

AD_____

Award Number: W81XWH-11-1-0119

TITLE: RNA Chimeras as a Gene Signature of Breast Cancer

PRINCIPAL INVESTIGATOR: D. Joshua Liao

CONTRACTING ORGANIZATION: University of Minnesota, Twin Cities
Austin, MN 55912

REPORT DATE: May 2013

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE May 2013		2. REPORT TYPE Annual		3. DATES COVERED 15 April 2012 – 14 April 2013	
4. TITLE AND SUBTITLE RNA chimeras as a gene signature of breast cancer				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER W81XWH-11-1-0119	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) D. Joshua Liao				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Minnesota, Twin Cities Austin, MN 55912				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT This project is set to test a hypothesis that breast cancer may express many RNA chimeras not only because there are fusion genes derived from chromosomal translocation but also because of abnormal trans-splicing of RNA transcripts. Many of these RNA chimeras may influence the behaviors of breast cancer via undiscovered mechanisms. We initially planned to establish the world's first comprehensive list of breast cancer specific fusion RNAs. However, last year we found out that our European competitor, Dr. Rolf I. Skotheim, had just submitted a patent application for making the same microarray chip (Patent application number: 20100279890; http://www.faqs.org/patents/app/20100279890#b). Several hundreds of fusion RNAs on their array list overlap with those on our list. While this latest development strengthens the importance of this project, it also forces us to forgo our task of building a similar chip for legal consideration. Since understanding of what chimeric RNAs are formed in breast cancer and how they are formed is still important for us to disclose mechanisms for breast cancer formation and progression, we continue to identify chimeric RNAs from different databases and have obtained several novel findings in the past year, as summarized below.					
15. SUBJECT TERMS- none provided					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU		19b. TELEPHONE NUMBER (include area code)

Table of Contents

	<u>Page</u>
Introduction.....	3
Body.....	3
Key Research Accomplishments.....	6
Reportable Outcomes.....	6
Conclusion.....	6
References.....	7
Appendices.....	8

Introduction

This project is set to test a hypothesis that breast cancer may express many RNA chimeras not only because there are fusion genes derived from chromosomal translocation but also because of abnormal trans-splicing of RNA transcripts. Many of these RNA chimeras may influence the behaviors of breast cancer via undiscovered mechanisms. We initially planned to establish the world's first comprehensive list of breast cancer specific fusion RNAs. However, last year we found out that our European competitor, Dr. Rolf I. Skotheim, had just submitted a patent application for making the same microarray chip (Patent application number: 20100279890; <http://www.faqs.org/patents/app/20100279890#b>). Several hundreds of fusion RNAs on their array list overlap with those on our list. While this latest development strengthens the importance of this project, it also forces us to forgo our task of building a similar chip for legal consideration. Since understanding of what chimeric RNAs are formed in breast cancer and how they are formed is still important for us to disclose mechanisms for breast cancer formation and progression, we continue to identify chimeric RNAs from different databases and have obtained several novel findings in the past year, as summarized below.

Body

Sequence analyses of chimeric RNAs from databases (Task 1):

Since last year's report, the number of putative chimeric RNAs in different databases continues increasing, which on one hand provides us more candidates but on the other hand increases the burden of task 1a. We have analyzed about 2,000 more expression sequence tags (ESTs) of putative chimeras in the past year. The three categories of chimeric or trimeric RNAs, i.e. those with a gap or an overlap or with two partner genes directly joined, have the similar frequencies as reported last year. Also as reported last year, conduction of task 1a leads to a serendipitous finding of trimeric RNA, i.e. one RNA sequence containing three different partners. We have found more such trimers and enlarge our database (Table 1). Moreover, we identified several ESTs that contain four genes' elements, coined as tetrameric RNA or tetrameras (Fig. 1).

Table 1: Trimeric ESTs.

AI335862	BF826714	BF826602	BF764896	BF762577	BF744644	BF331329	BF306729	BE694080
AI924910	AU132130	BF109407	AV744183	AV729389	AV725012	BE814336	BE762537	BE876577
AW608255	BE715872	BE715869	BE715858	BE709675	BE694009	BE696199	BQ348968	AA514694
AW956968	BF803049	BF764896	BF331329	AU142287	BE172179	X93499	BM824189	BG995785
AW999004	BQ689257	BQ689139	BU539467	AI925024	BF814512	BG003110	R19361	BE898652
BC064904	M77198	BM915020	BI004882	BF995070	BF878278	BF109407	AW994480	BE716966
BE074730	BE876742	BE937759	BF987118	BM691077	BM703781	BX109950		

Note: Each of these ESTs contains three sequence elements.

BE762537

TGATAGATTGGTCCAATTGGGTGTGAGGAGTTCACTTATATGTTGGGATTTTTAGGTAGTGGGTGTTGAGCTTGAACGCTTTCTCGATGGGTGTC
GGCAAAGATCCAGGATAAGGAAGGCATTCTCTGATCAGCAGAGTTGATCTTTGCCGGAACAGCTGgGAGGGATGCCTTCCTTGTCTTGGATC
TTTGCTTGACATTCTCAATGGTGTCTACTGTCTATTCTGACCCAGCTCATGGAATTTTTTCATCTTATACTGAGCTCCAGAAAGGACGTAACCTT
AGCATGGATCACAATCAATCAAAAAATAAATAACTACTAAGGATTGGAGAACTCATAGAACAGGTGAAAGACATG

Fig. 1: BE762537 is a tetrameric RNA. Its first 86 nt (boldfaced) belongs to the 2198-2283rd region of the L-strand of mitochondrial (mt) RNA. The following 11 nt sequence (italicized and underlined) is an unmatchable gap. Its 98-168th nt sequence is part of alternatively spliced exon 2 of the UBC mRNA from chromosome (chr) 12, which is followed by a 53-nt (the 268-220th) UBC antisense fragment. Both sense and antisense fragments of UBC, which overlap at the 168th nt (the lowercase "g"), have multiple repeats in the UBC mRNA. The last 149 nt (the 221-369th nt) is part of the ENPP6 mRNA from chr 4.

Identification of chimera and trimers that contain mitochondrial sequence

Conduction of task 1a also led to serendipitous identification of some chimeric and trimeric RNAs that contain mitochondrial (mt) RNA sequence (table 2 as Appendix 1). If any of these

for RNA fusion, because transportation of nRNA from the nucleus to the mitochondria or fusion of the mtRNA to the nRNA should happen after the nuclear transcription has been terminated and splicing has been completed. An important additional support for this new mechanism is our finding that in some ESTs, the downstream partner is fused to the poly-A tail of the upstream partner (Fig. 3), because polyadenylation occurs after pre-mRNA cleavage and splicing completion. Moreover, it remains possible that this new mechanism occurs outside the nucleus, whereas both transcription slippage and trans-splicing of nRNA occur in the nucleus [10] .

AU142287
CAGTGTGTGCTCCCGAGCCTCAGCCCCAAGCTGATTCTTATCTGGAATGTTACTGAATTCTCTGGGTGGCTTTCTGTGGCCCCATGGGA
TGCAGCGTGGGGGCTGTCTGAAGGACCCTGCTTTTCCAGGGGCGAGGGGCTGCCTTTCCTTT**AAAAAAAAAAAAAAAAAGATTTTTACCTGAG**
TAGGCCTAGAAATAAACATGCTAGCTTTTATTCCAGTTCTAACCAAAAAATAAACCTCGTTCCACAGAAAGCTGCCATCAAGTATTTCTCACGC
AAGCAACCGCATCCATAATCCTTCTAATAGCTATCCTCTTCAACAATATACTCTCCGACAATGAACCATAACCAATACTACCAATCAATACTCAT
CATTAAATAATCATAATGGCTATAGCAATAAACTAGGAATAGCCCCCTTTCACTTCTGAGTCCCAGAGGTTACCCAAGGCACCCCTCTGACATCCG
GCCTGCTTCTTCTCACATGACAAAACTAGCCCCCATCTCAATCATgcttttctctCTCTCTTTCCTGCAAGGCGGGCGGAGGAGAGGTTGTGGT
GCTAGTTTCTCTAAGCCATCCAGTGCCATCCTCGTCTGTCAGCGACACAGCTCTCGCCGCCGCCATGACTGAGCAGATGACCTTCGTGGCACC
CTCAAGGGCCACACGGCTGGGTAACCCAGATCGCTACTACCCGCGAGTCCCAGGACATGATCCTCTCCGCTCTCGAGATAAGACCATCATCATG
TGGAAACTGACAGGGATGAGACCAACTATGGAATCCAGCGGCTCTCGGGGGCACTNCCACTTTGTagnngg

Fig. 3: The 1-160th nt matches the 429-588th nt of the last exon, which should have 2248 bp, of the POLDIP3 mRNA from chr 22, indicating an early transcriptional termination followed by polyadenylation (boldfaced and underlined sequence). Following the poly-A sequence, i.e. the 178-526th nt region (italicized grey sequence), is part of the ND2 mRNA from mitochondria. The 527-836th nt sequence belongs to the first two exons of the GNB2L1 mRNA from chr 5, with the first 10 nt (underlined lowercase letters) alternatively initiated from the -10 bp of the GNB2L1 gene. The last 85 nt (underlined) has a few deleted mismatch to the first 88 nt of exon 2 of the GNB2L1. The last 5 nt (agngg) is unmatchable and might belong to the cloning vector.

Identification of a new spliced mtRNA

When conducting tasks 2 and 3, we found that most ESTs could not be detected in the cDNA libraries we constructed from 15 pairs of samples of breast cancers and adjacent normal tissues. This enforces our previous conclusion that the majority of ESTs may be technical artifacts, while the majority of the truly existing chimeric RNAs are derived from chromosomal translation. Several of the possible mechanisms for the formation of such artifacts are proposed in our new publication (Appendex 2). In addition, during experimental verification of ESTs, we identified a novel mtRNA derived from cis-splicing (Fig. 4).

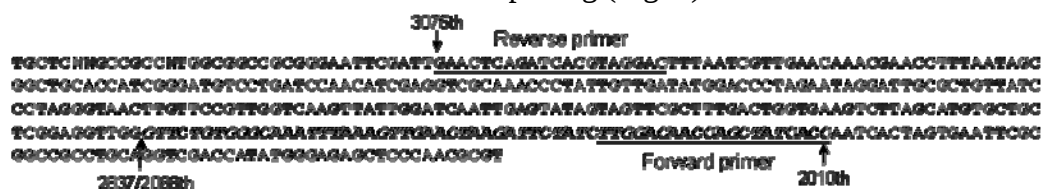


Fig. 4: RT-PCR followed by T-A cloning and sequencing identifies a spliced mtRNA from HEK293 cells in which the 2069-2836th nt region of the mtRNA was deleted during splicing, when aligned with mtDNA by using UCSC browser. The sequence is reverse-complementary to mtDNA. The boldfaced region is the downstream mt exon. Both forward and reverse primers used were underlined. The number of nucleotide (nt) in the mtDNA nt was based on the UCSC browser, with the position of the last nt in the reverse primer and the first nt in the forward primer indicated. The sequences before the reverse primer and after the forward primer belong to the cloning vector.

Technical renovations

When conducting tasks 2 and 3, we encountered two major technical difficulties for experimental verification and detection of RNA. One is that many genes are expressed in association with their antisense transcripts, in line with the literature report that for over 63% of the genes, their RNA transcripts are accompanied by antisense counterparts [11]. Another hurdle is a surprising finding that routine reverse transcription (RT) of RNA can be primed by endogenous primers

existing in the RNA samples. It took us great efforts to solve these technical obstacles by establishing two new cloning methods and a novel method called “cDNA protection assay” that replaces the traditional RNA protection for verification of an RNA. All these methods are now published in RNA Biology (Appendix 2).

It is worth mentioning that our novel strategy to detect RNA by protecting cDNA has several merits. DNA/RNA hybrid has its unique structure and compositions that are distinguishable from DNA/DNA or RNA/RNA hybrid, in part because DNA/DNA contains dA and dT, RNA/RNA contains rA and rU, while DNA/RNA contains all four. These differences should provide us with unique strategies to develop sensitive methods and instruments for the detection and quantification of those DNA/RNA hybrids that are at very low abundance. Such strategies should be applicable and thus intriguing, as endogenous DNA/RNA hybrids in eukaryotic cells are many fewer than the DNA/DNA and RNA/RNA hybrids, especially when a larger DNA/RNA fragment is designed for protection.

Key Research Accomplishments

1. In the past two years we have analyzed over 70,000 putative chimeras from different databases, much more than we originally planned. However, RT-PCR verification suggests that the majority of them may be artifacts, and major technical reasons are described in our publication (Appendix 2).

2. We have identified, for the first time, mt-sequence in chimeric, trimeric and even tetrameric RNAs, suggesting that mtRNAs, especially those transcribed from the non-coding regions of the L-strand, may be combined with nuclear RNA to enlarge the RNA repertoire.

3. We found that human mtRNAs also undergo cis- and trans-splicing, and have identified a novel cis-splicing derived mtRNA.

4. We have established several cloning methods and a new strategy for verification of the existence of RNA, coined as “cDNA protection assay”. These methods are published (Appendix 2)

5. The reference gene methods mentioned in the last year’s report has also been published. (Appendix 3).

6. The PI (Dr. Liao) and the postdoctoral fellows whose salary and position are supported by this grant have also published several other papers. (Appendix 4)

Reportable outcome

1. A longer list of trimeric RNA as a database is made.
2. A paper published to show peers new reference genes and primers for RT-PCR methods.
3. A paper published to describe new cloning methods and RNA verification.
4. Two postdoctoral fellows are supported by the funds and they have publications in the past year.

Conclusion

So far there probably have been over a million of putative chimeric RNAs reported in the literature or deposited in different databases. However, the vast majority of these chimeras remain unverified and therefore are still meaningless to us. For this reason, our work to verify them and to clone their full length sequence is of importance. After analyzing over 70,000 chimeric sequences and determining the expression status of about 3000 selected candidates, we conclude that the majority of putative chimeric RNAs in different databases may be technical artifacts. In a publication we propose major technical reasons for how the artifacts are made. We also conclude that most of those truly existing chimeras are associated with a corresponding change in the genome. Of those chimeric, trimeric and even tetrameric RNAs that truly exist and

occur at the RNA level, mitochondrial RNAs, especially those transcribed from the non-coding regions of the L-strand, participate in their formation. In other words, human mitochondrial RNAs also undergo cis- and trans-splicing and fuse with nuclear RNAs to enlarge the cellular RNA repertoire, which implies a previously unaware mechanism for RNA fusion that may occur at the cytoplasm, but not the nucleus.

References cited

1. Jackson,C.J. and Waller,R.F. (2013) A widespread and unusual RNA trans-splicing type in dinoflagellate mitochondria. *PLoS.One.*, **8**, e56777.
2. Szczesny,R.J., Wojcik,M.A., Borowski,L.S., Szewczyk,M.J., Skrok,M.M., Golik,P., and Stepien,P.P. (2013) Yeast and human mitochondrial helicases. *Biochim.Biophys.Acta*.
3. Burzio,V.A., Villota,C., Villegas,J., Landerer,E., Boccardo,E., Villa,L.L., Martinez,R., Lopez,C., Gaete,F., Toro,V., Rodriguez,X., and Burzio,L.O. (2009) Expression of a family of noncoding mitochondrial RNAs distinguishes normal from cancer cells. *Proc.Natl.Acad.Sci.U.S.A*, **106**, 9430-9434.
4. Villegas,J., Burzio,V., Villota,C., Landerer,E., Martinez,R., Santander,M., Martinez,R., Pinto,R., Vera,M.I., Boccardo,E., Villa,L.L., and Burzio,L.O. (2007) Expression of a novel non-coding mitochondrial RNA in human proliferating cells. *Nucleic Acids Res.*, **35**, 7336-7347.
5. Villegas,J., Zarraga,A.M., Muller,I., Montecinos,L., Werner,E., Brito,M., Meneses,A.M., and Burzio,L.O. (2000) A novel chimeric mitochondrial RNA localized in the nucleus of mouse sperm. *DNA Cell Biol.*, **19**, 579-588.
6. Rackham,O., Mercer,T.R., and Filipovska,A. (2012) The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley.Interdiscip.Rev RNA.*, **3**, 675-695.
7. Woischnik,M. and Moraes,C.T. (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Res.*, **12**, 885-893.
8. Ramos,A., Barbena,E., Mateiu,L., del Mar,G.M., Mairal,Q., Lima,M., Montiel,R., Aluja,M.P., and Santos,C. (2011) Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. *Mitochondrion.*, **11**, 946-953.
9. Tsuji,J., Frith,M.C., Tomii,K., and Horton,P. (2012) Mammalian NUMT insertion is non-random. *Nucleic Acids Res.*, **40**, 9073-9088.
10. Gingeras,T.R. (2009) Implications of chimaeric non-co-linear transcripts. *Nature*, **461**, 206-211.
11. Katayama,S., Tomaru,Y., Kasukawa,T., Waki,K., Nakanishi,M., Nakamura,M., Nishida,H., Yap,C.C., Suzuki,M., Kawai,J., Suzuki,H., Carninci,P., Hayashizaki,Y., Wells,C., Frith,M., Ravasi,T., Pang,K.C., Hallinan,J., Mattick,J., Hume,D.A., Lipovich,L., Batalov,S., Engstrom,P.G., Mizuno,Y., Faghihi,M.A., Sandelin,A., Chalk,A.M., Mottagui-Tabar,S., Liang,Z., Lenhard,B., and Wahlestedt,C. (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**, 1564-1566.

Table 2: Chimeric, trimeric or tetrameric ESTs that contain mitochondrial sequence

#	Mitochondria			Chromosome		Fusion
	Access #	M strand/region	M-span (nt)	NUMT	Partner	
1	DB324119.1	L7212-7515	304	1	M-10	Chimera
2	BE898652.1	L8462-8527, L8569-8969	62, 401	1, 1	M-M-19	Trimer
3	AU142287.1	H4547-4895	349	1	22-M-5	Trimer
4	BF378297.1	L8443-8642	200	1	M-22	Chimera
5	BE716966.1	H7587-7790	204	1	M-1-11	Trimer
6	BE899119.1	H8399-8936	538	1	M-22	Chimera
7	AV744183.1	H4333-4399	67	1	M-19-9	Trimer
8	BF762577.1	L6897-7014, L9645-9779	118,135	1, 1	7-M-M	Trimer
9	BF764896.1	H11165-11248	84	5	2-M-6	Trimer
10	BE709292.1	L10705-10865	161	5	M-7	Chimera
11	BE709354.1	L10705-10864	160	5	M-7	Chimera
12	BE899559.1	H10677-1170	494	5	6-M	Chimera
13	BF083248.1	H14061-14231	171	5	10-M	Chimera
14	BF083272.1	H14061-14233	173	5	10-M	Chimera
15	AV751897.1	H10511-10874	364	5	M-19	Chimera
16	BF306729.1	H10637-10864	227	5	5-M	Chimera
17	BF744644.1	H15540-15721, H14092-14295	182, 204	5	M-M-11	Trimer
18	BE762537.1	L2198-2283	86	17	M-12-12-4	Tetramer
19	BE876577.1	H10689-11361, H84455-8680	493, 226	5, 1	M-M-11	Trimer
20	AV702773.1	H2639-3082	444	17	14-M	Chimera
21	AA514694	L9176-9208	33	1	M-8-8	Trimer
22	AA581515	L7399-7453	55	1, 17	M-17	Chimera
23	AA679609	L7399-7520	122	1	M-22	Chimera
24	AA679609	L7399-7520 (?AA679609)	122	1	M-22	Chimera
25	AI925024	H2037-2239	203	3, 11, 5	14-M-13	Trimer
26	AW134795	L1619-1671	53	11	M-1	Chimera
27	AW370799	H8966-9070	105	1, 5	19-M	Chimera
28	AW753072	L14400-14455	56	18, 5, 17, 5	M-9	Chimera
29	AW821349	L12015-12081	67	5	M-16	Chimera
30	AW898803	H2647-2773	127	None	X-M	Chimera
31	AW950200	H7276-7520	245	1	5-M	Chimera
32	BE074730	L6707-6844	48	1, x, 2	M-6-17	Trimer
33	BE162186	H5117-5322	205	1	M-13	Chimera
34	BE876742	H7398-7457	60	1	19-M-1	Trimer
35	BE937759	L9080-9144	65	1	9-M	Chimera
36	BF852160	L4703-4822	120	1	17-M	Chimera
37	BF987118	H6568-6604	37	1	M-21-7	Trimer
38	BF988359	L6876-6938, L6950-7025	63, 76	1, 1	7-M-M	Chimera
39	BG995785	H10473-10589, L15168-15217	117, 50	5, 5	M-M-1	Trimer
40	BM691077	H14103-14159	57	5	17-M-1	Trimer
41	BM703781	H9567-9605	39	1	4-M	Chimera
42	BM997144	L7189-7520	332	1	M-2	Chimera
43	BP348380	H2852-3018	167	5	19-M	Chimera
44	BQ300150	H8936-8995	60	1	3-M	Chimera
45	BQ348968	L12338-12505	168	None	M-8-2	Trimer
46	BQ638079	H11651-11698	48	5, 5	5-5-5-M	Tetramer
47	CV385666	L2620-2837	218	11	M-7	Chimera
48	DA086571	H7201-7521	321	1	M-1	Chimera
49	DA182598	H2417-2733	317	11,3,6,17,5	M-11	Chimera
50	DA365070	H1613-1672	60	11, 7, 5	M-7	Chimera
51	DA511096	H7192-7533	342	1	M-1	Chimera
52	DA757571	H2421-2762	342	None	M-1	Chimera
53	DB314922	L7393-7515	123	1	M-1	Chimera
54	DB324119	L7212-7515	304	1	M-10	Chimera
55	AW994480	L2654-2804, L2814-2880, L2894-2984	63, 67, 91	None	M-2-8	Trimer
56	BE694080	L2654-2984 (?AW994480)	63, 67, 91	None	M-2-8	Trimer
57	BF826602	L11053-11116	64	5, 5	12-15-M	Trimer

Note: The first and last nt positions at the mtDNA of an mt sequence (M) are indicated, based on UCSC browser, while its length (span in the number of nt) may not always be calculated due to possible deletion of several nt. The chromosome or chromosomes that harbor an NUMT homologous to the mt sequence are indicated. The order of each partner in the chimera, trimer or tetramer is shown in the 5'-to-3' orientation.

