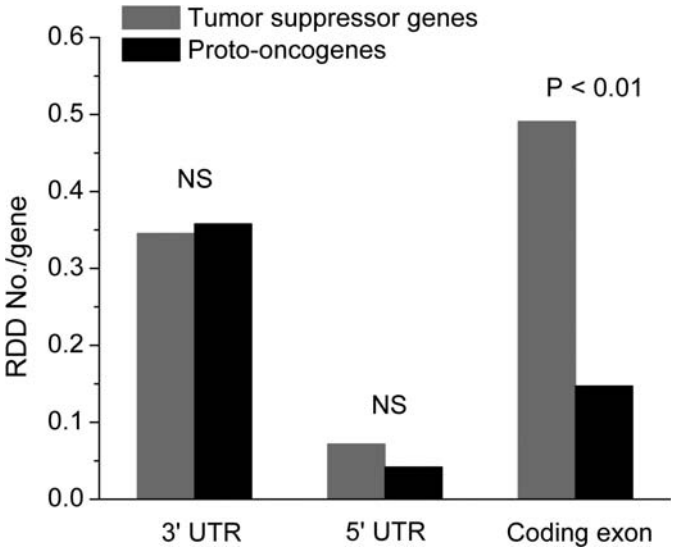

Figure 1 | RNA-DNA differences (RDDs) are rarer in proto-oncogenes than in tumor suppressor genes. The number of coding-region RDDs is significantly lower in proto-oncogenes than in tumor suppressor genes, while RDDs at 3′ UTR and 5′ UTR show no significant difference.

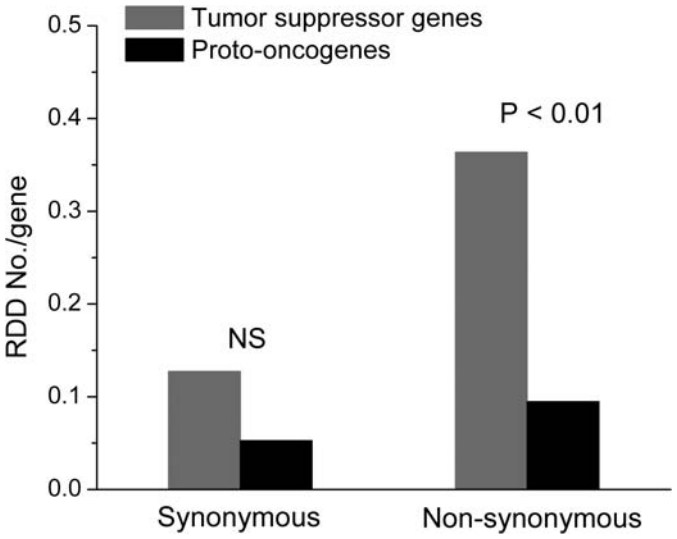

Figure 2 | Significantly lower number of non-synonymous RDDs in proto-oncogenes than in tumor suppressor genes.

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 reference[5]. DRDD can be accessed from the website: http://tubic.tju.edu.cn/drdd.

Discussion

Some mechanisms, such as transcriptional errors[6] and RNA editing[7,8], are known to explain the exceptions for the complete fidelity from genomic DNA to mRNA. But transcription errors are very uncommon because of proofreading and repair mechanism[9] and RNA editing mainly only involves A to G transition[8]. The RDDs identified in[1] 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 spectrometry[1].

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 2006[4], 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 hypothesis[10], which suggests that both alleles that code for

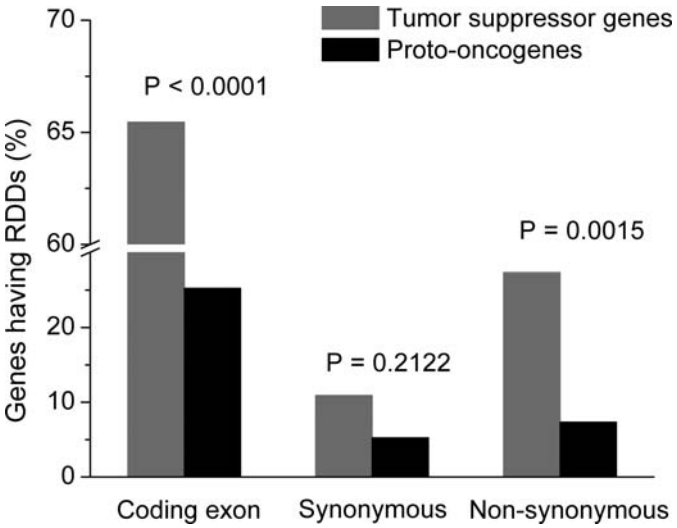

Figure 3 | Significantly lower percentage of RDD-containing genes in proto-oncogenes than in tumor suppressor genes.