New methods as alternative or corrective measures for the pitfalls and artifacts of reverse transcription and polymerase chain reactions (RT-PCR) in cloning chimeric or antisense-accompanied RNA

Chengfu Yuan,¹ Yongming Liu,^{2,*} Min Yang³ and D. Joshua Liao^{1,*}

¹Hormel Institute; University of Minnesota; Austin, MN USA; ²Department of Biochemistry and Molecular Biology; Guilin Medical University; Guilin, China; ³School of Laboratory Medicine; Chengdu Medical College; Chengdu, Sichuan, China

Keywords: antisense, chimeric RNA, cDNA protection assay, reverse transcription, polymerase chain reaction, cDNA cloning, RNA protection assay

We established new methods for cloning cDNA ends that start with reverse transcription (RT) and soon proceed with the synthesis of the second cDNA strand, avoiding manipulations of fragile RNA. Our 3'-end cloning method does not involve poly-dT primers and polymerase chain reactions (PCR), is low in efficiency but high in fidelity and can clone those RNAs without a poly-A tail. We also established a cDNA protection assay to supersede RNA protection assay. The protected cDNA can be amplified, cloned and sequenced, enhancing sensitivity and fidelity. We report that RT product using gene-specific primer (GSP) cannot be gene- or strand-specific because RNA sample contains endogenous random primers (ERP). The gene-specificity may be improved by adding a linker sequence at the 5'-end of the GSP to prime RT and using the linker as a primer in the ensuing PCR. The strand-specificity may be improved by using strand-specific DNA oligos in our protection assay. The CDK4 mRNA and TSPAN31 mRNA are transcribed from the opposite DNA strands and overlap at their 3' ends. Using this relationship as a model, we found that the overlapped sequence might serve as a primer with its antisense as the template to create a wrong-template extension in RT or PCR. We infer that two unrelated RNAs or cDNAs overlapping at the 5'- or 3'-end might create a spurious chimera in this way, and many chimeras with a homologous sequence may be such artifacts. The ERP and overlapping antisense together set complex pitfalls, which one should be aware of.

Introduction

A recent advance in RNA research suggests that virtually the entire non-repeat part of the human genome is transcribed, at least at some times or in some cell types,^{7,17} with a tally of 161,000 transcripts so far.²⁵ Moreover, it is estimated that over 63% of RNA transcripts are accompanied by antisense counterparts,²⁴ and the Unigene database of the National Center for Biotechnology Information (NCBI) contains over 123,000 human antisense entries.²⁶ One meaning of these figures is that, for most genomic loci, both strands of the DNA double helix are transcribed.^{6,19} Most antisense transcripts may be non-coding, but there are still many that do encode proteins. For example, the DNA strand opposite to the one encoding the THRA (17q11.2), CDK4 (12q14.1), CCND1 (11q13) and GAPDH (12p13) genes harbors the NR1D1, TSPAN31, LOC100996515 and LOC100996356 protein-coding genes, respectively, as shown in

the NCBI. Cloning cDNA often involves reverse-transcription (RT) and polymerase chain reactions (PCR), but the situation wherein antisense is also expressed often sets pitfalls and hurdles, which are widely neglected, in our way of cloning the 5'- or 3'-end of cDNA or determining from which DNA strand an RNA is transcribed. For instance, it may not be easy to clone the 5' and 3' ends of the so-called ncRNA_{CCND1}⁴³ and to determine whether it is transcribed from the same strand as the CCND1 or as the LOC100996515.

Another ribonomic advance suggests that transcripts from about 65% of the human genes form chimeric RNA with a transcript from another gene. This other gene in most cases is nearby on the same chromosome but can also be located on another chromosome.^{7,17} Actually, modern RNA-sequencing technologies have provided us with thousands of RNA chimeras.¹⁴ A tiny number of them are known to be transcribed from fusion genes that are formed due to genetic alterations, such as chromosomal

*Correspondence to: Yongming Liu and Joshua Liao; Email: liuym@glmc.edu.cn and djliao@hi.umn.edu
Submitted: 02/22/13; Revised: 03/31/13; Accepted: 04/05/13
<http://dx.doi.org/10.4161/rna.24570>

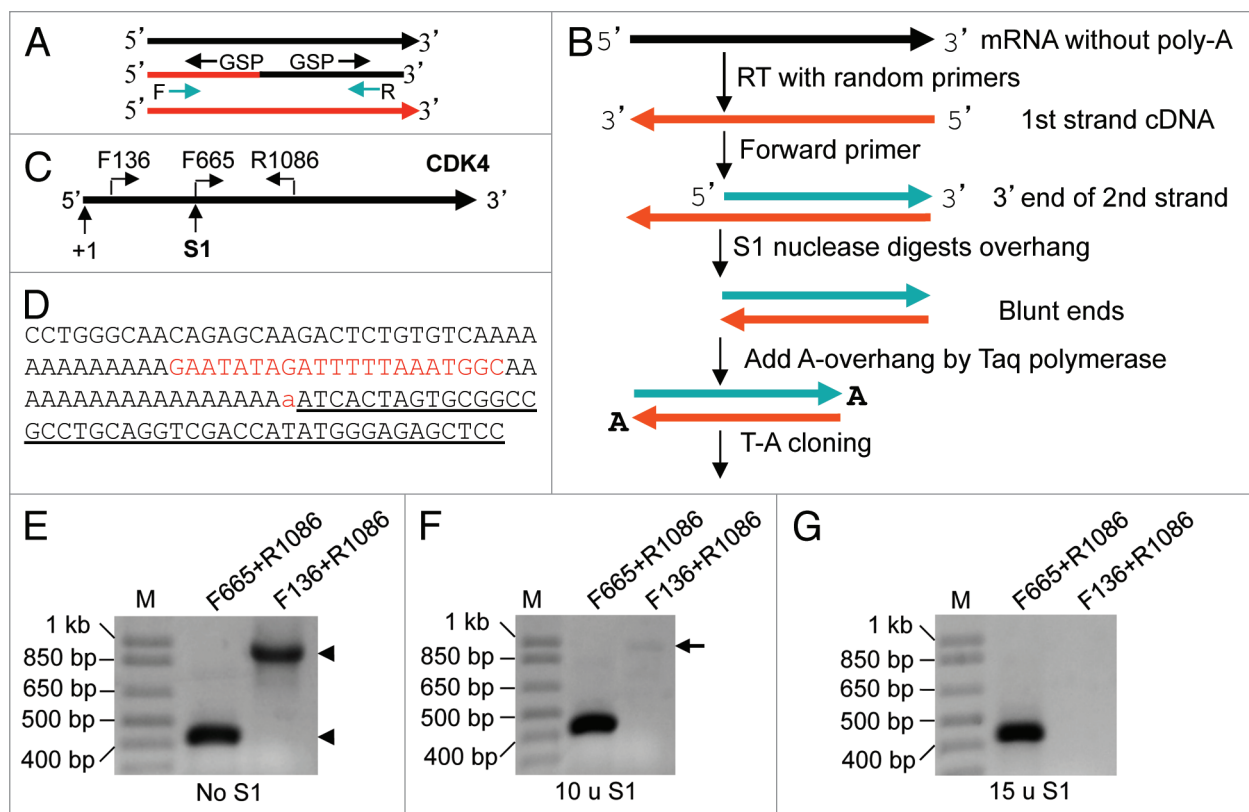


Figure 1. Cloning RNA 3' end. (A) Two hurdles for cloning the 5' or 3' end, and for PCR amplification, of chimeric cDNA: (1) Gene-specific primer (GSP) used in RACE amplifies the 5' or 3' end of not only the chimeric cDNA but also the cDNA of a parent gene (black or red line). (2) Forward (F) or reverse (R) primer primes not only the chimeric cDNA but also the cDNA of a parent gene, making the first several cycles of PCR less efficient. (B) Our strategy for cloning RNA 3' end: After RT with random hexamers, a forward primer of the gene of interest and Taq are used to synthesize the 3' part of the second cDNA strand. S1 is added to cut the 3'-overhang of the first strand. The cDNA blunt ends are then appended with a dA by Taq, followed by T-A cloning. (C) Illustration of the locations of primers and the S1 cutting site in the CDK4 mRNA. (D) Part of the 3' sequence obtained, in which the lowercase "a" is added by Taq and the underlined sequence belongs to the T-A vector. The sequence matches completely to the CDK4 mRNA. Note that there is an internal poly-A sequence 21 nt upstream of the authentic poly-A tail. (E–G) Both pairs of primers (F665+R1086 and F136+R1086) designed to amplify the 5' and the middle regions of the mRNA, respectively, can amplify the RT product (cDNA) of RNA from HeLa cells without S1 digestion (arrowheads). F665+R1086 can still yield a band from the cDNA digested with 10 or 15 units of S1, while the F136+R1086 yielded only a very faint band (arrow) when 10 units were used and no band when 15 units were used.

translocation and genomic DNA deletion or amplification.⁹ Unfortunately, the vast remaining majority, i.e., those not associated with a known genomic alteration, remain putative and are not very meaningful to us so far, because their existence has not been verified with a vigorous method and because their full-length sequence has not been cloned and, thus, their open reading frame is unclear.¹⁶ These weaknesses are due mainly to the lack of reliable and efficient approaches of cloning and verification. Current RNA sequencing technologies are reliant on RT or on the principles similar to RT,¹³ provide only short sequences, and have poor strand-specificity,³⁵ thus only suitable for screening, but not for verification, of long RNA. Cloning methods involving RT-PCR may result in artificial chimeric cDNA^{8,11,23,30,35,36,38} in part because template switching may occur during RT^{11,23,33} and mis-priming can occur in PCR. RNA protection assay is the most commonly used method to verify the true existence of an RNA, in which an in vitro synthesized complementary RNA (cRNA) is used to hybridize with the parental RNA in solution, followed by RNase digestion of the non-hybridized RNA.

This method does not involve RT or PCR, which on one hand increases its reliability, but on the other hand makes the method very inefficient.¹³ Also problematically, the protected RNA cannot be directly sequenced to confirm its identity. For these reasons, most attempts to verify chimeric RNAs still, unfortunately, use problematic RT-PCR. Moreover, it needs to distinguish a true chimeric cDNA end cloned with a routine 5' or 3' RACE (rapid amplification of cDNA end) method from the cDNA end of the parent mRNA, as depicted in Figure 1A. Actually, because in most cases the mRNA of each parent gene is more abundant, the cloned cDNA end is more likely to belong to the parent mRNA.

We attempt to develop methods that are devoid of the above-described weaknesses in cloning and verifying chimeric or antisense-accompanied RNA. Although not yet reaching this aim, we have established new methods for cloning cDNA ends and have established a cDNA protection assay to supersede RNA protection assay, as described in this report. Sometimes "RNA," but not "mRNA," is termed herein because our methods are also suitable for cloning those long RNAs and chimeric RNAs that

Table 1. Primers used

Primer	Sequence	Primer	Sequence
NewA	5'-GTGGAGTCTACGCGAACTTGTCTT-3'	CDK4R822	5'-TCCACATGTCCACAGGTGTTGC-3'
NewC	5'-GTGGAGTCTACGCGAACTTGTCC-3'	CDK4F933	5'-GATGACTGGCCTCGAGATGT-3'
NewB mixture	5'-TCAGGATTGATGGTGCCTACAGC13V-3' (V = A,G,T)	CDK4R1086	5'-AGGCAGAGATTGCTGTTGT-3'
NewD	5'-TCAGGATTGATGGTGCCTACAGC-3'	CDK4F1096	5'-TGCAGCACTCTTATCTACATAAGGAT-3'
HPRT1F123	5'-CTTCCTCCTCTGAGCAGTC-3'	TSPAN31F73	5'-AAGCTGTCGGGGTCTGGAA-3'
HPRT1R683	5'-AACACTTCGTGGGGTCTTT-3'	TSPAN31F647	5'-CTTAAGCATTACAGACGAAGC-3'
CCND1F70	5'-TAGCAGCGAGCAGCAGAGTC-3'	TSPAN31R860	5'-ACCTAGATATCCCTAAGG-3'
CCND1F183	5'-CCCAGCTGCCAGGAAGAGC-3'	TSPAN31R1668	5'-CTTGGAAGAAGGGACTTTCC-3'
CCND1R981	5'-TTGACTCCAGCAGGGCTTCG-3'	MYCF125	5'-GCGCTGAGTATATAAAGCCGGTT-3'
CCND1R1067	5'-TGTGCAAGCCAGGTCCACCT-3'	MYCR838	5'-CCACCGCGCTGTTGTCTCC-3'
CDK4F136	5'-GTATGGGGCCGTAGGAACCG-3'	BCAS4E1F	5'-TCCTGATGCTGCTGTTGGAC-3'
CDK4F665	5'-TCTGGTGACAAGTGGTGGA-3'	BCAS3E25R	5'-CATACACAGGGACCGAGCTT-3'
NewDCF933	5'-TCAGGATTGATGGTGCCTACAGCGATGACTGGCCTCGAGATGT-3'		
NewDTF647	5'-TCAGGATTGATGGTGCCTACAGCCTTAAGCATTACAGACGAAGC-3'		

Note: The number in the primer indicates the first (for forward) or the last (for reverse) nucleotide of that primer in the position of the mRNA. Thus, the range between the F and R numbers should normally be the size of the RT-PCR amplified DNA fragment in agarose gel.

are non-coding. Some pitfalls and artifacts of RT and PCR that are widely neglected in the literature are also described to alert the peers.

Results

Cloning RNA 3' end. The 3' end of long RNAs is usually cloned by using a poly-dT oligo to prime the poly-A tail of the RNA or by ligating a linker sequence to the 3' end since about 50% of the long RNAs lack a poly-A tail,³⁷ although they likely have a poly-A signal. In our strategy, RNA can be primed with random hexamers in RT. Taq DNA polymerase (Taq for brevity) and a forward primer of the interested gene are used to synthesize the 3' part of the second cDNA strand, which is also the 3' part of the parental RNA. S1 nuclease (S1 for brevity) is added to cut off the 3'-overhang of the 1st cDNA strand (Fig. 1B), since S1 digests single-stranded, but not double-stranded, DNA or RNA. The blunt ends of the double-stranded cDNA are then appended with a dA by Taq to allow cloning the fragment into a T-A vector (Fig. 1B).

As an example, unDNased RNA from HeLa cells was converted in RT to the first cDNA strand by random hexamers. The CDK4F665 forward primer (all primers listed in Table 1) and PCR Mastermix were mixed with the RT product to synthesize the second strand of CDK4 cDNA by incubation at 72°C for 10 min. S1 was added to cut off the single-stranded part of the 1st cDNA strand upstream of the F665 primer (Fig. 1C). After inactivation of S1 and purification of the double-stranded cDNA, PCR of the cDNA with F136+R1086 primers did not yield signal, which confirmed that the region upstream of F665 (including the F136 sequence) had been digested by S1, while PCR with F665+R1086 yielded a band (Fig. 1E–G), indicating that the double-stranded part could withstand the S1. The amount of S1 might need to be optimized for different target genes, due

to the difference in the residuals of RNAs and single stranded cDNAs to be digested, since herein 10 units of S1 was not sufficient (Fig. 1F). A portion of the purified double-stranded cDNA was then cloned into a T-A vector. Sequencing a resultant plasmid and aligning the sequence with the CDK4 mRNA sequence (NM_000075.3) revealed that the canonical 3' end, including the whole poly-A tail (Fig. 1D), was fully cloned.

Cloning RNA 5' end by G-tailing. Tailing the 3' end of a DNA with terminal deoxynucleotidyl transferase (TdT) is a traditional method for several different purposes, including cloning of the 3' end of a cDNA. We prime the RNA of interest with a reverse primer in RT to convert the RNA to the first cDNA strand (Fig. 2). After removal of dNTP and short oligos, TdT and dGTP are used to append a poly-dG tail to the 3' end of the cDNA, which is the RNA 5' end. A poly-dC mixture is used to prime the poly-dG for synthesis of the second cDNA strand. This poly-dC mixture, referred to as NewB, contains four oligos with a linker sequence (dubbed as NewD) at the 5' end and with one of the four bases at the 3' end, so that one of the four oligos can be anchored on the last nucleotide (nt) of the cDNA. A reverse primer and NewD will then be used in PCR to amplify the double-stranded 3' part of the cDNA (Fig. 2).

As an example, unDNased RNA from HeLa cells was converted in RT with a HPRT1 reverse primer (R683) to the first cDNA strand, followed by removal of dNTP, primers and other short oligos by running the reaction through a RapidTip2, followed by washing and precipitation with ethanol. The cDNA was then tailed with a poly-dG using TdT and dGTP. NewB was used to prime the poly-dG tail for synthesis of the second cDNA strand. NewD and the R683 were used as the primers in PCR to amplify the double-stranded part of the HPRT1, which resulted in a fuzzy band. Purification of this fuzzy band from agarose gel followed by second PCR resulted in a clear band of the correct size, which was cloned into a T-A vector. Sequencing

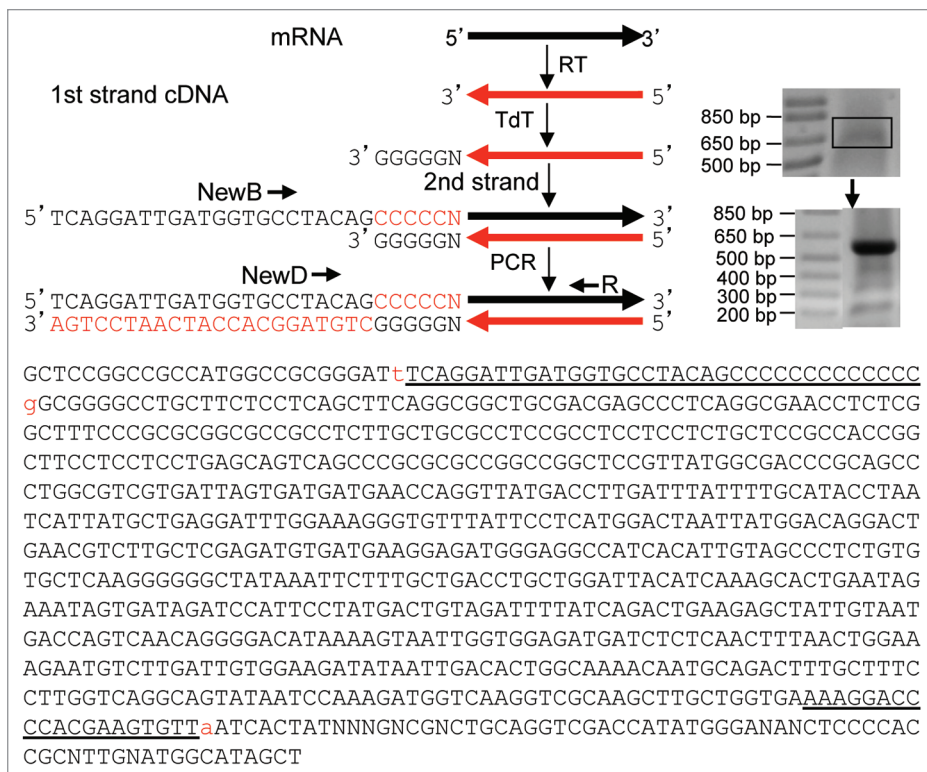


Figure 2. Cloning RNA 5' end. In our strategy, RNA is converted in RT to the first strand of cDNA with a reverse primer of the interested gene. TdT and dGTP are used to append a poly-dG to the cDNA 3' end, which is the 5' end of the RNA. NewB is used to prime the synthesis of the second cDNA strand with PCR Mastermix. NewB is a mixture of four poly-dC oligos with a linker sequence (NewD) at the 5' end and one of the four bases at the 3' end (Table 1) and, thus, can be anchored on the last nt of the cDNA. The NewD and a reverse primer of the desired gene are then used in PCR to amplify the double-stranded cDNA, followed by T-A cloning. As an example, RNA from HeLa cells was RT with the HPRT1R683 primer. The HPRT1 cDNA was tailed with a poly-dG, followed by PCR with NewD+R683 that yielded a fuzzy band. Excision and purification of this band (boxed) as the template for a second round of PCR with the same primers resulted in a dominant band and several smaller bands. Cloning and sequencing the dominant band confirm that it is G-tailed 5' end of HPRT1 mRNA. In the sequence obtained, the lowercase "t" before the NewB (underlined) and the "a" (added by Taq) after the R683 (underlined) were the cloning sites. The sequences before the "t" and after the "a" belong to the T-A vector. The lowercase "g" after NewB is the first nt of HPRT1 mRNA anchored by the NewB.

one resultant plasmid followed by alignment of the sequence with the HPRT1 mRNA sequence (NM_000194.2) confirms that a NewB primer has indeed anchored on the last nt of the HPRT1 (Fig. 2).

cDNA protection assay. We established a new strategy using cDNA to supersede cRNA and accordingly using S1 to replace RNase in RNA protection assay. In this strategy, an RNA aliquot is primed by random hexamers in RT and converted to cDNA. An aliquot of the cDNA is used to hybridize with a commensurate amount of the RNA (Fig. 3A). S1 is added to digest the non-hybridized cDNA and RNA and then is inactivated. PCR with gene-specific primers (GSP) ensues to amplify a fragment of the RNA-protected cDNA. As a negative control, an equal aliquot of the cDNA is directly digested with the same amount of S1 to ensure that without hybridization, no cDNA is left to be amplified in PCR. The amount of S1 may need to be optimized for each target gene because too-low enzyme activity may not be sufficient to remove all single-stranded cDNAs and may thus

cause false positivity, whereas too much enzyme is known to have weak activity toward double-stranded DNA.

In the MCF7 human breast cancer cell line, the BCAS4 (breast cancer amplified sequence 4) at the 20q13 of the human genome forms a fusion gene with the BCAS3 at the 17q23, which is transcribed and alternatively spliced to different chimeric RNAs, most of which contain exon 1 of BCAS4 and exons 24 and 25 of BCAS3.^{5,21,29} We performed RT with random hexamers and then PCR with primers at exon 1 of BCAS4 and exon 25 of BCAS3, which resulted in a dominant band at about 650 bp and several minor and smaller bands (Fig. 3B). Pretreatment of the RNA sample with DNase I followed by inactivation of the enzyme (as discussed later) caused partial losses of the minor bands, which redistributed the primers and thus increased the abundance of the dominant band (Fig. 3B). An aliquot of the cDNA was hybridized with an equivalent amount of the RNA sample. Moreover, the RT product (1/20) was also used in PCR to amplify a fragment of CCND1, which was purified from agarose gel. Half of the purified CCND1 cDNA was added into the hybridization reaction as an indicator of whether double-stranded DNA could withstand the hybridization and the S1 digestion. After hybridization, S1 was added to digest the non-hybridized RNAs and cDNAs, while the same amount of the cDNA as used in hybridization was also

S1-treated as a negative control (Fig. 3C). After inactivation of S1, PCR with the BCAS primers yielded a band from the S1-treated cDNA/RNA hybrids, but not from the non-hybridized cDNA, as expected (Fig. 3C). Similarly, PCR amplification of CCND1 as done in the above also yielded the anticipated band from the hybridized cDNA but not from the non-hybridized counterpart (Fig. 3C). Cloning the BCAS4-BCAS3 band and sequencing three resultant plasmid clones reveal that in clones 1 and 2, the canonical 3' end of exon 1 of BCAS4 was fused to the canonical 5' end of exon 24 of BCAS3, whereas the clone 3 lacked the last 2 nt (i.e., AG) of exon 1 of BCAS4 and the whole exon 24 of BCAS3 (Fig. 3C).

RT product primed by endogenous random primers in RNA samples. When performing RT, we often set up a reaction without adding primers as a negative control, but this reaction still and always yielded cDNAs. As examples, PCR with 1/20 (1 μ l) of such non-primer RT product as the template could amplify CCND1, CDK4 or HPRT1 cDNA (Fig. 4A). The same

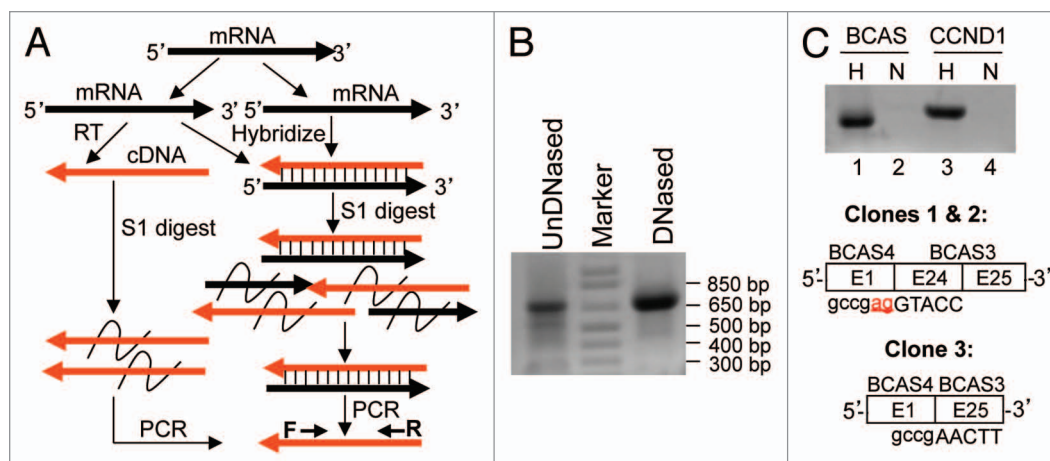


Figure 3. cDNA protection assay. (A) In this strategy, an RNA aliquot is converted to cDNA in RT with random hexamers. An aliquot of the cDNA is hybridized with an equivalent amount of RNA, followed by digestion of the non-hybridized cDNA and RNA with S1. S1 is inactivated and PCR with gene-specific primers (F and R) ensues to amplify the RNA-protected cDNA. As a negative control, an equal aliquot of cDNA is digested with S1 to ensure that without hybridization, no cDNA is left for amplification by PCR. (B) RT with random hexamers and with RNA sample from MCF7 cells that was not treated or was treated with DNase I followed by inactivation of the enzyme. PCR with BC54F1+BCAS3R25 primers detects a dominant band at about 650 bp and several minor and smaller bands. (C) An equal amount of RT product (cDNA) was hybridized (H) or non-hybridized (N) with a commensurate amount of RNA, followed by S1 digestion. PCR with BCAS4F1+BCAS3R25 primers detected the dominant BCAS4-BCAS3 cDNA (BCAS) in hybridized, but not in non-hybridized, RNA aliquot. As a control, a CCND1 PCR fragment was added into the hybridization reaction as an indicator that double-stranded DNA could withstand the hybridization and the S1 digestion as it can be amplified by PCR, whereas single-stranded CCND1 cDNA in non-hybridized RT product was digested by S1 and thus could not be amplified by PCR. Cloning the BCAS band and sequencing three randomly selected plasmid clones reveal that in clones 1 and 2, the 3' end of exon 1 of BCAS4 was fused to the 5' end of exon 24 of BCAS3, whereas clone 3 lacks the last two nt (underlined "ag" shown in clones 1 and 2) of exon 1 of BCAS4 and the whole exon 24 of BCAS4.

CDK4 primers did not produce any band when the template was replaced by water (lanes 1 vs. 3 in Fig. 4A), confirming that the PCR reagents were not contaminated by cDNA templates.

In our routine practice, we often treated RNA samples with low concentration (several units) of DNase I, followed by inactivation of the enzyme with different methods including protein extraction with phenol/chloroform, so that genomic DNA residuals would not be mis-primed in the ensuing RT-PCR.⁴¹ RNA samples from HeLa, PC-3 and ZR75-1 cell lines that were pre-treated with DNase I followed by inactivation of the enzyme with 15 mM EDTA at 72°C for 15 min were used in RT without adding primer. PCR with the RT product as the template could still amplify several genes' cDNA (Fig. 4B). Treatment of RNA samples with a much larger amount of DNase I could not eliminate the PCR-amplified bands, although the DNase activity could not be completely inactivated and the remaining activity decreased the detected level (data not shown). These results suggest that RNA specimens contain endogenous random primers (ERP) for RT that cannot be removed by DNase treatment.

Antisense-caused RT-PCR artifacts. We infer that when an antisense is expressed and overlaps with the sense RNA at their 5' or 3' ends, any cloning approaches that involve RT-PCR, including our method, may create artifacts, although many peers still use RT-PCR in cloning under this situation. We used the CDK4/TSPAN31 relationship to test this hypothesis, since the CDK4 mRNA and the TSPAN31 mRNA overlap at their last 517 nt (Fig. 5A). We designed a forward primer at the penultimate exon of CDK4 (CF933) or TSPAN31 (TF647); one set of these two primers also contained newD sequence at the 5' end as a linker

(NewDCF933 and NewDTF647). Other primers are illustrated in Figure 5A. Forward primer of one strand can also serve as reverse primer of the opposite strand. UnDNased RNA from HeLa cells was used in RT with the NewDTF647 as the GSP, which should specifically convert the CDK4 mRNA to cDNA (RT-B in Fig. 5B) if the mRNA reaches the TF647 region as we hypothesized. PCR using this RT-B product as the template and the CF136+R822 as the primer pair could indeed amplify a correct CDK4 band as expected (lane 4 in Fig. 5B). PCR with NewD as the reverse and the CF1096 as the forward primers resulted in a 1.5 kb band (Lane 6 in Fig. 5B). Cloning and sequencing this band confirmed that it was part of CDK4 mRNA containing the 84-bp intron 5 of TSPAN31. These results together indicate that some CDK4 mRNAs reach at least the TF647 region, making the overlapped region much longer than what is shown in the NCBI (thick arrow in Fig. 5A).

Similarly, RT primed by the NewDCF933 should specifically convert the TSPAN31 mRNA to cDNA if the mRNA reaches the CF933 region (RT-A in Fig. 5B). PCR using this RT-A product as the template and the TF73+TR860 as the primer pair yielded the correct TSPAN31 band (the top band in lane 1 in Fig. 5B). However, a smaller band was also produced that was confirmed by T-A cloning and sequencing to be the LMTK2 mRNA from chromosome 7, but not the TSPAN31. Alignment of the LMTK2 and CDK4 sequences suggests that the TMTK2 cDNA is more likely to be primed by an endogenous primer in the RNA sample as described above, but not by the NewDCF933, indicating that RT using GSP is not so gene- and strand-specific as it is supposed to be.

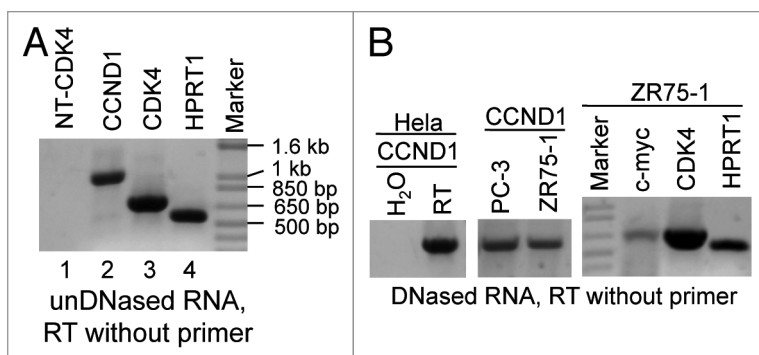


Figure 4. RT primed by endogenous random primers (ERP). **(A)** UnDNased RNA sample from HeLa cells was used in RT without adding primers. The RT product (1/20) was used as the template in PCR with the F70+R1067 for CCND1, F136+R822 for CDK4 and F123+R683 for HPRT1, respectively. As a negative control, the same CDK4 primers were used in a PCR with H₂O to replace the RT product as the template (lanes 1 vs. 3). **(B)** RNA samples from HeLa, PC-3 and ZR75-1 cell lines were treated with DNase I, followed by inactivation of the enzyme. RT was then performed without adding primers. The RT product was used as the template in PCR amplification of CCND1, CDK4 and HPRT1 as in **(A)** or of c-myc with the F125+R838 primers. In an addition, CCND1 PCR, the RT product, was superseded by H₂O as the template.

PCR using the RT-A product as the template and the NewD+TF647 as the primer pair did not yield any band, which was discrepant to the results in lane 1 (lanes 1 vs. 5 in Fig. 5B). More surprisingly, PCR using TF73+TR860 as the primers and RT-B as the template yielded the same two bands as when RT-A was used as the template (Lanes 1 vs. 3 in Fig. 5B), although the RT-B primed by NewDTF647 was not supposed to convert the 5' part of TSPAN31, or any RNA from this region, to cDNA. Similarly, PCR using CF136+R822 as the primers and RT-A as the template generated the same band as when RT-B was used as the template (Lanes 2 vs. 4 in Fig. 5B), although the RT-A primed by NewDCF933 was not supposed to convert the 5' part of CDK4, or any RNA from this region, to cDNA. A reasonable explanation for these inconsistent results is that some CDK4 and TSPAN31 mRNAs have an unprotected 3'-end at the overlapped region, serving as the primer to extend its cDNA as illustrated in Figure 5C. This extension may happen in RT or in the ensuring PCR as discussed later. In other words, the bands in lanes 2 and 3 in Figure 5B do not have the corresponding RNA as the original template and thus are wrong-template artifacts. The TSPAN31 band (the top one) in lane 1 of Figure 5B might be such wrong-template artifact as well, which explains why NewD+TF647 failed to yield a PCR product (lane 5 in Fig. 5B). In line with this conjecture, a PCR with the RT-A product as the template and the CF933+TR1668 as the primer pair did not produce any band (data not shown).

Discussion

Features of our cloning methods. Routine 5' or 3'RACE usually can only clone short cDNA fragments, sometimes making it unclear whether the cloned cDNA end belongs to a chimeric RNA or to an mRNA of the parent gene (Fig. 1A), in part

because the first and last exons are often very large. Moreover, 5'RACE is difficult as its first several steps are manipulations of fragile RNA. Our cloning methods start with RT and soon proceed with the synthesis of the second cDNA strand, with all later steps involving only double-stranded cDNA that is much more stable. Since almost the entire second strand can be synthesized, either one of our cloning methods can clone virtually the entire cDNA. A difference is that our 5' cloning method involves PCR amplification and, thus, is more efficient, whereas our 3' cloning method is a non-amplified approach with low efficiency but high fidelity. In addition, our 3' method does not require a poly-dT primer, allowing cloning those RNAs without a poly-A tail and eliminating mis-priming to an internal poly-A sequence. Actually, we once used 3'RACE to clone the 3' end of TSPAN31 with a poly-dT primer that contains a linker (coined as NewA; Table 1), followed by PCR with the linker sequence as a primer (coined as NewC; Table 1). The 3' end cloned lacks the last 17-nt sequence (GAC CAT TAA AAA AAA AA) because there is a 14-adenine sequence in front of it (data not shown). If needed, however, our method can still use poly-dT

primers in RT, alone or in combination with PCR, to enhance the cloning efficiency.

In our practice of molecular cloning, poly-dT primer is often used to prime poly-A tail in RT, whereas a poly-dG oligo longer than a hexamer (GGG GGG) is technically difficult to be synthesized, purified and verified and, thus, is much more expensive. Therefore, tailing a cDNA with poly-dG followed by priming it with a poly-dC oligo becomes the only practical choice for our 5' end cloning method. The length of the poly-dG tail may be different among tailed targets, but one of the four oligos in the NewB mixture (Table 1) should be anchored on the last nt of the targeted cDNA, regardless of the length of the poly-dG tail and whether the 3' end of the first cDNA strand has been added with several nt by the MMLV during the RT.^{4,18} However, it remains possible that the NewB mis-primers an internal poly-G sequence, which has actually happened in our practice.

Similar to routine 5' and 3' RACEs, our methods only use a single GSP and, thus, may still cause unspecific bands as shown in Figure 2. Moreover, cDNA may have breakages and, similar to routine RACEs, our methods cannot distinguish a genuine cDNA end from a spurious one. In addition, transcription may be initiated from or terminated at alternative sites. Therefore, cloning multiple bands and sequencing multiple plasmid clones are strongly recommended, not only to avoid artifacts but also to increase the chance of identifying alternative 5' or 3' ends.

Merits of cDNA protection assay. The strategy to protect a cDNA instead of the parental RNA has four major merits: (1) After being protected by the parental RNA, the cDNA can be PCR-amplified, which dramatically increases the sensitivity. If part of the cDNA is an RT artifact, it will not be protected because the single-stranded part of the cDNA or of the parental RNA will be digested by S1. Single-stranded DNA is about 5-fold more sensitive to S1 than RNA, as stated in the

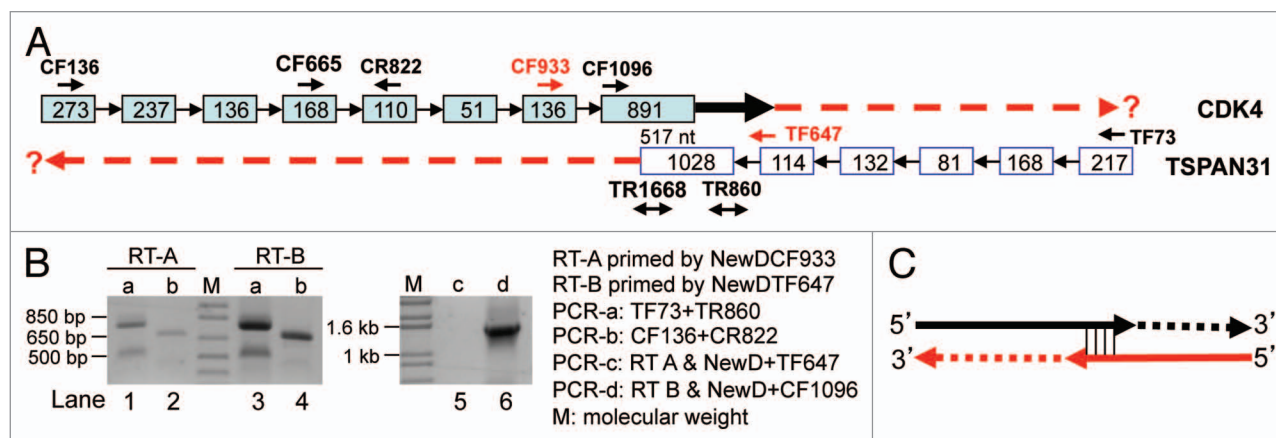


Figure 5. RT with linker-containing GSP to detect CDK4 (NM_000075.3) and TSPAN31 (NM_005981.3). **(A)** Illustration of the CDK4/TSPAN31 relationship according to the NCBI, with the locations of CDK4 forward (CF) or reverse (CR) primers and TSPAN31 forward (TF) or reverse (TR) primers indicated. Boxes represent exons with their length indicated as the number of nt. Note that the two mRNAs overlap at their last 517 nt. **(B)** RT of unDNased RNA from HeLa cells primed by the NewDCF933 should specifically convert the TSPAN31 mRNA to cDNA (RT-A), whereas RT primed by the NewDTF647 should specifically convert the CDK4 mRNA to cDNA (RT-B), if the mRNAs reach the regions of these primers. These two RT products were used as the template in PCR with either the TF73+TR860 (PCR-a, lanes 1 and 3) or the CF136+CR822 (PCR-b, lanes 2 and 4) as the primer pair or in PCR with the primer pair of NewD+TF647 (PCR-c, lane 5) or NewD+CR1096 (PCR-d, lane 6). The results in lanes 4 and 6 together with sequence data suggest that some CDK4 transcripts reach the TF647 region (thick black arrow in **A**). **(C)** When sense or antisense RNA has an unprotected 3' end overlapping with the other, the overlapping sequence may serve as a primer in RT to extend its 3' end with its antisense as the template. The extension may also occur in the ensuing PCR, as depicted in **Figure 6A**. Because the mRNA does not really have this extended part (dashed lines), the corresponding PCR product, such as the bands in lanes 1, 2 and 3 in **(B)**, is a wrong-template artifact.

supplier's datasheet of S1 nuclease. (2) The protected cDNA can be directly cloned and sequenced to confirm its identity, whereas in RNA protection assay, the protected RNA still needs to be converted to cDNA if a long fragment needs to be sequenced at a high quality. (3) It is still technically difficult to determine from which DNA strand an RNA is transcribed. Use of strand-specific DNA oligos to supersede cDNA in our protection assay may be the best way for this purpose, as discussed later. (4) DNA/RNA hybrid has its unique structure and compositions that are distinguishable from DNA/DNA or RNA/RNA hybrid,⁴⁰ in part because DNA/DNA contains dA and dT, RNA/RNA contains rA and rU, while DNA/RNA contains all four. These differences should provide us with unique strategies to develop sensitive methods and instruments for the detection and quantification of those DNA/RNA hybrids that are at very low abundance. Such strategies should be applicable and, thus, intriguing, as endogenous DNA/RNA hybrids in eukaryotic cells are many fewer than the DNA/DNA and RNA/RNA hybrids, especially when a larger DNA/RNA fragment is designed for protection.

In most assays the probe is used at great excess compared with the target. We suggest that if our method is used mainly to verify the true existence of an RNA transcript, the RNA sample should be considered as the probe and, thus, used in great excess, relative to the cDNA. Conversely, if the aim is to quantify the RNA expression level, the cDNA should be regarded as the probe and used in great excess. A set of nested PCR, including those with one primer in the S1-digested region, should help in authenticating the RNA and, thus, is highly recommended, especially when T-A cloning and sequencing the resultant plasmids are omitted due to whatever considerations.

Unvanquished obstacles set by ERP in RT. Although retrovirus uses cellular tRNA to prime mRNA for reverse transcriptases to synthesize the 1st DNA strand,^{31,32} endogenous small RNAs such as mRNA fragments can efficiently prime cDNA synthesis by reverse transcriptases.^{12,20,22} RNA samples contain a huge number of short RNA fragments, such as degraded RNAs, excised introns and other processed mRNAs that are known to us recently,¹ which can serve as ERP for RT. This is likely the reason why RT can occur without addition of primers, a phenomenon coined by others as "background priming."^{2,15} This also explains why DNase treatment of RNA samples cannot eliminate cDNA generation in the ensuing RT. Actually, during DNase treatment and inactivation, some RNAs are likely degraded to be ERP. Besides, short genomic and mitochondrial DNA fragments resulting from degradation or incomplete DNase digestion can also serve as ERP.