Table 1 | Non-synonymous RNA-DNA differences in tumor suppressor genes and proto-oncogenes [a,b]

	No.	Gene Abbr.	Gene name	Ensembl ID	RDD coordinate	RDD type	AA change
Tumor suppressor gene	1	*APC*	Adenomatous polyposis coli	ENSG00000134982	chr5: 112204309	T to G	S to A
	2	*BLM*	Bloom syndrome	ENSG00000197299	chr15: 89155557	A to C	R to S
	3	*BRCA1*	breast cancer 1	ENSG00000012048	chr17: 38453188	A to C	H to P
					chr17: 38497514	G to T	S to I
	4	*BUB1*	Budding uninhibited by benzimidazoles 1 homolog	ENSG00000169679	chr2: 111140447	A to G	H to R
					chr2: 111112186	A to G	N to D
	5	*CDC2*	Cyclin-dependent kinase 1	ENSG00000170312	chr10: 62217991	T to A	H to Q
	6	*CDKN2A*	Cyclin-dependent kinase inhibitor 2A	ENSG00000147889	chr9: 21960979	T to G	S to A
	7	*DAPK1*	Death-associated protein kinase 1	ENSG00000196730	chr9: 89510940	T to C	L to P
					chr9: 89486358	A to G	T to A
	8	*FANCA*	Fanconi anemia	ENSG00000187741	chr16: 88358799	C to G	H to Q
					chr16: 88404654	T to C	S to P
	9	*MAD2L1*	MAD2 mitotic arrest deficient-like 1	ENSG00000164109	chr4: 121206413	A to G	I to V
	10	*MLL3*	Myeloid/lymphoid or mixed-lineage leukemia 3	ENSG00000055609	chr7: 151476867	T to C	W to R
	11	*MNAT1*	Menage a trois homolog 1	ENSG00000020426	chr14: 60344826	A to G	D to G
	12	*NBN*	Nibrin	ENSG00000104320	chr8: 91027572	A to G	Q to R
					chr8: 91045867	T to G	V to G
	13	*NKTR*	Natural killer-tumor recognition sequence	ENSG00000114857	chr3: 42636164	T to A	N to K
	14	*REST*	RE1-silencing transcription factor	ENSG00000084093	chr4: 57490953	G to T	C to F
	15	*TSC1*	Tuberous sclerosis 1	ENSG00000165699	chr9: 134766013	G to T	M to I
Proto-oncogene	1	*ABL1*	C-abl oncogene 1	ENSG00000097007	chr9: 132740222	T to C	L to P
	2	*ETV6*	ETS translocation variant 6	ENSG00000139083	chr12: 11928688	A to C	D to A
	3	*FLI1*	Friend leukemia virus integration 1	ENSG00000151702	chr11: 128180503	T to C	W to R
	4	*JUNB*	Jun B proto-oncogene1	ENSG00000171223	chr19: 12765027	A to G	H to R
	5	*MAP3K8*	Mitogen-activated protein kinase kinasekinase 8	ENSG00000107968	chr10: 30787136	T to A	Y to N
					chr10: 30776876	T to A	C to S
	6	*MYB*	V-mybmyeloblastosis viral oncogene homolog	ENSG00000118513	chr6: 135557262	A to G	N to D
	7	*RALA*	V-ral simian leukemia viral oncogene homolog A	ENSG00000006451	chr7: 39696651	A to G	E to G
					chr7: 39702910	C to A	A to E

[a]Other tumor suppressor genes include *ATM, ATR, BRCA2, CASP3, CD82, CDH1, CDK4, CDKN1A, CDKN2B, CHEK2, DCC, FBXW7, FHIT, MCC, MDC1, MLH1, MLH3, MSH2, MSH6, NF1, NF2, PMS1, PMS2, PTCH, PTEN, RB1, RBM5, SMAD4, SMARCA3, SMARCB1, TGFBR2, THBS1, TIMP3, TP53, TP53BP2, TREX1, TUSC3, VHL, WRN* and *WT1*.

[b]Other proto-oncogenes include *AKT1, AKT2, ARAF, ARHGEF5, AXL, BCL2, BCL3, BCL6, BRAF, CCND1, CRK, CRKL, CSF1R, CTTN, CXCL2, CXCL3, DDX6, ECT2, EGFR, ELK1, EMP1, EPHA1, EPHA3, ERBB2, ERBB3, ERG, ETS1, EWSR1, FER, FES, FGF3, FGF4, FGF6, FGR, FLT1, FOS, FOSL1, FURIN, FUS, FYN, GLI1, GLI2, HCK, HRAS, IGF1, IGF2, IRF4, JUN, KIT, KRAS, MAF, MAS1, MCF2, MOS, MPL, MYC, MYCN, MYEOV, NFKB2, NOTCH1, NOV, NRAS, NTRK1, PARK7, PDGFB, PIM1, RAF1, RALB, REL, RET, ROS1, SPI1, SRC, TAL1, TAL2, TCL1A, TERT, THPO, THRA, TLX1, VAV1, VAV2, VEGF, WNT1, WNT2, WNT3, WNT5A* and *YES1*.

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 non-synonymous amino acid changes. This result suggests that proto-oncogenes 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 publication[11] 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 reference[1] 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 reference[1]. 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 reference[1]. 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) database[4], 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).
5. 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).
6. Sydow, J. F. & Cramer, P. RNA polymerase fidelity and transcriptional proofreading. *Curr Opin Struct Biol* **19**, 732–9 (2009).
7. Powell, L. M. *et al.* A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **50**, 831–40 (1987).
8. 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).
9. Thomas, M. J., Platas, A. A. & Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* **93**, 627–37 (1998).
10. 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.

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).