The presence of ERP should not affect the RT results from random hexamers, and may not affect the results from poly-dT primer either if polyadenylation is not a specific concern. However, the results from GSP may no longer be gene- and strand-specific, not only because GSP may mis-prime, which is familiar to us, but also because ERP can prime other RNAs, including the antisense of the interested RNA if it exists. The gene-specificity may be improved by adding a linker sequence, herein NewD, to the 5' end of the GSP and using it as one primer in the ensuing PCR. The strand-specificity may also be improved in this way if the antisense RNA level is relatively low and the amount of Linker-GSP is carefully managed, as shown in lanes 5 vs. 6 in **Figure 5B** and as depicted in **Figure 6A**. However, this strategy may not always work well. When we tested this strategy by determining the existence of LOC100996515 RNA

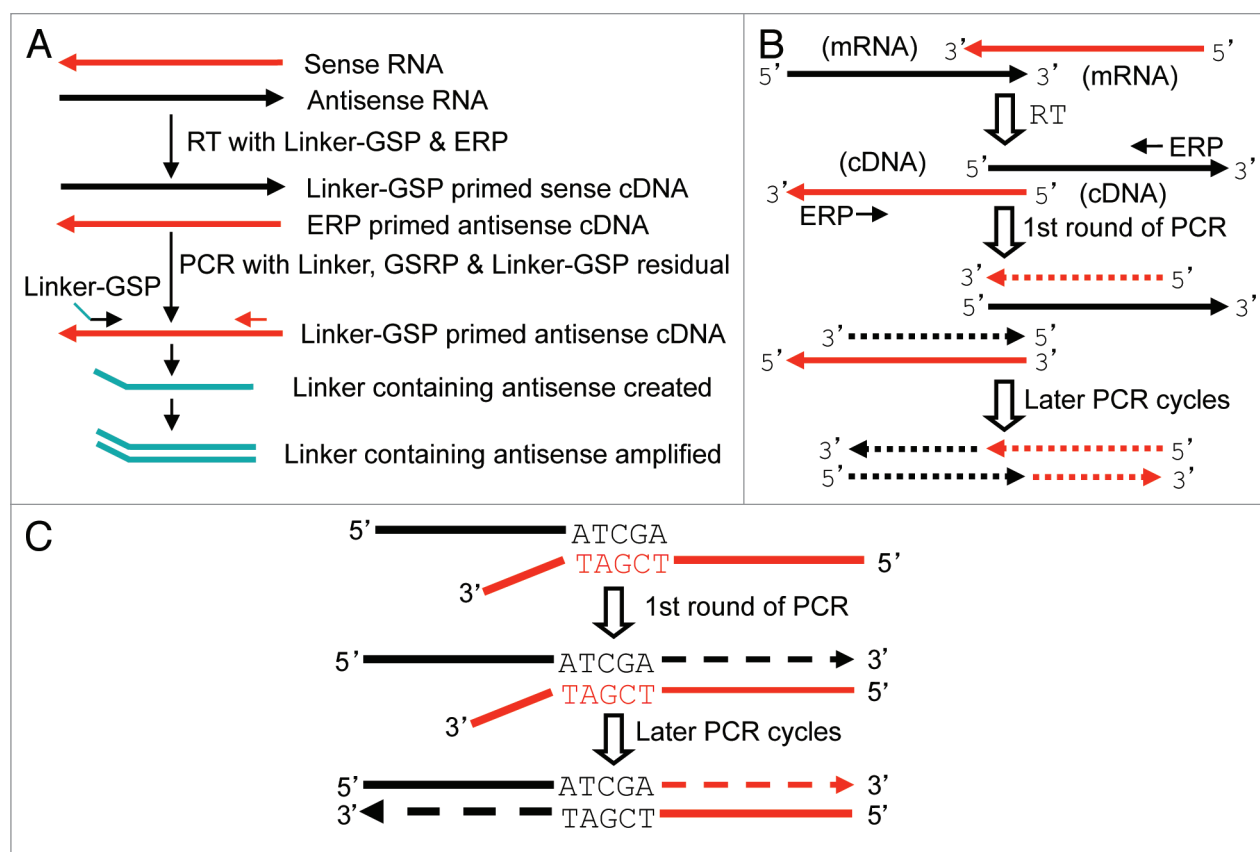


Figure 6. Depiction of artifacts caused by ERP or by 5' or 3' overlapping. **(A)** Because the RNA sample contains ERP, RT with Linker-GSP will also generate the first cDNA strand of the antisense RNA, besides the cDNA of the desired sense RNA. When the RT product (usually only 1 μ l) is added into the PCR mixture as the template, some Linker-GSP residual is transferred together, which primes the synthesis of a linker containing antisense fragment. The fragment is amplified in the later PCR cycles. However, because the PCR mixture contains many more copies of the GSP and the gene-specific reverse primer (GSRP) than the Linker-GSP residual, the first PCR cycle should generate many more copies of the desired sense cDNA, which titrates out the antisense in later PCR cycles, unless the antisense RNA is expressed at a much higher level than the sense. We use as small an amount of linker-GSP as possible in the RT to minimize its residual in the RT product. **(B)** Two cDNAs that overlap at their 5' ends, no matter whether they are unrelated or are originated respectively from a sense and an antisense transcripts that overlap at their 3' ends (like CDK4 and TSPAN31), can be converted to 3'-overlapped counterparts after one round of PCR by GSP or ERP, which, in turn, creates wrong-template extension in later PCR cycles as depicted in **Figure 5C**. **(C)** If a cDNA has an unprotected 3' end that is reverse-complementary to an unrelated cDNA (in red color), this matched part (e.g., ATCGA/TAGCT) and this other cDNA may serve in PCR as the primer and the template, respectively, to create a spurious chimera.

with primers at the region overlapped by CCND1, CCND1 as its antisense was often detected because of its much higher abundance. A better way to ensure the strand-specificity may be to use strand-specific, probably labeled, DNA oligos to replace cDNA probes in hybridization with the RNA of interest. Such strand-specific DNA oligos can be in vitro synthesized like a primer or made by other ways,¹⁰ including PCR with one biotinylated primer followed by capture with streptavidin-coated magnetic beads or PCR with one 5'phosphorylated primer¹⁰ followed by digestion of the useless, 5'phosphorylated strand using lambda exonuclease.^{10,34}

Although GSP has been widely used in RT-PCR for decades, to our knowledge none of the published studies has addressed the possible spuriousness and provided a corrective measure as did we herein. Since routine GSP-primed RT is, likely, neither gene- nor strand-specific, whether those published data need to be reevaluated or reinterpreted becomes an uncomfortable but unavoidable question that peers need to bear in mind, in our humble opinion.

Unsolved, overlap-caused artifacts of cDNA end and chimeras. The NCBI has updated several times the sequences of CDK4 and TSPAN31 and keeps extending the overlapped region. One of the CDK4 sequences we obtained is longer than the latest NCBI version. We surmise that all old and new sequences may be correct, representing different variants with different lengths of the overlap. However, cloning the 3' end of each of these mRNAs is technically difficult due to two major reasons: (1) ERP will result in cDNA of the antisense in RT (**Fig. 6A**). (2) One of the mRNA molecules, either CDK4 or TSPAN31, may have an unprotected 3' end at the overlapped region, due to reasons such as degradation (breakage), premature transcription, deadenylation, early termination of RT, etc., occurring either as a physiological event or as an artifact. This unprotected 3' end will serve as a primer to extend its cDNA with an antisense RNA as the template in RT, creating a wrong-template artifact (**Fig. 5C**). In this situation, the 3' end cloned by any RT-PCR involved approach, including our method, could be an artifact. This artifact may also occur in

PCR, if not already in RT, because one round of PCR, primed by either GSP or ERP, will convert two 5'-overlapped cDNAs to two 3'-overlapped ones (Fig. 6B). This pitfall should be particularly alerted to the biomedical society because so often RT-PCR is used to clone RNA without preclusion of the existence of overlapped antisense RNA. So far we are still unable to get out from this trap and, thus, unable to clone the genuine 3'-end of the extended CDK4 mRNA shown in lane 6 of Figure 5B, and to determine whether TSPAN31 also has mRNA variant(s) extended beyond the latest NCBI version.

Our results also alert us to another pitfall that if routine or quantitative RT-PCR is used to determine the expression level of an RNA that is accompanied by an overlapped antisense transcript, PCR with primers at the overlapped region starts with two templates and, thus, will falsely double the expression level. Therefore, the locations of the primers matter, and primers at different regions of the RNA should be used. Since over 63% of RNA transcripts may be accompanied by antisense counterparts,²⁴ peers should be alerted to the pitfalls described above.

A huge number of putative chimeric RNAs encompass a short homologous sequence shared by the two partners.²⁷ The reason is unknown but it has led to discussions on how such chimeras are formed.^{3,27,39} Our observation of wrong-template extension created by overlap at the 5' or 3' end enlightens us in that some, likely many, of this type of chimeras may simply be RT-PCR spuriousness: If, as described above, an RNA or a cDNA has an unprotected 3' end that is reversely complementary to an unrelated (i.e., not its antisense) RNA or cDNA, a chimeric sequence may be generated in RT or PCR (Fig. 6C). Since a pentamer can prime RT or PCR efficiently, the homologous part can be a 5-nt sequence, although whether a shorter oligo still has some priming ability is not so clear. Because the RNA repository in any human cell contains numerous such short homologous sequences, we tend to believe that many of those chimeras containing a short homologous sequence and obtained by approaches that involve RT, PCR or similar methods are such technical artifacts.

Materials and Methods

RNA preparation, DNase I treatment and RT. Total RNA was extracted from indicated cell lines using TRIzol (Invitrogen, Cat. 15596-026). In some experiments, RNA was treated with DNase I (1–3 units) to remove genomic DNA residuals, followed by inactivation of the DNase with 15 mM EDTA at 72°C for 15 min. An aliquot (4–5 µg) of total RNA was reverse-transcribed to the first strand of cDNA with indicated primers and M-MLV Reverse Transcriptase (Promega, Cat #. M1705; www.promega.com), following the manufacturer's instruction, but in a 20–25 µl volume.

Primer nomenclature. We used “F” and “R” to indicate a forward and a reverse primer, respectively. Each primer's name ends with a number that indicates the first (for F) or the last (for R) nt of that primer in the position, i.e., the distance from the first nt, of the mRNA. Thus, the F-to-R range is the size of an RT-PCR amplified DNA fragment in agarose gel. All primers are

listed in Table 1. More details of the primer design principle were described before.⁴¹

Purification of DNA and T-A cloning. PCR-amplified cDNA fragment was fractionated in 1% agarose gel and visualized by ethidium bromide staining. The desired band was then excised out and purified with UltraClean Gel DNA Extraction Kit (ISC BioExpress; www.bioexpress.com) following the manual, or with a simple method we described before.⁴² The purified DNA was ligated into a pGEM-T Easy Vector (Promega; www.promega.com).

RNA 3' end cloning. RT was performed using RNA from HeLa cells and primed by random hexamers, with other conditions as described above. The 3' part of the second strand of CDK4 cDNA was synthesized using 1/3 to 1/2 of the RT products, 100 nM CDK4 F665 primer and 1x PCR Mastermix, with one cycle of 95°C for 5 min, 60°C for 2 min and 72°C for 15 min in a thermocycler. Ten or 15 units of S1 nuclease (Cat # 18001-016; www.invitrogen.com) was added, followed by incubation at room temperature for 60 min to digest the 3' overhang of the first cDNA strand and all single stranded cDNAs or mRNAs. EDTA was added to a final concentration of 10–15 mM with incubation at 72°C for 15 min to inactivate S1. To remove EDTA, the reaction was transferred to an Eppendorf tube with additions of 0.35 ml H₂O and 1.2 ml 95% ethanol, followed by precipitation at –20°C for 20 min and then centrifugation at 13,000 rpm at 4°C for 15 min. The ethanol was discarded and the cDNA pellet was suspended with 14 µl H₂O. To ensure that the 3' overhang of the first cDNA strand had been removed by S1 but the double-stranded fragment was protected, 2 µl of the recovered double-stranded cDNA was used as the template to run 40 cycles of PCR with the F136+R1086 or the F665+R1086 primer pair (Fig. 1). The remaining (10 µl) double-stranded cDNA was then added with 10 µl PCR Mastermix, followed by incubation at 72°C for 10 min to append a dA at the cDNA blunt ends. A portion (herein 6 µl) of the dA-appended cDNA was cloned into a T-A vector. The resultant plasmid clones were first confirmed by PCR with the F665+R1086 primers and then sequenced with a vector primer.

RNA 5' cloning with G-tailing. RT was performed using RNA from HeLa cells as above-described, but with HPRT1R683 as a gene specific reverse primer. After being run through RapidTip2 (cat # RT050-096; www.midsci.com) to remove primers, enzymes, dNTP and debris, an aliquot (1/4) of the RT product was transferred into a 500-µl tube with additions of 30 units of TdT (www.promega.com; Cat# M828C), two units of RNase H, 4 mM dGTP, and 2 mM MnCl₂ in a final volume of 25 µl, followed by incubation at 37°C for 30–60 min to synthesize a poly-dG tail. The TdT was inactivated by heating to 72°C for 10 min. About half of the dG-tailed product was primed by a NewB mixture (Table 1), with 1x PCR Mastermix in a 20-µl volume, to synthesize the second cDNA strand by one cycle of 95°C for 5 min, 60°C for 2 min and 72°C for 15 min in a thermocycler. About 1/4 of the double-stranded HPRT1 cDNA was then used as the template to run PCR with the NewD+HPRT1R683 primer pair for 40 cycles of 95°C for 30 sec, 60°C for 30 sec and 72°C for 60 sec. The PCR product appeared as a fuzzy band in

agarose gel and, thus, was excised out and purified as the template for a second round of PCR, followed by excision and purification of the dominant band for T-A cloning (Fig. 2).

cDNA protection assay. RT was performed using RNA from MCF7 cells in a 25- μ l volume and primed with random hexamers. The RT product was incubated at 72°C for 15 min with about 10–15 mM EDTA to inactivate RNase H and DNA polymerase activities of the MMLV. To remove EDTA, the RT product was transferred to an Eppendorf tube with additions of 0.35 ml H₂O and 1.2 ml 95% ethanol, followed by precipitation of the cDNA as described above. The cDNA was suspended in 20 μ l of H₂O. In a 500- μ l tube, the hybridization was set up with 1/10–1/5 (2–4 μ l) of the cDNA and an equivalent amount of the RNA sample in a 50- μ l solution containing 25% formamide (v/v), 600 mM NaCl, 30 mM Tris-HC (pH 7.5), 0.1% SDS, 10 mM DTT and 4 mM EDTA, as described before.²⁸ Moreover, 1 μ l (1/20) of the cDNA was also used in PCR to amplify the CCND1 cDNA with the F70+R1067 primer pair, and the PCR product was purified from agarose gel. About half of the purified CCND1 PCR product was added into the hybridization reaction as an indicator of whether the hybridization and the ensuing S1 digestion degrade double-stranded DNA. After topping with 35 μ l mineral oil (purchased from a Walmart store; product #831432DB1) to prevent evaporation, the hybridization reaction was performed at 68°C for 8 h or longer. After transfer to an Eppendorf tube, the reaction was diluted and precipitated with additions of 0.35 ml H₂O and 1.2 ml 95% ethanol as described above. The hybrids were suspended in 18 μ l H₂O and divided to three aliquots for digestion with 0, 10 or 15 units of S1 in a final volume of 20 μ l at room temperature for 60 min. As a negative control (Fig. 2), a separate S1 digestion was set up with equal amounts of cDNA (the RT product) and S1. The S1 was then inactivated with 10–15 mM EDTA at 72°C for 15 min. The EDTA was removed by dilution and precipitation with additions of 0.35 ml H₂O and 1.2 ml 95% ethanol at –20°C as described above. The cDNA/RNA hybrids were suspended in 16 μ l H₂O, 3 μ l of which was used to run PCR with the BCAS4E1F+hBCAS3E25R primers and the CCND1F183+R1067 primers to ensure that the BCAS4-BCAS3 and the CCND1 cDNAs had been protected. The BCAS4-BCAS3 band was then purified from agarose gel and cloned into a T-A vector for sequencing verification.

References

1. Affymetrix ENCODE Transcriptome Project; Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* 2009; 457:1028-32; PMID:19169241; <http://dx.doi.org/10.1038/nature07759>.
2. Adrover MF, Muñoz MJ, Baez MV, Thomas J, Kornblihtt AR, Epstein AL, et al. Characterization of specific cDNA background synthesis introduced by reverse transcription in RT-PCR assays. *Biochimie* 2010; 92:1839-46; PMID:20709138; <http://dx.doi.org/10.1016/j.biochi.2010.07.019>.
3. Al-Balool HH, Weber D, Liu Y, Wade M, Guleria K, Nam PL, et al. Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant. *Genome Res* 2011; 21:1788-99; PMID:21948523; <http://dx.doi.org/10.1101/gr.116442.110>.

4. Alldred MJ, Che S, Ginsberg SD. Terminal continuation (TC) RNA amplification without second strand synthesis. *J Neurosci Methods* 2009; 177:381-5; PMID:19026688; <http://dx.doi.org/10.1016/j.jneumeth.2008.10.027>.
5. Bärklund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, et al. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosomes Cancer* 2002; 35:311-7; PMID:12378525; <http://dx.doi.org/10.1002/gcc.10121>.
6. Beiter T, Reich E, Williams RW, Simon P. Antisense transcription: a critical look in both directions. *Cell Mol Life Sci* 2009; 66:94-112; PMID:18791843; <http://dx.doi.org/10.1007/s00018-008-8381-y>.

Summary

We describe two new methods for cloning cDNA ends and a cDNA protection assay to supersede RNA protection assay. We also report that GSP-primed RT product is neither gene- nor strand-specific because the RNA sample contains ERP. The gene-specificity may be improved by adding a linker sequence to the GSP and then using the linker as a primer in the ensuing PCR, whereas the strand-specificity may be improved by using strand-specific DNA oligos as the probe in our protection assay. Using the CDK4/TSPAN31 relationship as a model, we find that when sense and antisense RNAs overlap at their 3' ends, the overlapped sequence might serve as a primer with its antisense as the template to create a wrong-template extension in RT or PCR, resulting in a spurious 3' end. This result edifies us that two unrelated RNAs or cDNAs that overlap at the 5'- or 3'-end may also create a chimeric sequence in this way. Therefore, many chimeric RNAs containing a short homologous sequence and obtained by approaches involving RT or PCR may be such artifacts and, thus, need to be vigorously verified with, such as, our protection assay. The ERP and the 5'- or 3'-overlapping antisense together set more complex pitfalls in our way of RNA cloning, which should be highly alerted to the peers. Our methods cannot fully circumvent these traps but should be good alternative or corrective measures to the available ones for cloning chimeric or antisense-accompanied RNA, both together constituting the majority of the cellular RNA repository.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We want to thank Fred Bogott, MD, Ph.D., at Austin Medical Center, for his excellent English editing of this manuscript. This work was supported by a grant from the Department of Defense of United States (DOD Award W81XWH-11-1-0119) to D.J.L. The funding agency had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

7. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; 447:799-816; PMID:17571346; <http://dx.doi.org/10.1038/nature05874>.
8. Brakenhoff RH, Schoenmakers JG, Lubsen NH. Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res* 1991; 19:1949; PMID:2030976; <http://dx.doi.org/10.1093/nar/19.8.1949>.

9. Celestino R, Sigstad E, Løvf M, Thomassen GO, Grøholt KK, Jørgensen LH, et al. Survey of 548 oncogenic fusion transcripts in thyroid tumors supports the importance of the already established thyroid fusions genes. *Genes Chromosomes Cancer* 2012; 51:1154-64; PMID:22961909; <http://dx.doi.org/10.1002/gcc.22003>.
10. Civit L, Fragoso A, O'Sullivan CK. Evaluation of techniques for generation of single-stranded DNA for quantitative detection. *Anal Biochem* 2012; 431:132-8; PMID:22995064; <http://dx.doi.org/10.1016/j.ab.2012.09.003>.
11. Cocquet J, Chong A, Zhang G, Veitia RA. Reverse transcription template switching and false alternative transcripts. *Genomics* 2006; 88:127-31; PMID:16457984; <http://dx.doi.org/10.1016/j.ygeno.2005.12.013>.
12. Colett MS, Larson R, Gold C, Strick D, Anderson DK, Purchio AF. Molecular cloning and nucleotide sequence of the pestivirus bovine viral diarrhea virus. *Virology* 1988; 165:191-9; PMID:2838957; [http://dx.doi.org/10.1016/0042-6822\(88\)90672-1](http://dx.doi.org/10.1016/0042-6822(88)90672-1).
13. Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, et al. Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One* 2012; 7:e28213; PMID:2238572; <http://dx.doi.org/10.1371/journal.pone.0028213>.
14. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29:15-21; PMID:23104886; <http://dx.doi.org/10.1093/bioinformatics/bts635>.
15. Frech B, Peterhans E. RT-PCR: 'background priming' during reverse transcription. *Nucleic Acids Res* 1994; 22:4342-3; PMID:7524039; <http://dx.doi.org/10.1093/nar/22.20.4342>.
16. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, et al. Chimeras taking shape: potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* 2012; 22:1231-42; PMID:22588898; <http://dx.doi.org/10.1101/gr.130062.111>.
17. Gingeras TR. Implications of chimeric non-co-linear transcripts. *Nature* 2009; 461:206-11; PMID:19741701; <http://dx.doi.org/10.1038/nature08452>.
18. Ginsberg SD, Che S. RNA amplification in brain tissues. *Neurochem Res* 2002; 27:981-92; PMID:12462399; <http://dx.doi.org/10.1023/A:1020944502581>.
19. Grinchuk OV, Jenjaroenpun P, Orlov YL, Zhou J, Kuznetsov VA. Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns. *Nucleic Acids Res* 2010; 38:534-47; PMID:19906709; <http://dx.doi.org/10.1093/nar/gkp954>.
20. Gubler U. Second-strand cDNA synthesis: mRNA fragments as primers. *Methods Enzymol* 1987; 152:330-5; PMID:3309563; [http://dx.doi.org/10.1016/0076-6879\(87\)52038-9](http://dx.doi.org/10.1016/0076-6879(87)52038-9).
21. Hahn Y, Bera TK, Gehlhaus K, Kirsch IR, Pastan IH, Lee B. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. *Proc Natl Acad Sci USA* 2004; 101:13257-61; PMID:15326299; <http://dx.doi.org/10.1073/pnas.0405490101>.
22. Herzog E, Voronin N, Hizi A. The removal of RNA primers from DNA synthesized by the reverse transcriptase of the retrotransposon Tf1 is stimulated by Tf1 integrase. *J Virol* 2012; 86:6222-30; PMID:22491446; <http://dx.doi.org/10.1128/JVI.00009-12>.
23. Houseley J, Tollervey D. Apparent non-canonical splicing is generated by reverse transcriptase in vitro. *PLoS One* 2010; 5:e12271; PMID:20805885; <http://dx.doi.org/10.1371/journal.pone.0012271>.
24. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, et al.; RIKEN Genome Exploration Research Group; Genome Science Group (Genome Network Project Core Group); FANTOM Consortium. Antisense transcription in the mammalian transcriptome. *Science* 2005; 309:1564-6; PMID:16141073; <http://dx.doi.org/10.1126/science.1112009>.
25. Kowalczyk MS, Higgs DR, Gingeras TR. Molecular biology: RNA discrimination. *Nature* 2012; 482:310-1; PMID:22337043; <http://dx.doi.org/10.1038/482310a>.
26. Li K, Ramchandran R. Natural antisense transcript: a concomitant engagement with protein-coding transcript. *Oncotarget* 2010; 1:447-52; PMID:21311100.
27. Li X, Zhao L, Jiang H, Wang W. Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J Mol Evol* 2009; 68:56-65; PMID:19089307; <http://dx.doi.org/10.1007/s00239-008-9187-0>.
28. Liao D, Porsch-Hällström I, Gustafsson JA, Blanck A. Sex differences at the initiation stage of rat liver carcinogenesis--influence of growth hormone. *Carcinogenesis* 1993; 14:2045-9; PMID:8222052; <http://dx.doi.org/10.1093/carcin/14.10.2045>.
29. Løvf M, Thomassen GO, Bakken AC, Celestino R, Fioretos T, Lind GE, et al. Fusion gene microarray reveals cancer type-specificity among fusion genes. *Genes Chromosomes Cancer* 2011; 50:348-57; PMID:21305644; <http://dx.doi.org/10.1002/gcc.20860>.
30. Mader RM, Schmidt WM, Sedivy R, Rizovski B, Braun J, Kalipciyan M, et al. Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J Lab Clin Med* 2001; 137:422-8; PMID:11385363; <http://dx.doi.org/10.1067/mlc.2001.115452>.
31. Mak J, Kleiman L. Primer tRNAs for reverse transcription. *J Virol* 1997; 71:8087-95; PMID:9343157.
32. Marquet R, Isel C, Ehresmann C, Ehresmann B. tRNAs as primer of reverse transcriptases. *Biochimie* 1995; 77:113-24; PMID:7541250; [http://dx.doi.org/10.1016/0300-9084\(96\)88114-4](http://dx.doi.org/10.1016/0300-9084(96)88114-4).
33. McManus CJ, Duff MO, Eipper-Mains J, Graveley BR. Global analysis of trans-splicing in *Drosophila*. *Proc Natl Acad Sci USA* 2010; 107:12975-9; PMID:20615941; <http://dx.doi.org/10.1073/pnas.1007586107>.
34. Null AP, Hannis JC, Muddiman DC. Preparation of single-stranded PCR products for electrospray ionization mass spectrometry using the DNA repair enzyme lambda exonuclease. *Analyst* 2000; 125:619-26; PMID:10892018; <http://dx.doi.org/10.1039/a908022h>.
35. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011; 12:87-98; PMID:21191423; <http://dx.doi.org/10.1038/nrg2934>.
36. Pääbo S, Irwin DM, Wilson AC. DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem* 1990; 265:4718-21; PMID:2307682.
37. Ponting CP, Belgard TG. Transcribed dark matter: meaning or myth? *Hum Mol Genet* 2010; 19(R2):R162-8; PMID:20798109; <http://dx.doi.org/10.1093/hmg/ddq362>.
38. Qiu X, Wu L, Huang H, McDonel PE, Palumbo AV, Tiedje JM, et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl Environ Microbiol* 2001; 67:880-7; PMID:11157258; <http://dx.doi.org/10.1128/AEM.67.2.880-887.2001>.
39. Ritz K, van Schaik BD, Jakobs ME, Aronica E, Tijssen MA, van Kampen AH, et al. Looking ultra deep: short identical sequences and transcriptional slippage. *Genomics* 2011; 98:90-5; PMID:21624457; <http://dx.doi.org/10.1016/j.ygeno.2011.05.005>.
40. Shaw NN, Arya DP. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* 2008; 90:1026-39; PMID:18486626; <http://dx.doi.org/10.1016/j.biochi.2008.04.011>.
41. Sun Y, Li Y, Luo D, Liao DJ. Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. *PLoS One* 2012; 7:e41659; PMID:22927912; <http://dx.doi.org/10.1371/journal.pone.0041659>.
42. Sun Y, Sriramajayam K, Luo D, Liao DJ. A quick, cost-free method of purification of DNA fragments from agarose gel. *J Cancer* 2012; 3:93-5; PMID:22359530; <http://dx.doi.org/10.7150/jca.4163>.
43. Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, et al. Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. *Nature* 2008; 454:126-30; PMID:18509338; <http://dx.doi.org/10.1038/nature06992>.

Pseudogenes as Weaknesses of ACTB (Actb) and GAPDH (Gapdh) Used as Reference Genes in Reverse Transcription and Polymerase Chain Reactions

Yuan Sun^{1,2}, Yan Li¹, Dianzhong Luo^{2*}, D. Joshua Liao^{1*}

¹ Hormel Institute, University of Minnesota, Austin, Minnesota, United States of America, ² Department of Pathology, Guangxi Medical University, Nanning, Guangxi, People's Republic of China

Abstract

The genes encoding β -actin (ACTB in human or Actb in mouse) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH in human or Gapdh in mouse) are the two most commonly used references for sample normalization in determination of the mRNA level of interested genes by reverse transcription (RT) and ensuing polymerase chain reactions (PCR). In this study, bioinformatic analyses revealed that the ACTB, Actb, GAPDH and Gapdh had 64, 69, 67 and 197 pseudogenes (PGs), respectively, in the corresponding genome. Most of these PGs are intronless and similar in size to the authentic mRNA. Alignment of several PGs of these genes with the corresponding mRNA reveals that they are highly homologous. In contrast, the hypoxanthine phosphoribosyltransferase-1 gene (HPRT1 in human or Hprt in mouse) only had 3 or 1 PG, respectively, and the mRNA has unique regions for primer design. PCR with cDNA or genomic DNA (gDNA) as templates revealed that our HPRT1, Hprt and GAPDH primers were specific, whereas our ACTB and Actb primers were not specific enough both vertically (within the cDNA) and horizontally (compared cDNA with gDNA). No primers could be designed for the Gapdh that would not mis-prime PGs. Since most of the genome is transcribed, we suggest to peers to forgo ACTB (Actb) and GAPDH (Gapdh) as references in RT-PCR and, if there is no surrogate, to use our primers with extra caution. We also propose a standard operation procedure in which design of primers for RT-PCR starts from avoiding mis-priming PGs and all primers need be tested for specificity with both cDNA and gDNA.

Citation: Sun Y, Li Y, Luo D, Liao DJ (2012) Pseudogenes as Weaknesses of ACTB (Actb) and GAPDH (Gapdh) Used as Reference Genes in Reverse Transcription and Polymerase Chain Reactions. PLoS ONE 7(8): e41659. doi:10.1371/journal.pone.0041659

Editor: Arun Rishi, Wayne State University, United States of America

Received: April 13, 2012; **Accepted:** June 25, 2012; **Published:** August 22, 2012

Copyright: © 2012 Sun et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by a grant from the United States Department of Defense (DOD Award W81XWH-11-1-0119) to DJL. The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: luodianzhong@yahoo.com.cn (DZL); djliao@hi.umn.edu (DJL)

Introduction

Determination of mRNA level of an interested gene in eukaryotic cells often involves conversion of the mRNA to cDNA by reverse transcription (RT), followed by polymerase chain reactions (PCR). This RT-PCR approach is much more sensitive than other methods such as Northern blot, because PCR amplifies the cDNA in an exponential manner. So often, the RT-PCR products need to be compared between two, or among more, samples to determine whether some sample(s) have a different mRNA level of the interested gene from the others [1]. In this case, a reference gene is needed for sample normalization, i.e. for assessing that an equal amount of the RT products from all the samples is used as template in the PCR. Because RT-PCR is so sensitive that it can detect the mRNA level even in a single cell, variation in the expression level of the reference gene needs to be tightly controlled, otherwise a bias may be produced. Ideally, expression of the reference gene should be constant in all situations and should be refractory to all the changes in the experimental conditions. At least, its expression should not be changed by the to-be-studied situation.

There has been a long list of genes that have been used as references in RT-PCR [2–6], of which the genes encoding β -actin (ACTB in human or Actb in mouse, according to the NCBI

nomenclature of genes) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH in human and Gapdh in mouse) are the two most frequently used ones [1]. The hypoxanthine phosphoribosyltransferase-1 (HPRT1 in human and Hprt in mouse) is also used often [5,7,8]. The reason for having so many reference genes is because none of them really meets the above mentioned ideal criteria, and therefore researchers have to select different ones according to their to-be-studied situations [9–11]. The weaknesses of most reference genes have been discussed in the literature and pertain mainly to the stability or variation of their expression in different situations [12–14], with only very few concerning the influence of pseudogene (PG) in their fidelity as references [15–17]. ACTB, Actb, GAPDH or Gapdh had been used as references in RT-PCR long before the human and mouse genomes were fully sequenced. Although individual PGs of these genes in the human and mouse were reported a long time ago, these genes continue serving as references, because some peers do not realize that these genes have PGs while many others consider that most PGs are not transcribed, which actually has been a generally accepted concept for a long time.

According to a report from the human genome project, about 1.1% of the human genomic DNA belongs to exons while 24% belongs to introns, together about a quarter of the genome owned

by the protein coding genes [18]. However, the majority of the remaining three quarters is not junk and is actually also transcribed at least in some cell types or at some times [19]. After sequencing the RNA transcripts from about 1% of the human genome, the ENCODE pilot project reports that 93% of the bases in this 1% of the genome are transcribed [20]. The human genome contains only about 20,000 protein coding genes but about 19,000 PGs, although probably less than 20% of the PGs are commonly transcribed [21]. Like most other non-coding RNAs, many PG transcripts may be functional, such as in regulation of the expression of their parental genes [21–23]. There are even further lessons that in some situations processed PGs are transcriptionally activated but are mistaken as the turn-on of the authentic genes that are actually inactive [24–26]. These facts actually arouse in us a concern as to whether genetic knockout of one gene would change its PG expression or, after some latent period of time, even trigger expression of some of its PGs that are otherwise silent in some tissues or cell types, since reports on gene-knockout animals hardly address this aspect. In short, these latest advances in RNA biology are revolutionary to biomedical science as they not only challenge the definition of “gene” [27] and many other fundamental concepts in our mind but also require us to reevaluate some experimental methodologies. As an example, in this report we provide bioinformatic data showing that ACTB (Actb) and GAPDH (Gapdh) have many PGs in the human and mouse genomes, which may affect the fidelity of these genes as references for RT-PCR.

Materials and Methods

In the database of National Center for Biotechnology Information (NCBI) of the United States, the mRNA sequence of all genes is presented as DNA sequence, i.e. uracil (U) is replaced by thymine (T). According to the nomenclature of the NCBI, the name of human genes should be fully capitalized whereas the name of mouse genes should be capitalized only the first letter. We pulled out the mRNA sequences of ACTB, Actb, GAPDH, Gapdh, HPRT1 and Hprt from the NCBI database; the gene identification (gi) number and mRNA access number were provided in Fig. S1, prior to the corresponding sequence. PGs were identified using online software and databases as indicated. An online software (<http://biotools.umassmed.edu/bioapps/primer3www.cgi>) was used for primer design. Insilico PCR was performed with two different online software packages (<http://insilico.ehu.es/PCR/> and <http://genome.csdb.cn/cgi-bin/hgPcr>). The cell lines from which data were presented include GI101A human breast cancer cell line as well as Panc-1, Panc-28, CooLo357 and L3.5pL human pancreatic cancer cell lines; all these cell lines are well documented in the literature. E6E7st non-transformed and E6E7st/ras transformed human pancreatic ductal epithelial cell lines were provided by Dr. Paul Campbell [28]. M8 mouse pancreatic [29] and ND5 mouse breast [30] cancer cell lines were established by us previously. All cell lines were cultured with DMEM containing 5% bovine serum and were harvested when the cells reached about 70% confluence. Isolation of total RNA from the cells was performed using Trizol (Invitrogen, Cat. 15596-026; www.invitrogen.com), following the manual. Genomic DNA (gDNA) was isolated with the traditional phenol-chloroform method. The gDNA samples were treated with RNase A whereas the RNA samples were treated with DNase I, both followed by extraction with phenol and chloroform to remove the enzyme. The DNA or RNA samples were then precipitated and washed with ethanol at a final concentration of 70%.

An aliquot of RNA from each cell line was reverse-transcribed to cDNA with random hexamers and M-MLV Reverse Transcriptase (Promega, Cat. M1705; www.promega.com), following the manual. Forty cycles of PCR were performed to ensure that the reactions entered into the plateau of the amplification of the authentic cDNA and that possible PGs were detectable. PCR products were separated in 1% agarose gel, visualized with ethidium bromide staining, and photographed with Kodak Digital Campture DC290 Camera under a UV light.

Results

Identification of PGs of the ACTB, Actb, GAPDH, Gapdh, HPRT1 and Hprt

The mRNA (actually shown as DNA) sequences of the ACTB, Actb, GAPDH, Gapdh, HPRT1 and Hprt, with their gene identity and mRNA access numbers, are shown in figure S1. The GAPDH has an mRNA variant (NM_001256799.1) that is transcribed from an alternative initiation site and thus differs from the wild type mRNA (NM_002046.4) at the 5'-part (Fig. S1). Like many RNA transcripts [19], Actb and Hprt mRNAs contain only poly-A signal but lack a long poly-A tail (Fig. S1), which is a reason for us to perform RT with random hexamers. We used these mRNA sequences, after deleting the poly-A tail from those having it, as a bait to fish out their PGs from the corresponding (human or mouse) genome in the UCSC Genome Browser Database (<http://genome.ucsc.edu/>) by performing Blat search [31]. The UCSC Genome Browser scores similarity according to not only sequence identity but also sequence length, gap, etc, with a higher score indicating a generally better similarity. The results identified 64, 69, 67, 197, 3 and 1 PGs for the ACTB, Actb, GAPDH, Gapdh, HPRT1 and Hprt, respectively (table 1), which score over 200 and have over 80% identity to the bait. Those genomic sequences that score less than 200 are not counted in, although they still have over 83% identity to the bait and span several hundred nucleotides (nt) on the corresponding chromosome. The details of these PGs, such as their chromosomal locations, sizes, starting and ending nt, homologues, etc, are shown in figures 1 and 2 as well as figures S2, S3, S4 and S5. Most of these PGs are processed, i.e. intronless, as they are similar in size to the bait. Use of other tools such as Blast of the NCBI database (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>), search from other sources such as the PG database (<http://www.pseudogene.org/>), or imposition of different criteria for the cutoff may result in different numbers of PGs. For instance, another study identified only 56 PGs of the GAPDH and 166 PGs of the Gapdh [32]. However, the conclusion remains the same that there are many PGs of these genes in the human and mouse genomes.

Table 1. Number of putative pseudogenes.

Gene	Human			Mouse		
	ACTB	GAPDH	HPRT1	Actb	Gapdh	Hprt
Number	64	67	3	69	197	1

Note: Only those putative pseudogenes that score over 200 are counted, with details are presented in S-Fig. 2, 3, 4, and 5.
doi:10.1371/journal.pone.0041659.t001



Figure 1. Identification of HPRT1 PGs for primer design. **Top panel:** Blat search using HPRT1 mRNA sequence (1405-bp long after deletion of the poly-A tail) as the bait pulls out three putative PGs, besides the authentic HPRT1 genomic sequence that spans 40514 nt on the plus strand of X chromosome. The three putative PGs match the 193–1383rd, the 269–1413th, and the 213–1143rd nt regions of the HPRT1 mRNA as illustrated. There are three additional very short fragments, spanning only 28, 35 and 35 bp, respectively, that also match parts of HPRT1 mRNA but are not considered as PGs. **Bottom panel:** We pulled out the sequence of each PG (by clicking “details”) and assembled those homologous parts to construct the “cDNA” of the PGs on chromosomes 11 and 4. Alignment of the three sequences with the HPRT1 mRNA reveals that the HPRT1 mRNA has some unique regions. The forward and reverse primers (table 2) we designed are underlined.

HPRT1 or Hprt primer design starting from discrimination against PGs

The result of Blat search shows that the HPRT1 mRNA, which is a 1405-bp sequence after its poly-A tail is deleted, spans over 40514 nt in the plus strand of the X chromosome (Fig. 1, top panel). The 93723936–93732367th nt region of the minus strand of chromosome 11 has an 88.3% identity to the 193–1384th nt region of the HPRT1 mRNA. This putative PG spans 8432 nt and is thus unprocessed, i.e. containing intron(s). Another putative PG at chromosome 5 has an 86.2% identity to the 269–1404th nt region of the HPRT1 mRNA. The third putative PG is at chromosome 4 and is homologous to the 213–1143rd nt region of the HPRT1 mRNA; it spans 1274 nt, longer than 930 bp (213–1143rd), and thus may be unprocessed as well. There are three other genomic fragments that are highly homologous to parts of the HPRT1 mRNA but are too small (spanning only 28, 35 and 42 nt, respectively) to be considered as PGs (Fig. 1, top panel).

During Blat search, we clicked the “details” of each PG shown in figure 1 to display the whole sequence. For the unprocessed PGs on chromosomes 11 and 4, we assembled together the parts that are homologous to the HPRT1 mRNA to construct the putative “cDNA”. Alignment of the HPRT1 mRNA with the original sequence or the assembled sequences of the three PGs revealed that the 1–192nd nt and the 586–710th nt regions of the HPRT1 mRNA are lacking in the three PGs (Fig. 1, bottom panel). We designed a forward primer at the 123–142nd nt and a reverse primer at the 664–683rd nt regions (table 2). The HPRT1 mRNA

also has other unique regions that may be used for primer design as well (Fig. 1, bottom panel).

Use of the Hprt mRNA sequence as a bait to fish in the mouse genome identified one putative PG that locates at chromosome 17 and is homologous to the 240–1248th nt region of the Hprt mRNA (Fig. 2, top panel). We displayed the sequence of this PG by clicking “details” during Blat search and assembled the parts that are homologous to the Hprt mRNA to construct the “cDNA”. Alignment of this “cDNA” with the Hprt mRNA shows that the 1–240th nt, the 270–535th nt, and several other regions of the Hprt mRNA are unique to the Hprt (Fig. 2, bottom panel). We designed a forward primer at the 56–75th nt and a reverse primer at the 482–503rd nt regions (table 2).

Design of ACTB, Actb and GAPDH primers that discriminate against PGs

By a quick glance at the figure S2, one could immediately realize that most PGs of the ACTB are similar in length to the ACTB mRNA and thus are intronless. The ACTB mRNA lacks a unique region, making it difficult to design primers that would not mis-prime the PGs. We thus pulled out the sequences of six best-scored PGs (in the red box in Fig. S2) and aligned them with the bait sequence. The results confirm that the ACTB mRNA does not contain a unique part that is long enough for a primer (Fig. 3). The best regions we could find for forward and reverse primers that might have some discrimination against the six PGs are the 1452–1473rd nt and the 1678–1697th nt regions of the ACTB

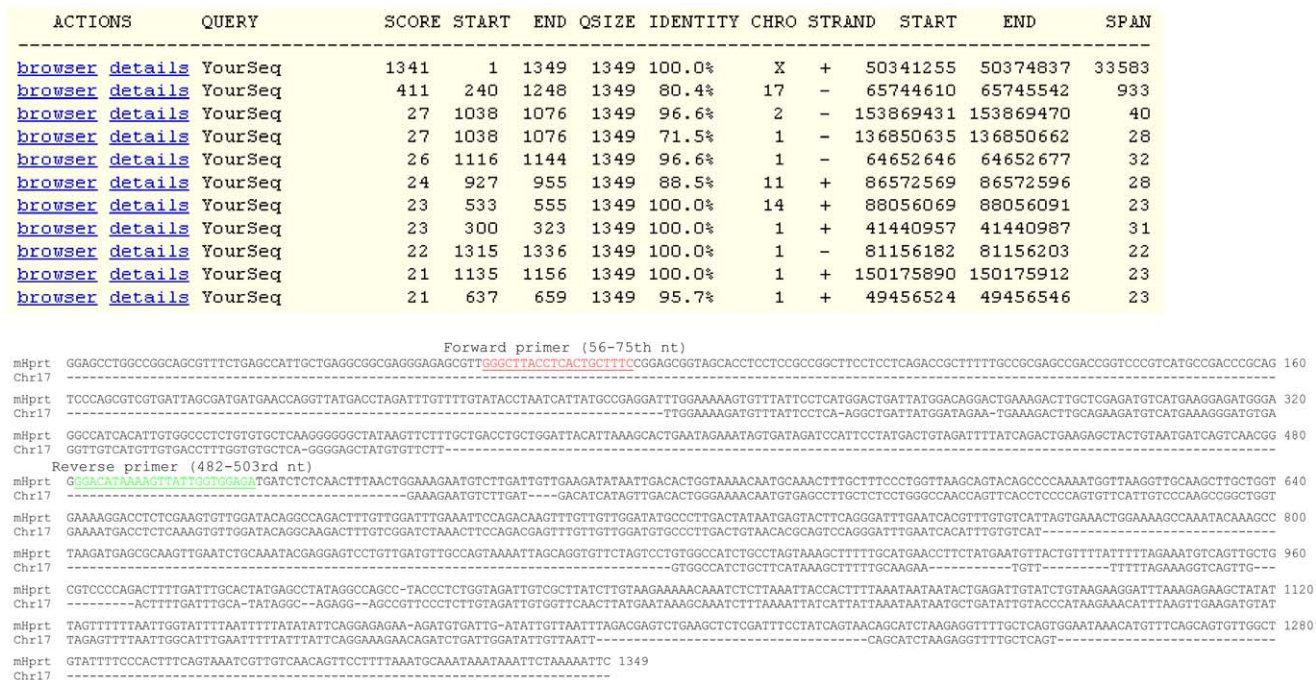


Figure 2. Identification of Hprt PG for primer design. **Top panel:** Blat search using Hprt mRNA sequence as the bait pulls out only one putative PG, besides the authentic Hprt genomic sequence that spans 33583 bp in the plus strand of the mouse X chromosome. This PG matches the 240–1248th nt of the Hprt mRNA and spans 933 nt on the mouse chromosome 17. **Bottom panel:** We pulled out the PG sequence and assembled the parts that are homologous to the Hprt mRNA to construct a cDNA. Alignment of the assembled cDNA with the Hprt mRNA reveals that the Hprt mRNA has several unique regions. The forward and reverse primers (table 2) we designed in some unique regions are underlined. doi:10.1371/journal.pone.0041659.g002

mRNA, respectively (Fig. 3 and table 2). However, since there are a total of 64 putative PGs and only six of them were aligned, it remains possible that these primers may match better some of the other PGs than the six aligned.

Similar to its human counterpart, the mouse Actb mRNA does not have any unique sequence either, relative to its PGs (Fig. S3). We pulled out the sequences of six best-scored PGs (in the red box in Fig. S3) and aligned them with the Actb mRNA. The results confirm that the Actb mRNA has no unique part that is long enough to be a primer (Fig. 4). Nevertheless, we selected the 1471–1489th nt and the 1852–1869th nt regions of the Actb mRNA as forward and reverse primers, respectively (table 2), which might better discriminate against the six PGs than the other parts of the Actb mRNA, although it remains possible that these primers may match better some of the other PGs than the six aligned.

By a quick glance at figure S4, one could immediately find that the first 26 nt of the wild type GAPDH mRNA is a unique region. We pulled out the sequences of seven best-scored PGs (in the red box in Fig. S4) and aligned them with the GAPDH mRNA. The results not only confirm the uniqueness of the 1–26th nt but also show that the 685–705th nt region has the most mismatches to most PGs, except the first X-linked PG that only has one nt mismatched to the GAPDH (Fig. 5). We used these two regions as the forward and reverse primers, respectively (table 2), in part because this pair of primer will not amplify the variant 2 of GAPDH (NM_001256799.1), transcriptional feature of which is unknown.

Impossibility of designing Gapdh specific primers

A quick glance at figure S5 immediately leads us to the fact that there are so many processed PGs of the mouse Gapdh which are

100% or almost 100% identical to the Gapdh mRNA. Indeed, alignment of the Gapdh mRNA with seven best-scored PGs (in the red box in Fig. S5) showed several-nt mismatches only, making it impossible to design any primer that can discriminate against the PGs (Fig. 6).

Verification of the ACTB, GAPDH and HPRT1 primers

PCR results showed that the authentic ACTB band of 246 bp (table 2) in agarose gel was amplified easily, as expected, in the cDNA sample from a panel of human cell lines (Fig. 7A). Although both forward and reverse primers locate at the same exon (exon 6) and thus should also amplify the authentic gene, the same band was detected only weakly (not stronger than nonspecific bands) from the gDNA sample of the same cell lines (Fig. 7A). Thus, this pair of primers meets our purpose to discriminate against gDNA, both the authentic gene and the PGs, as discussed later. However, an additional band that was about 100-bp larger (~350 bp) than the 246-bp ACTB cDNA and another one of about 650 bp were also amplified from the cDNA (but not the gDNA) samples, suggesting that our primers also mis-prime cDNA of two unknown genes, expression of which, as expected, varied among different cell lines, as manifested by different ratios to the 246-bp ACTB band that serves as the internal reference in the same cell line (Fig. 7A). This pair of primers also detected many nonspecific bands from gDNA samples that differed in size from the 246-bp band.

The authentic band of the wild type GAPDH (700-bp) was detected only in the cDNA, but not the gDNA, samples, although the primers also detected cDNA of two unknown genes at smaller (~500 and 400 bp, respectively) sizes and detected several gDNA fragments of different sizes (Fig. 7B). Our HPRT1 primers

Table 2. Primer information.

Primer Name	Sequence	Size	Location	Fragment	Region
hACTB-F1452	5'-TTAATAGTCATCCAAATATGA-3'	22-mers	exon 6	246 bp	1452–1473rd nt
hACTB-R1697	5'-GGGACAAAAAGGGGAAGG-3'	20-mers	exon 6		1678–1697th nt
hGAPDH-F6	5'-GAGCCCGCAGCTCCCGCTT-3'	20-mers	exon 1	700 bp	6–25th nt
hGAPDH-R705	5'-CCCGCGGCCATCAGCCACAG-3'	21-mers	exon 8		685–705th nt
mActb-F1471	5'-GACTTTGTACATTGTTTG-3'	19-mers	exon 6	382 bp	1471–1489th nt
mActb-R1870	5'-TGCACTTTTATTGGTCTCA-3'	19-mers	exon 6		1870–1852nd nt
hHPRT1-F123	5'-CTTCCTCTCTCTGAGCAGTC-3'	20-mers	exon 1	561 bp	123–142nd nt
hHPRT1-R683	5'-AACACTTCGTGGGGTCTTT-3'	20-mers	exon 7		664–683rd nt
mHprt1-F56	5'-GGGCTTACCTCACTGCTTTC-3'	20-mers	exon 1	448 bp	56–75th nt
mHprt1-R503	5'-TCTCCACCAATAACTTTATGTCC-3'	24-mers	exon 4		482–503rd nt

Note: The “h” or “m” in front of each primer’s name indicates the human or mouse origin. “F” or “R” indicates a forward or reverse primer. The number after “F” indicates the position of the first nucleotide (nt) of that primer in the mRNA sequence, whereas the number after “R” indicates the position of the last nt of that primer in the mRNA sequence. Thus, the “R” number minus the “F” number and then plus one is the size of the DNA fragment amplified by PCR.

doi:10.1371/journal.pone.0041659.t002

amplified only the anticipated band (at 561-bp) from cDNA samples, although the primers also amplified several bands of different sizes from gDNA samples (Fig. 7C).

We had obtained similar results from cDNA and gDNA samples of many other human cell lines and tissues by using the same ACTB, GAPDH and HPRT1 primers (data not shown). Based on our experience, the PCR conditions for these primers are recommended as initial denature at 95°C for 5 min, followed by each cycle of melting at 95°C for 30 sec, primer-annealing at 58°C for 30 sec, and elongating at 72°C for 30 sec. To maintain the reaction at the linear portion, the number of cycles should be significantly decreased to less than 30 cycles, unless a very small amount of cDNA template is used. The reaction should be terminated at 72°C for 10 min.

Verification of the Actb and Hprt primers

Our Hprt primers (table 2) amplified only the anticipated band from the cDNA sample, without mis-priming gDNA, from the M8 and ND5 mouse cell lines, thus confirming the specificity of the primers (Fig. 7D and 7E). In contrast, our Actb primers (table 2) produced not only the authentic Actb cDNA band of 382-bp (star in Fig. 7D and 7E) but also several other bands that were less abundant and differed between the cDNA samples from M8 and ND5 cells, indicating that the primers also mis-prime cDNA of other unknown genes expression of which differs between cell lines. gDNA samples also produced a band that was similar in size to the Actb cDNA but was very fuzzy, likely because it was a mixture of different bands, including some that were actually slightly smaller than the Actb cDNA (arrow in Fig. 7D and 7E). Likely, this or these fuzzy bands are derived also from some PGs, besides the authentic Actb gene that should be amplified as both primers locate at the same exon (exon 6). The Actb primers also amplified other bands from gDNA samples that differed between the two cell lines as well (Fig. 7D and 7E). We had also studied many other mouse cell lines and tissues and obtained very similar results (data not shown).

Discussion

Our bioinformatic analyses show that the ACTB (Actb) and GAPDH (Gapdh) genes have 64–197 putative PGs (table 1) that score over 200 and have over 80% identity to the corresponding

parental mRNA, based on the UCSC Genome Browser [31]. There are some more genomic fragments scored lower than 200 and thus not accounted in as PGs, but they span over several hundred nt, have over 83% identity to the authentic mRNA, and may still be mis-primed. If, like its 1% that has been studied [20], the human genome has 93% of its bases being transcribed, we may have to accept the new concept that most of the genome is transcribed at least in some cell types or at some times. Because each of these PGs resides at a different chromosomal site from the authentic gene, if any of them is expressed, it is controlled by different transcription-regulatory elements and thus is actually a different gene. Therefore, before the ACTB (Actb) or GAPDH (Gapdh) can be used as a reference in RT-PCR, it needs to be confirmed that none of their 64–197 putative PGs is transcribed in the particular cell (tissue) or situation of interest. To determine whether many processed PGs are expressed or not, the only strategy we can think of is to clone the RT-PCR products from each interested cell line or tissue into a vector, followed by sequencing a large number of plasmid clones to ensure that none of the clones has a PG sequence. If some PGs are found to be expressed, how their transcription is regulated needs to be determined, which needs to use another gene as the reference, so as to determine whether they meet the criteria of a reference gene, including refractoriness to the to-be-studied situation. Without saying, it is practically impossible to perform such a tedious and cumbersome sideshow to determine the expression status of so many PGs and to determine their transcriptional features, especially when a study involves multiple cell lines (tissues) or multiple experimental situations. It is much simpler to forgo these genes and elect someone else, such as the HPRT1 or Hprt.

In eukaryotic genomes, 2.7–97.7% of the genes are intronless [33], such as about 50% of the G protein coupled receptor genes in the human [34], while many other genes have processed PGs. RT-PCR amplification of the RNA transcripts from these two classes of genes requires a perfect DNase digestion to remove not only genomic DNA residuals but also mitochondrial DNA from the RNA sample, since some chromosomal genes are highly homologous to their ancestors in mitochondria [35,36]. According to our experience, complete DNase digestion that leaves no traceable DNA residual in the RNA sample is actually not so easy, because PCR is supersensitive. PG may cause artifact without

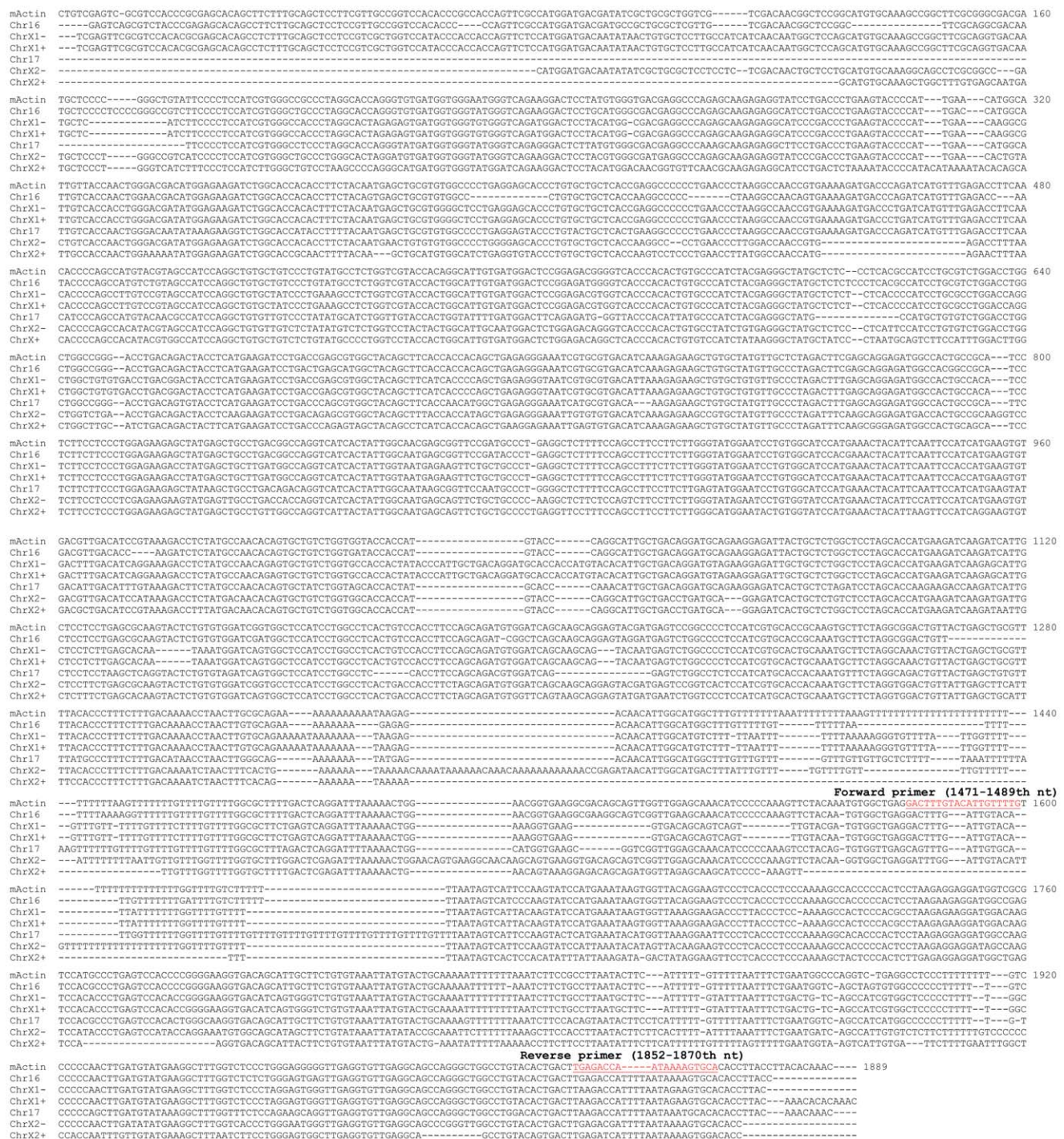


Figure 4. Alignment of the Actb mRNA with the six PGs that are the most homologous to the Actb mRNA shows that the Actb mRNA has no unique region that is long enough to be a primer. The 1471–1489th nt and 1852–1870th nt regions of the Actb mRNA (underlined) have the most mismatches compared with other regions and thus are used as the forward and reverse primers, respectively. doi:10.1371/journal.pone.0041659.g004

have them [21]. The simplest way is to use the to-be-studied mRNA sequence, after removing its poly-A tail if there is one, as a bait to fish out PGs from the corresponding genome. Primers should be assigned to the regions unique to the mRNA, such as the first 26 nt of the wild type GAPDH (NM_002046.4). If there is no unique region, a second step should be taken to align the mRNA with the PG sequence, so as to identify regions that encompass the

most mismatches, such as the region used as our GAPDH reverse primer. Once such a region is selected, the whole mRNA sequence, but not just the selected region, should be used to run the routine primer designing software to further analyze whether a primer can be identified within the selected region. It merits mention that one type of quantitative PCR technique uses a gene-specific oligo (with modified backbone) probe, besides the two



Figure 5. Alignment of the human GAPDH mRNA with seven PGs that are the best homologous to the wild type GAPDH mRNA shows that the first 26 nt of the mRNA is the only unique region while the 685–705th nt region of the GAPDH has the most mismatches (shown as underlined lowercase letters in PGs). We select the 6–25th and the 685–705th regions as the forward and reverse primers (underlined), respectively.
doi:10.1371/journal.pone.0041659.g005

primers, to increase the specificity. This strategy should also enhance the preference to the authentic cDNA if, like the two primers, the probe is also assigned to a region containing mismatches. However, it remains possible that like the two primers, the probe also mis-annoates with some PGs, due to the high sequence similarity. It also merits mention that because mouse genome varies hugely among different strains [37], the real risk of mis-priming may be higher in the cells or tissues from some strains of mice.

Specificity is an important criterion of primer design for PCR, which hitherto is concerned only vertically: appearance of additional band(s) above or below the expected one(s) in the same lane of the agarose gel is indicative of non-specificity of the primer pair, as seen in the PCR results from our ACTB and Actb primers (Fig. 7A, 7D and 7E). We now propose to consider the specificity also horizontally as part of the SOP: gDNA sample should be included in PCR as template and in the ensuing gel electropho-

resis, so as to determine whether there are processed PGs amplified. This is needed if it is preferred not to determine whether a PG is transcribed in the to-be-studied cell type or situation. If the expected band appears also in the gDNA samples, the primers need to be redesigned or a complete DNase digestion of RNA sample needs to be ensured. As an example of the horizontal criteria, the slight difference in the expected band of Actb between cDNA and gDNA samples is discerned when they were loaded into the agarose gel in a side-by-side manner, i.e. the band from gDNA samples is fuzzy and seems to be mixed bands (Fig. 7D and 7E), suggesting that the primers may not be specific horizontally, although they were selected already under the best combination of PGs. A caveat needs to be given that it is generally more difficult to perform PCR with gDNA as template, in part because long chromatin DNAs are highly wound and difficult to be denatured by heating and annealed by primers. This may be one of the reasons why results from gDNA samples may differ

Chr4-1	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr15	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
mGAPDH	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
ChrX	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr5	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr11	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr2	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr4-2	AGAGACAGCGCGATCTTCTTGTGAGTGCAGAGCTCGTCCGTAGACAAATGGTGAAGTGGTGTGAACGGATTGGGCGGCTTGGTACAGGGGTGCCATTGTCAGTGGCAAGTGGAGATTGTTGCCATCAACGACCCCTTCATTGA
Chr4-1	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr15	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
mGAPDH	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
ChrX	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr5	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr11	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr2	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr4-2	CCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTACGCGCAATTCACGCGCAGCTCAAGCGCGAGATGGGAAGCTTGTCTCAACGGGAAGGCCATCACCATCTCCAGGAGCGAGACCCCACTACATCAATGGGGTAGGCGCGTGTCT
Chr4-1	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr15	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
mGAPDH	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
ChrX	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr5	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr11	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr2	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr4-2	GAGTATGTCGTGGAGTCTACTGGTGTCTTCCACCACATGGAAGAGCGCGGGGCCACTTGAAGGTTGGAGCAAAAGGTCATCATCTCCGCCCTTCTGCCGATGCCCATGTTTGTGATGGGTGTGAACACAGAGAAATGACAACTCACTCAAGA
Chr4-1	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr15	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
mGAPDH	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
ChrX	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr5	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr11	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr2	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr4-2	TTTGACAGATGCATCTGACACACCAACTGCTTAGCCCCCTGGCCAGGTCATCATGACAACTTTGGCATTTGTGAAGGGCTCATGACACAGTCCATGCCATCTGCCACCCAGAGACTGTGATGGCCCTCTGGAAGCTGTGGCGTGTATGG
Chr4-1	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr15	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
mGAPDH	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
ChrX	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr5	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr11	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr2	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr4-2	CGTGGGGTGGCCAGAACATCATCCCTGCATCCATGGTGTGCCAAGGCTGTGGGCAAGTGCATCCAGAGCTGAACGGGAAGCTCACTGGCATGGCCTTCGTGTCTACCCCCAATGTGTCGTGGTGTGATGTCGCGCGCTGGAGAACT
Chr4-1	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr15	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
mGAPDH	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
ChrX	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr5	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr11	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr2	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr4-2	GCCAAATGATGATGACATCAAGAGGTGGTGAAGCAGGACATGAGGGGCCACTGAAGGGCATCTTGGGCTACACTGAGGACAGGTTGTCTCTGCGACTTCAACAGCACTCCCACTCTTCCACTTTCGATGCGGGGCTGGCATTTGCTCTCAATGACA
Chr4-1	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr15	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
mGAPDH	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
ChrX	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr5	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr11	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr2	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr4-2	ACTTTGTCAAGCTCATTTCTCGGTATGACAATGAATACGGCTACAGCAACAGGGTGTGGACCTCATGGCTCATAGGCTCCAGAGGATTAAGAAACCTGGACACCCACCCAGCAGGACATGAGCAAGAGAGAGGCCATCTCCCACTCGGCC
Chr4-1	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
Chr15	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
mGAPDH	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
ChrX	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
Chr5	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
Chr11	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
Chr2	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-
Chr4-2	CAACATGAGCATCTCCCTCACAATTTCCATCCAGACCCCAATAAACAGAGAGGGGCTAGGAGGCCCTCCCTACTCTTCTTGAATACCATCAATAAAGTTCGCTGACACCA-

Figure 6. Alignment of the mouse *Gapdh* mRNA with seven PGs that are the best homologous to the *Gapdh* mRNA shows that these PGs are almost identical to the *Gapdh* with only several mismatches. No region of the *Gapdh* can be used as a primer that can significantly discriminate against any of the PGs.

doi:10.1371/journal.pone.0041659.g006

among different cell lines for our ACTB and GAPDH primers and why our ACTB primers, both of which locate at the same exon (exon 6), seem to amplify cDNA samples more easily than gDNA samples (Fig. 7A). PCR results of the ACTB (Actb) and GAPDH (Gapdh) shown in the literature, including those from us previously, may not be specific if evaluated horizontally. Indeed, we randomly performed bioinformatic analyses of quite a few PCR primers of these genes reported in the three most prestigious journals, i.e. Nature, Science and Cell; without exception, all those primers match well and amplify the corresponding PGs by *in silico* PCR. Whether or not the previously published RT-PCR results that involve these genes (especially the *Gapdh*) as the reference need to be reevaluated or reinterpreted is an uncomfortable but unavoidable question, and should be left to the corresponding authors to decide accordingly.

An additional reason to abandon ACTB is that the DNA fragment amplified by our primer pair, which is the only one we can find in consideration of the specificity horizontally, is only 246-bp long (table 2). This size may be suitable for real-time RT-PCR but is too short to sensitively reflect a difference in copy numbers of the ACTB mRNA, if the RT-PCR products are evaluated by visualization in an agarose gel, because it requires more copies of a small DNA fragment to reach a visible amount of

nucleotides in a gel. For instance, one more copy of a 246-bp double-stranded DNA fragment only adds 492 nucleotides in the gel, which is not as visible as one more copy of a 1-kb DNA fragment that adds 2,000 nucleotides. For this technical reason, we usually set, if possible, primer pairs to amplify DNA fragments of 400–700 bp and to minimize the difference in the fragment sizes between the reference gene and the interested gene.

In summary, in this study we provide bioinformatic data showing that the genes encoding β -actin and glyceraldehyde-3-phosphate dehydrogenase have many PGs in the human and mouse genomes. These PGs may affect the fidelity of ACTB (Actb) or GAPDH (Gapdh) as a reference in RT-PCR by their genomic DNA or, if some of them are expressed, by their RNA transcript, because a large copy number in the genome may amplify the artifact derived from genomic DNA residual in the RNA sample during PCR whereas their RNA transcript may contribute to the yield of RT-PCR. We suggest to peers to forgo these genes, especially the *Gapdh*, as a reference in RT-PCR or, if there is no suitable surrogate, to use with extra caution our primers and PCR conditions provided herein that may better avoid mis-priming PGs, relative to most primers described in the literature. We also propose an SOP in which design of primers for RT-PCR starts from avoiding mis-priming PGs and all primers need be tested

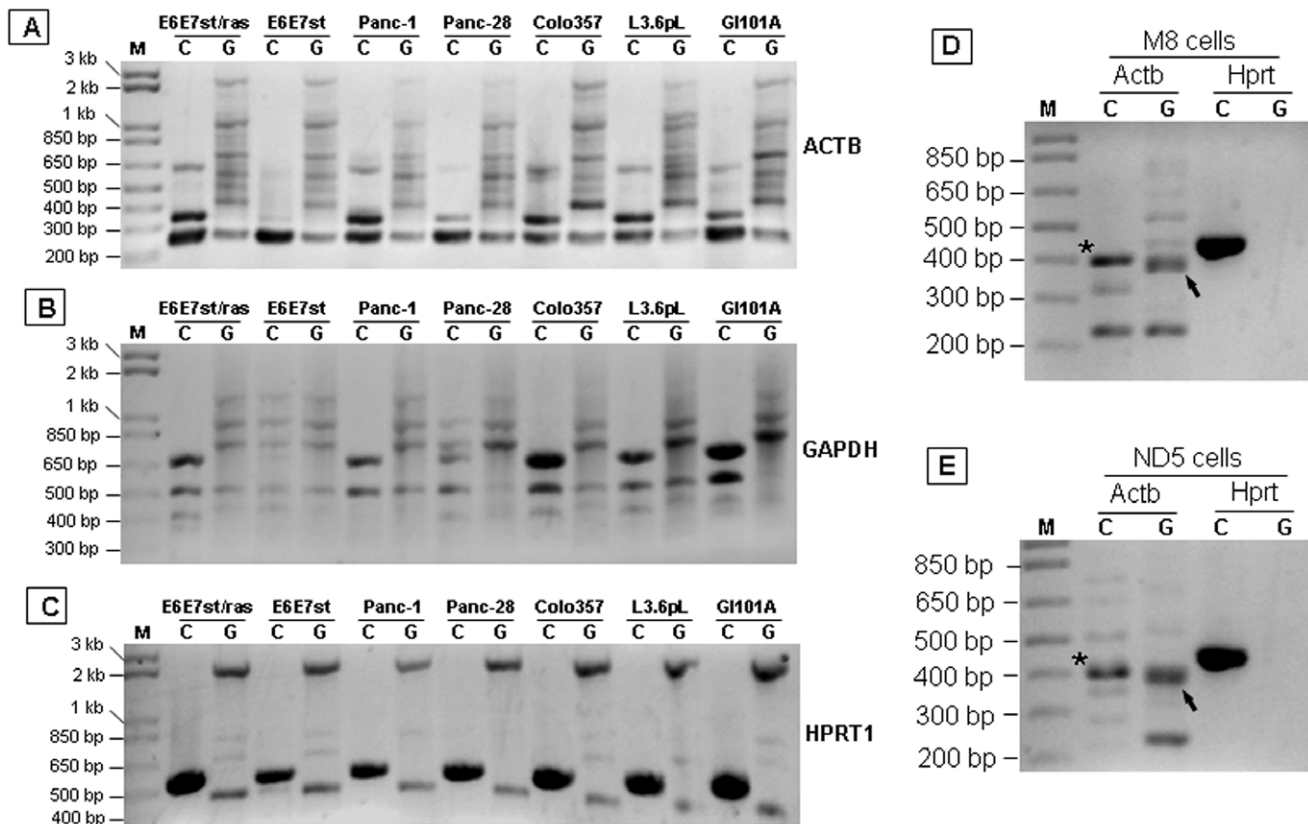


Figure 7. Determination of the primer specificity for PCR both vertically and horizontally. A–C: gDNA (G) and cDNA (C) samples from a panel of human cell lines (described in materials and methods) were amplified by PCR with conditions of initial denature at 95°C for 5 min and 40 cycles of melting at 95°C for 30 sec, primer-annealing at 58°C for 30 sec, and elongating at 72°C for 30 sec. The reaction was determined at 72°C for 10 min. M is molecular weight marker. D and E: gDNA (G) and cDNA (C) samples from M8 and ND5 mouse cell lines were amplified by PCR under the conditions described above. Stars indicate the authentic Actb cDNA band whereas arrows indicate its counterpart from gDNA samples. doi:10.1371/journal.pone.0041659.g007

with not only cDNA but also gDNA to ensure their specificity both vertically and horizontally.

Supporting Information

Figure S1 ACTB, Actb, GAPDH, Gadph, HPRT1 and Hprt mRNA sequences pulled out from the NCBI database that presents all mRNA as DNA sequence, i.e. with thymine replacing uracil. (DOC)

Figure S2 Putative PGs of the ACTB identified by Blat search using the ACTB mRNA sequence (after deletion of the poly-A tail). The top sequence that has 100% identity to the bait is the authentic ACTB gene on chromosome 7. The six genomic DNA fragments that have the highest scores to the bait were used in the alignment with the bait sequence shown in figure 3. (DOC)

Figure S3 Putative PGs of the Actb identified by Blat search using the Actb mRNA sequence. The top sequence that has 100% identity to the bait is the authentic Actb gene on mouse chromosome 5. The six genomic DNA fragments that have the highest scores to the bait were used in the alignment with the bait sequence shown in figure 4. (DOC)

Figure S4 Putative PGs of the GAPDH identified by Blat search using the GAPDH mRNA sequence (after deletion of the poly-A tail). The top sequence that has 100% identity to the bait is the authentic GAPDH gene on human chromosome 12. The seven genomic DNA fragments in the red box that have the highest scores to the bait were used in the alignment with the bait sequence shown in figure 5. (DOC)

Figure S5 Putative PGs of the Gapdh identified by Blat search using the Gapdh mRNA sequence (after deletion of the poly-A tail). The top sequence that has 100% identity to the bait is the authentic Gapdh gene on the mouse chromosome 7. The seven genomic DNA fragments in the red box that have the highest scores to the bait were used in the alignment with the bait sequence shown in figure 6. (DOC)

Acknowledgments

We would like to thank Fred Bogott, M.D., Ph.D., at Austin Medical Center, Austin of Minnesota, for his excellent English editing of this manuscript.

Author Contributions

Conceived and designed the experiments: DJL DZL. Performed the experiments: YS. Analyzed the data: YL. Wrote the paper: DJL.

References

- Huggett J, Dheda K, Bustin S, Zumla A (2005) Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun* 6: 279–284.
- Cinar MU, Islam MA, Uddin MJ, Tholen E, Tesfaye D, et al (2012) Evaluation of suitable reference genes for gene expression studies in porcine alveolar macrophages in response to LPS and LTA. *BMC Res Notes* 5: 107-doi: 10.1186/1756-0500-5-107
- Guo R, Ki JS (2011) Evaluation and validation of internal control genes for studying gene expression in the dinoflagellate *Prorocentrum minimum* using real-time PCR. *Eur J Protistol* -doi: org/10.1016/j.ejop.2011.11.001
- Zhang G, Zhao M, Song C, Luo A, Bai J, et al (2012) Characterization of reference genes for quantitative real-time PCR analysis in various tissues of *Anoectochilus roxburghii*. *Mol Biol Rep* 39: 5905–5912.
- Everaert BR, Boulet GA, Timmermans JP, Vrints CJ (2011) Importance of suitable reference gene selection for quantitative real-time PCR: special reference to mouse myocardial infarction studies. *PLoS One* 6: e23793-doi: 10.1371/journal.pone.0023793
- Sun JH, Nan LH, Gao CR, Wang YY (2012) Validation of reference genes for estimating wound age in contused rat skeletal muscle by quantitative real-time PCR. *Int J Legal Med* 126: 113–120.
- Dupont M, Goldsborough A, Levayer T, Savare J, Rey JM, et al (2002) Multiplex fluorescent RT-PCR to quantify leukemic fusion transcripts. *Biotechniques* 33: 158–60, 162, 164.
- Stephens AS, Stephens SR, Morrison NA (2011) Internal control genes for quantitative RT-PCR expression analysis in mouse osteoblasts, osteoclasts and macrophages. *BMC Res Notes* 4: 410-doi: 10.1186/1756-0500-4-410
- Penna I, Vella S, Gignoli A, Russo C, Cancedda R, et al (2011) Selection of candidate housekeeping genes for normalization in human postmortem brain samples. *Int J Mol Sci* 12: 5461–5470.
- Ledderose C, Heyn J, Limbeck E, Kreth S (2011) Selection of reliable reference genes for quantitative real-time PCR in human T cells and neutrophils. *BMC Res Notes* 4: 427-doi: 10.1186/1756-0500-4-427
- Manjarin R, Trottier NL, Weber PS, Liesman JS, Taylor NP, et al. (2011) A simple analytical and experimental procedure for selection of reference genes for reverse-transcription quantitative PCR normalization data. *J Dairy Sci* 94: 4950–4961.
- Suzuki T, Higgins PJ, Crawford DR (2000) Control selection for RNA quantitation. *Biotechniques* 29: 332–337.
- Stecyk JA, Couturier CS, Fagermes CE, Ellefsen S, Nilsson GE (2012) Quantification of heat shock protein mRNA expression in warm and cold anoxic turtles (*Trachemys scripta*) using an external RNA control for normalization. *Comp Biochem Physiol Part D Genomics Proteomics* 7: 59–72.
- Kim I, Yang D, Tang X, Carroll JL (2011) Reference gene validation for qPCR in rat carotid body during postnatal development. *BMC Res Notes* 4: 440-doi: 10.1186/1756-0500-4-440
- Harper LV, Hilton AC, Jones AF (2003) RT-PCR for the pseudogene-free amplification of the glyceraldehyde-3-phosphate dehydrogenase gene (gapd). *Mol Cell Probes* 17: 261–265.
- Williams TK, Yeo CJ, Brody J (2008) Does this band make sense? Limits to expression based cancer studies. *Cancer Lett* 271: 81–84.
- Lehmann MH, Weber J, Gastmann O, Sigusch HH (2002) Pseudogene-free amplification of human GAPDH cDNA. *Biotechniques* 33: 766, 769–766, 770.
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al (2001) The sequence of the human genome. *Science* 291: 1304–1351.
- Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? *Hum Mol Genet* 19: R162–R168.
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816.
- Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, et al (2011) Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA* 17: 792–798.
- Gibb EA, Brown CJ, Lam WL (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer* 10: 38-doi: 10.1186/1476-4598-10-38
- Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146: 353–358.
- Liedtke S, Stephan M, Kogler G (2008) Oct4 expression revisited: potential pitfalls for data misinterpretation in stem cell research. *Biol Chem* 389: 845–850.
- Zhao S, Yuan Q, Hao H, Guo Y, Liu S, et al (2011) Expression of OCT4 pseudogenes in human tumours: lessons from glioma and breast carcinoma. *J Pathol* 223: 672–682.
- Guo X, Tang Y (2011) OCT4 pseudogenes present in human leukemia cells. *Clin Exp Med* -doi: 10.1007/s10238-011-0163-4
- Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, et al (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17: 669–681.
- Campbell PM, Groehler AL, Lee KM, Ouellette MM, Khazak V, et al (2007) K-Ras promotes growth transformation and invasion of immortalized human pancreatic cells by Raf and phosphatidylinositol 3-kinase signaling. *Cancer Res* 67: 2098–2106.
- Biliran H Jr, Banerjee S, Thakur A, Sarkar FH, Bollig A, et al (2007) c-Myc-induced chemosensitization is mediated by suppression of cyclin D1 expression and nuclear factor-kappa B activity in pancreatic cancer cells. *Clin Cancer Res* 13: 2811–2821.
- Wang Y, Thakur A, Sun Y, Wu J, Biliran H, et al (2007) Synergistic effect of cyclin D1 and c-Myc leads to more aggressive and invasive mammary tumors in severe combined immunodeficient mice. *Cancer Res* 67: 3698–3707.
- Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Rancey BJ, et al (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 40: D918–D923.
- McDonnell L, Drouin G (2012) The abundance of processed pseudogenes derived from glycolytic genes is correlated with their expression level. *Genome* 55: 147–151.
- Louhichi A, Fourati A, Rebai A (2011) IGD: a resource for intronless genes in the human genome. *Gene* 488: 35–40.
- Markovic D, Challiss RA (2009) Alternative splicing of G protein-coupled receptors: physiology and pathophysiology. *Cell Mol Life Sci* 66: 3337–3352.
- Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet* 6: e1000834.
- Calabrese FM, Simone D, Attimonelli M (2012) Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* 13 Suppl 4: S15.
- Keane TM, Goodstadt L, Danecek P, White MA, Wong K, et al (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477: 289–294.

Human genes

Human ACTB mRNA:

>gi|168480144|ref|NM_001101.3| Homo sapiens actin, beta (ACTB), mRNA:

ACCGCCGAGACCGCGTCCGCCCCGCGAGCACAGAGCCTCGCCTTTGCCGATCCGCCGCCCGTCCACACCCGCCCGCAGCTCACCATGGATGATGATATCGCCGCGCTCGTCG
TCGACAACGGCTCCGGCATGTGCAAGGCCGGCTTCGCGGGCGACGATGCCCCCGGGCGCTTCCCTCCATCGTGGGGCGCCCCAGGCACCAAGGCGTGATGGTGGGCAT
GGGTGAGAAGGATTCTATGTGGGCGACGAGGCCAGAGCAAGAGAGGCATCCTCACCTGAAGTACCCCATCGAGCACGGCATCGTCACCAACTGGGACGACATGGAGAAA
ATCTGGCACACACCTTCTACAATGAGCTGCGTGTGGCTCCCGAGGAGCACCCCGTGTGCTGACCGAGGCCCCCTGAACCCCAAGGCCAACCGCGAGAAGATGACCAGAA
TCATGTTTGAGACCTTCAACACCCAGCCATGTACGTTGCTATCCAGGCTGTGCTATCCCTGTACGCCCTTGCCCGTACCACTGGCATCGTGATGGACTCCGGTGACGGGGT
CACCCACACTGTGCCATCTACGAGGGGTATGCCCTCCCCATGCCATCCTGCGTCTGGACCTGGCTGGCGGGACCTGACTGACTACCTCATGAAGATCCTCACCGAGCGC
GGCTACAGCTTACCACCACGGCCGAGCGGGAATCGTGCGTGACATTAAGGAGAAGCTGTGTACGTGCGCTGGACTTCGAGCAAGAGATGGCCACGGCTGCTTCCAGCT
CCTCCCTGGAGAAGAGCTACGAGCTGCCGACGGCCAGGTATCACCATTGGCAATGAGCGGTTCCGCTGCCCTGAGGCACCTTCCAGCCTTCTTCTGGGCATGGAATC
CTGTGGCATCCACGAAACTACCTTCAACTCCATCATGAAGTGTGACGTGGACATCCGCAAGACCTGTACGCCAACACAGTGTGTCTGGCGGCACCACTATGACCTGGC
ATTGCCGACAGGATGCAGAAGGAGATCACTGCCCTGGCACCCAGCACAAATGAAGATCAAGATCATTGCTCCTCTGAGCGCAAGTACTCCGTGTGGATCGCGGGCTCCATCC
TGGCCTCGCTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTATGACGAGTCCGGCCCTCCATCGTCCACCGCAATGCTTCTAGCGGACTATGACTTAGTTGC
GTTACACCTTTCTTGACAAAACCTAATTGCGCAGAAAAACAAGATGAGATTGGCATGGCTTTATTTGTTTTTTTTGTTTTGTTTTGTTTTTTTTTTTTTTTTTGGCTTGAC
TCAGGATTTAAAAACTGGAACGGTGAAGGTGACAGCAGTCGGTTGGAGCGAGCATCCCCAAAGTTACAATGTGGCCGAGGACTTTGATTGCACATTGTTGTTTTTTAAT
AGTCATTCCAAATATGAGATGCGTTGTTACAGGAAAGTCCCTTGCCATCCTAAAAGCCACCCCACTTCTCTCTAAGGAGAATGGCCAGTCCTCTCCAAAGTCCACACAGGGG
AGGTGATAGCATTGCTTTCTGTGTAATTTATGTAATGCAAAATTTTTTAATCTTGCCTTAATACTTTTTTATTTTGTATTTTGAATGATGAGCCTTCGTGCCCCCCT
TCCCCCTTTTTTGCCCCCAACTTGAGATGTATGAAGGCTTTTGGTCTCCCTGGGAGTGGGTGGAGGCAGCCAGGGCTTACCTGTACACTGACTTGAGACCAGTTGAATAAA
AGTGACACCTTAAAAATGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

Human GAPDH mRNA:

>gi|83641890|ref|NM_002046.4| Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA:

GGCTGGGACTGGCTGAGCCTGGCGGGAGGCGGGGTCCGAGTCACCGCCTGCCGCCGCGCCCCGGTTTCTATAAATTGAGCCCGCAGCCTCCCGCTTCGCTCTCTGCTCCTC
CTGTTTCGACAGTCAGCCGCATCTCTTTTTCGCTCGCCAGCCGAGCCACATCGCTCAGACACCATGGGAAGGTGAAGGTGCGAGTCAACGGATTGGTCTGATTGGCGCCT
GGTCACCAGGGCTGCTTTAACTCTGGTAAAGTGGATATTGTTGCCATCAATGACCCCTTCATTGACCTCAACTACATGGTTTACATGTTCCAATATGATTCCACCATGGC
AAATTCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTATCAATGGAATCCCATCACCATCTTCCAGGAGCGAGATCCCTCCAAATCAAGTGGGGCGATGCTGGCG
CTGAGTACGTCGTGGAGTCCACTGGCGTCTTACCACCATGGAGAAGGCTGGGGCTCATTTGCAGGGGGGAGCCAAAAGGGTCATCATCTCTGCCCCCTCTGCTGATGCCCC
CATGTTTCGTATGGGTGTGAACCATGAGAAGTATGACAACAGCCTCAAGATCATCAGCAATGCCTCCTGCACCACCAACTGCTTAGCACCCCTGGCCAAGGTCATCCATGAC
AACTTTGGTATCGTGGAGGACTCATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCTCCGGGAAACTGTGGCGTGATGGCCGCGGGGCTCTCC
AGAACATCATCCTCGCTCTACTGGCGCTGCCAAGGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCTTCCTGTCCCCACTGCCAACGTTGCT
AGTGGTGAGCTGACCTGCCGTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAGGTGGTGAAGCAGGCGTCGGAGGGCCCCCTCAAGGGCATCCTGGGCTACACTGAG
CACCAGGTGGTCTCCTCTGACTTCAACAGCGACACCCACTCCTCCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGTCAAGCTCATTTCTCTGGTATGACA
ACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATGGCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACGAGCCAGCAAGAGACAAGAGGAAGAGAGAGACC
CTCACTGCTGGGAGTCCCTGCCACACTCAGTCCCCACCACACTGAATCTCCCTCCTCACAGTTGCCATGTAGACCCCTTGAAGAGGGGAGGGGCTAGGGAGCCGCACC
TTGTCATGTACCATCAATAAAGTACCCTGTGCTCAACCAAAAAAAAAAAAAAAAAAAA

Human GAPDH variant 2; NM_001256799.1:

GTGCGCAGCGGGTGATCCCTGTCCGGATGTGCGCCTGCGGTAGAGCGGCCGCCATGTTGCAACCGGGAAGGAAATGAATGGGCAGCCGTTAGGAAAGCCTGCCGGTGACT
AACCCTGCGCTCCTGCCTCGATGGGTGGAGTCGCGTGTGGCGGGGAAGTCAGGTGGAGCGAGGCTAGCTGGCCGATTTCTCCTCGGGTGATGCTTTTCTAGATTATTCTC
TGATTTGGTCTGATTGGCGCCTGGTACCAGGGCTGCTTTTAACTCTGGTAAAGTGGATATTGTTGCCATCAATGACCCCTTCATTGACCTCAACTACATGGTTTACATGT
TCCAATATGATTCCACCATGGCAAATTCATGGCACCGTCAAGGCTGAGAACGGGAAGCTTGTATCAATGGAATCCCATCACCATCTTCCAGGAGCGAGATCCCTCCAA
AATCAAGTGGGGCGATGCTGGCGCTGAGTACGTGCTGGAGTCCATGGCGTCTTACCACCATGGAGAAGGCTGGGGCTCATTTGCAAGGGGGAGCCAAAAGGGTCATCATC
TCTGCCCCCTCTGCTGATGCCCCCATGTTGCTCATGGGTGTGAACCATGAGAAGTATGACAACAGCCTCAAGATCATCAGCAATGCCCTCCTGCACCACCAACTGCTTAGCAC
CCCTGGCCAAGGTCATCCATGACAACTTTGGTATCGTGGAAAGGACTCATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCTCCGGGAAACTGTG
GCGTGATGGCCGCGGGGCTCTCCAGAACATCATCCTGCTCTACTGGCGTGGCAAGGCTGTGGGCAAGGTCATCCCTGAGCTGAACGGGAAGCTCACTGGCATGGCCTTC
CGTGTCCCCTGCAACGTTGCTAGTGGTGGACCTGACCTGCCGTCTAGAAAAACCTGCCAAATATGATGACATCAAGAAGGTGGTGAAGCAGGCGTCGGAGGGCCCCCTCA
AGGGCATCCTGGGCTACACTGAGCACCAGGTGGTCTCCTCTGACTTCAACAGCGACACCCACTCCTCCACCTTTGACGCTGGGGCTGGCATTGCCCTCAACGACCACTTTGT
CAAGCTCATTTCTGTTATGACAACGAATTTGGCTACAGCAACAGGGTGGTGGACCTCATGGCCACATGGCCTCCAAGGAGTAAGACCCCTGGACCACGAGCCCAAG
AGCACAAGAGGAAGAGAGAGACCCTCACTGCTGGGAGTCCCTGCCACACTCAGTCCCCACCACACTGAATCTCCCTCCTCACAGTTGCCATGTAGACCCCTTGAAGAGG
GGAGGGGCTAGGGAGCCGCACCTTGTATGTACCATCAATAAAGTACCCTGTGCTCAACCAAAAAAAAAAAAAAAAAAAA

Human HPRT1 mRNA:

>gi|164518913|ref|NM_000194.2| Homo sapiens hypoxanthine phosphoribosyltransferase 1 (HPRT1), mRNA:

GGCGGGGCTGCTTCTCCTCAGCTTCAGGCGGCTGCGACGAGCCCTCAGGCGAACCTCTCGGCTTTCCGCGCGGCGCGCCCTCTTGCTGCGCTCCGCTCCTCCTCTGCT
CCGCCACCGGCTTCTCCTCCTGAGCAGTCAGCCGCGCGCGGGCGGCTCCGTTATGGCGACCCGCGAGCCCTGGCGTCTGTGATTAGTGATGATGAACCAGGTTATGACCTT
GATTTATTTTGCATACCTAATCATTATGCTGAGGATTTGGAAGGGTGTTTATTTCTCATGGACTAATTATGGACAGGACTGAACGCTTCTGCTCGAGATGTGATGAAGGAGA
TGGGAGGCCATCACATTGTAGCCCTCTGTGTGCTCAAGGGGGGCTATAAATCTTTGCTGACCTGCTGGATTACATCAAAGCACTGAATAGAAATAGTGATAGATCCATTCC
TATGACTGTAGATTTTATCAGACTGAAGAGCTATTGTAATGACCAGTCAACAGGGGACATAAAAGTAATTGGTGGAGATGATCTCTCAACTTTAACTGGAAAGAAATGTCTTG
ATTGTGGAAGATATAATTGACACTGGCAAAACAATGCAGACTTTGCTTTCTTGGTCAGGCAGTATAATCCAAAGATGGTCAAGGTCGCAAGCTTGTGGTGAAGGACCC
CACGAAGTGTGGATATAAGCCAGACTTTGTTGGATTTGAAATTCAGACAAGTTTGTGTAGGATATGCCCTTGACTATAATGAATACTTCAGGGATTTGAATCATGTTTG
TGTCATTAGTGAAACTGGAAAAGCAAAATACAAAGCCTAAGATGAGAGTTCAGTTGAGTTTGGAAACATCTGGAGTCTATTGACATCGCCAGTAAATATCAATGTTCT
AGTTCTGTGGCCATCTGCTTAGTAGAGCTTTTGCATGTATCTTCTAAGAAATTTATCTGTTTGTACTTTAGAAATGTCAGTTGCTGCATTCTTAACTGTTTATTTGCAC

TATGAGCCTATAGACTATCAGTTCCTTTGGGCGGATTGTTGTTTAACTTGTAATGAAAAAATCTCTTAAACCACAGCACTATTGAGTGAAACATTGAACTCATATCTGT
AAGAAATAAAGAGAAGATATATTAGTTTTTTAATTGGTATTTTATATATGCAGGAAAGAATAGAAGTGATTGAATATTGTTAATTATACCACCGTGTGTTAGAAA
AGTAAGAAGCAGTCAATTTTCACATCAAAGACAGCATCTAAGAAGTTTTGTTCTGCTCGAATTATTTTAGTAGTGTTCAGTAATGTTGACTGTATTTTCCAACCTGTTT
AAATTATTACCAGTGAATCTTTGTCAGCAGTTCCTTTTAAATGCAATCAATAAATTCCTCAAAAAATTTAAAAA

Mouse genes

Mouse Actb mRNA:

>gi|145966868|ref|NM_007393.3| Mus musculus actin, beta (Actb), mRNA:

CTGTCGAGTCGCGTCCACCCGCGAGCACAGCTTCTTTGCGAGCTCCTCGTTGCGGGTCCACACCCGCCACAGTTCGCCATGGATGACGATATCGCTGCGCTGGTCTGTCGAC
AACGGCTCCGGCATGTGCAAAAGCCGGCTTCGCGGGCGACGATGCTCCCGGGCTGTATTCCCTCCATCGTGGGCCGCCCTAGGCCACAGGGTGTGATGGTGGGAATGGGTC
AGAAGGACTCCTATGTGGGTGACGAGGCCAGAGCAAGAGAGGTATCTGACCTGAAGTACCCATTGAACATGGCATTGTTACCAACTGGGACGACATGGAGAAGATCTG
GCACCACACCTTCTACAATGAGCTGCGTGTGGCCCTGAGGAGCACCTGTGCTGCTACCGAGGCCCCCTGAACCTAAGGCCAACCGTGAAGAGATGACCAGATCATG
TTTGAGACCTTCAACACCCAGCCATGTACGTAGCCATCCAGGCTGTGCTGTCCCTGTATGCCTCTGGTCTGACACAGGCATTGTGATGGACTCCGGAGACGGGGTCAACC
ACACTGTGCCATCTACGAGGGCTATGCTCTCCCTACGCCATCCTGCGTCTGGACCTGGCTGGCCGGGACCTGACAGACTACCTCATGAAGATCCTGACCGAGCGTG6CTA
CAGCTTACCACCACAGCTGAGCGGGAATCTGCGTGACATCAAAGAGAAGCTGTGCTATGTTGCTCTAGACTTCGAGCAGGAGATGGCCACTGCCGCATCCTTCTCTCC
CTGGAGAAAGAGCTATGAGCTGCCGTGACGACGATCACTATTGGCAACAGCGGTTCCGATGCCCTGAGGCTCTTTTCCAGCCTTCTTCTGGGTATGGAATCCTGTG
GCATCCATGAAACTACATTCAATTCCATCATGAAGTGTGACGTTGACATCCGTAAGACCTCTATGCCAACACAGTGTGCTGGTGGTACCACCATGTACCCAGGCATTGC
TGACAGGATCGAGAAGGAGATTACTGCTCTGGCTCCTAGCACCATAAGATCAAGATCATTGCTCCTCTGAGCGCAAGTACTCTGTGTGGATCGGTGGCTCCATCCTGGCC
TCACTGTCCACCTTCCAGCAGATGTGGATCAGCAAGCAGGAGTACGATGAGTCCGGCCCTCCATCGTGACCGCAAGTGCTTCTAGGCGGAGTGTACTGAGCTGCGTTTT
ACACCTTTCTTTGACAAAACCTAATTGCGCAGAAAAAATAAGAGACAACATTGGCATGGCTTTGTTTTTAAATTTTTTAAAGTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTAAGTTTTTTGTTTTGTTTTGGCGCTTTTGACTCAGGATTTAAAACTGGAACGGTGAAGGCGACAGCAGTTGGTTGGAGCAAACATCCCCAAAGTTCTA
CAAATGTGGCTGAGGACTTTGTACATTGTTTTGTTTTTTTTTTTTTTTTGTTTTGTCTTTTTTAATAGTCATTCCAAGTATCCATGAAATAAGTGTTTACAGGAAGTCCCT
CACCCTCCCAAAGCCACCCCACTCCTAAGAGGAGGATGGTGGCTCCATGCCCTGAGTCCACCCCGGGGAAGGTGACAGCATTGCTTCTGTGTAATATGTACTGCAAA
AATTTTTTAAATCTTCCGCTTAATACTTCATTTTTGTTTTAATTTCTGAATGGCCAGGTCTGAGGCCCTCCCTTTTTTTGTCCCCCAACTTGATGTATGAAGGCTTT
GGTCTCCCTGGGAGGGGTTGAGGTGTTGAGGCAGCCAGGGCTGGCTGTACACTGACTTGAGACCAATAAAAGTGACACCTTACCTTACACAAAC

Mouse Gapdh mRNA:

>gi|126012538|ref|NM_008084.2| Mus musculus glyceraldehyde-3-phosphate dehydrogenase (Gapdh), mRNA:

AGAGACGGCCGCATCTTCTGTGCAAGTGCCAGCCTCGTCCCGTAGACAAAATGGTGAAGGTGGTGTGAACGGATTGGCCGATTGGGCGCTGGTACCAGGGCTGCCAT
TTGCAAGTGGCAAAGTGGAGATTGTTGCCATCAACGACCCCTTCACTGACCTCACTACATGGTCTACATGTTCCAGTATGACTCCACTCACGGCAAATCAACGGCACAGTC
AAGGCCGAGAATGGGAAGCTTGTATCAACGGGAAGCCATCACCATCTCCAGGAGCGAGACCCACTAACATCAAATGGGGTGAGGCCGGTGTGAGTATGTCTGGAGT
CTACTGGTGTCTTACCACCATGGAGAAGGCCGGGCCCATTTGAAGGGTGGAGCCAAAGGGTCATCATCTCGGCCCTTCTGCCGATGCCCCATGTTTGTGATGGGTGT
GAACCACGAGAAATATGACAACCTCACTCAAGATTGTACGAATGCATCCTGCACCACCAACTGCTTAGCCCCCTGGCCAAGGTATCCATGACAACCTTGGCATTGTGGAA
GGGCTCATGACCACAGTCCATGCCATCACTGCCACCCAGAAGACTGTGGATGGCCCTCTGGAAGCTGTGGCGTGTGGCGTGGGGCTGCCAGAACATCATCCCTGCAT
CCACTGGTGTCTGCCAAGGCTGTGGGCAAGGTATCCCAGAGCTGAACGGGAAGCTCACTGGCATGGCTTCCGTGTTCTACCCCCAATGTGTCCGTGTGGATCTGACGTG
CCGCTGGAGAACTGCCAAGTATGATGACATCAAGAAGGTGGTGAAGCAGGCATCTGAGGGCCACTGAAGGGCATCTGGGCTACACTGAGGACCAGGTTGTCTCTCTGC
GACTTCAACAGCAACTCCACTCTTCCACCTTCGATGCCGGGGCTGGCATTGCTCTCAATGACAACCTTGTCAAGCTCATTTCTGGTATGACAATGAATACGGCTACAGCA
ACAGGGTGGTGGACCTCATGGCTTACATGGCTCCTCAAGGAGTAAGAAACCCTGGACCACCCACCCAGCAAGGACACTGAGCAAGAGAGGCCCTATCCCAACTCGGCCCA
ACACTGAGCATCTCCCTACAATTTCCATCCAGACCCCATATAAACAGGAGGGGCTAGGGAGCCCTCCCTACTCTCTGAATACCATCAATAAAGTTGCTGACCCAC
AAAAA

Mouse Hprt mRNA:

>gi|96975137|ref|NM_013556.2| Mus musculus hypoxanthine guanine phosphoribosyl transferase (Hprt), mRNA:

GGAGCCTGGCCGGCAGCGTTTCTGAGCCATTGCTGAGGCGGCGAGGGAGAGCGTTGGGCTTACCTCACTGCTTCCGGAGCGGTAGCACCTCCTCCGCGGCTTCTCTCTCA
GACCGCTTTTGGCGGAGCCGACCGGTCCTCATGCCGACCCGAGTCCACGCTCGTGATTAGCGATGATGAACAGGTTATGACCTAGATTTGTTTGTATACCTAAT
CATTATGCCGAGGATTTGGAAAAAGTGTATTCTCTCATGGACTGATTATGGACAGGACTGAAAGACTTGCTCGAGATGTCATGAAGGAGATGGGAGGCCATCACATTGTGG
CCCTCTGTGTCTCAAGGGGGGTATAAGTTCTTTGCTGACCTGCTGGATTACATTAAGCACTGAATAGAAATAGTGATAGATCCATTCTATGACTGTAGATTTTATCAG
ACTGAAGAGCTACTGTAATGATCAGTCAACGGGGGACATAAAAGTTATTGGTGGAGATGATCTCTCAACTTTAACTGGAAAGAATGTCTTGATTGTTGAAGATATAATTGAC
ACTGGTAAACAAATGCAAACTTTGCTTTCCCTGGTTAAGCAGTACAGCCCCAAATGGTTAAGGTTGAAGCTTGCTGGTGAAGAGGACCTCTGAAGTGTGGATACAGGC
CAGACTTTGTTGGATTGAAATTCAGACAAGTTGTTGTTGGATATGCCCTTGACTATAATGAGTACTTCAGGGATTGAATCACGTTTGTGTCTTATGAACTGGAAA
AGCCAAATACAAAGCCTAAGATGAGCGCAAGTTGAATCTGCAAAATACGAGGAGTCTGTTGATGTTGCCAGTAAAATTAGCAGGTGTTCTAGTCTGTGGCATCTGCCTAG
TAAAGCTTTTGCATGAACCTTCTATGAATGTTACTGTTTTATTTTAAAGATGTCAGTTGCTGCGTCCCAGACTTTTGATTGCACTATGAGCCTATAGGCCAGCCTACC
CTCTGGTAGATTGTCTGTTATCTTGAAGAAAAACAACTCTTAAATTAACACTTTTAAATAATAACTGAGATTGTATCTGTAAGAAGGATTTAAAGAGAAGCTATATT
AGTTTTTTAATTGGTATTTTAAATTTTATATATTCAGGAGAGAAAGATGTGATTGATATTGTTAATTTAGACGAGTCTGAAGCTCTCGATTTCCTATCAGTAACAGCATCTA
AGAGGTTTTGCTCAGTGAATAAACATGTTTACGAGTGTGGCTGTATTTCCCACTTTCAGTAAATCGTTGTCAACAGTTCCTTTTAAATGCAAAATAAATAAATCTAA
AATTC

Figure S1.

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
1799	1	1804	1804	100.0%	7	-	5566787	5570232	3446
1554	1	1801	1804	93.3%	5	+	77080627	77082421	1795
1446	1	1804	1804	90.9%	18	+	60109157	60110942	1786
1442	61	1804	1804	92.5%	1	-	224050852	224052577	1726
1441	77	1804	1804	92.2%	2	+	132021121	132022843	1723
1433	77	1804	1804	92.0%	2	-	131219802	131221524	1723
1430	1	1798	1804	90.7%	6	-	88985028	88986792	1765
1429	77	1804	1804	91.9%	2	+	131414315	131416037	1723
1423	77	1804	1804	92.3%	2	-	130831230	130832952	1723
1417	77	1804	1804	91.6%	2	+	132383704	132385425	1722
1415	24	1798	1804	91.0%	15	-	44280683	44282443	1761
1408	61	1804	1804	92.2%	22	-	41469582	41471267	1686
1407	77	1804	1804	91.1%	14	+	19584790	19586512	1723
1400	77	1804	1804	91.2%	14	-	19987124	19988846	1723
1394	77	1804	1804	91.0%	22	-	16254765	16256484	1720
1394	61	1804	1804	91.2%	9	+	6834432	6836161	1730
1375	21	1804	1804	90.9%	9	+	103493457	103495566	2110
1340	1	1802	1804	90.0%	5	+	97674364	97676110	1747
1311	20	1804	1804	90.8%	5	-	79594919	79598731	3813
1309	52	1804	1804	89.5%	11	-	1823104	1825405	2302
1293	109	1804	1804	90.0%	5	-	130994138	130995784	1647
1293	9	1804	1804	89.6%	3	-	175694709	175696448	1740
1279	28	1804	1804	89.3%	X	-	46146845	46148620	1776
1253	8	1804	1804	90.8%	1	-	78238666	78240600	1935
1240	28	1804	1804	88.5%	10	+	70782474	70784663	2190
1145	214	1800	1804	89.8%	15	-	34664450	35085764	421315
1110	61	1804	1804	88.1%	3	+	180538308	180540006	1699
1030	21	1804	1804	87.4%	19	+	9628762	9630853	2092
940	77	1804	1804	84.3%	8	+	85860820	85862503	1684
868	97	1219	1804	92.9%	17	-	79477709	79479368	1660
787	101	1205	1804	87.4%	6	+	46172731	46173831	1101
774	103	1211	1804	85.0%	Y	-	19868882	19869987	1106
765	123	1211	1804	86.5%	3	+	12111763	12112862	1100
710	194	1211	1804	86.2%	3	+	139212765	139213785	1021
708	202	1211	1804	85.7%	Y	+	20309887	20310893	1007

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
707	175	1211	1804	85.6%	X	-	53172007	53173004	998
686	175	1211	1804	84.8%	5	+	15450887	15451942	1056
680	205	1212	1804	88.4%	1	-	229567246	229568630	1385
675	154	1172	1804	83.2%	5	-	56777444	56778462	1019
660	115	1211	1804	84.9%	1	+	104112205	104113295	1091
646	245	1220	1804	83.5%	20	+	1141388	1142349	962
640	154	1211	1804	82.9%	16	-	50092062	50093073	1012
599	325	1208	1804	86.0%	17	+	17145337	17146201	865
595	195	1211	1804	81.4%	15	-	83395060	83396068	1009
562	121	1169	1804	80.2%	2	+	57992124	57993148	1025
557	175	1211	1804	80.6%	5	-	61127772	61128781	1010
535	337	1211	1804	86.7%	2	+	74135800	74146701	10902
507	200	1169	1804	82.5%	1	+	202842535	202843411	877
485	154	840	1804	88.2%	1	+	238090200	238090890	691
473	1111	1804	1804	89.7%	2	-	177490945	177491595	651
470	1234	1804	1804	93.2%	X	+	44653249	44654153	905
423	253	1156	1804	80.9%	6	-	101878295	101879166	872
385	337	1211	1804	86.2%	10	-	90694981	90703664	8684
370	1382	1799	1804	94.2%	6	+	146925279	146925692	414
355	743	1211	1804	88.0%	1	+	27651821	27652286	466
329	1366	1804	1804	91.5%	7	+	152462057	152462864	808
322	790	1211	1804	88.2%	1	-	92694571	92694992	422
319	1366	1804	1804	91.3%	7	-	149624173	149624972	800
307	1366	1804	1804	88.9%	18	+	57360280	57360708	429
293	1407	1804	1804	88.8%	1	-	142670565	142670960	396
293	1407	1804	1804	88.8%	1	+	143420546	143420941	396
291	1407	1804	1804	88.6%	1	-	143167020	143167415	396
279	1449	1804	1804	89.3%	4	-	49214646	49214997	352
259	1449	1779	1804	89.4%	4	+	49581493	49581819	327
249	1487	1801	1804	93.1%	9	-	110196946	110197567	622

Figure S2.

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
1884	1	1889	1889	100.0%	5	-	143664795	143668403	3609
1578	6	1879	1889	94.9%	16	+	37882559	37884341	1783
1435	4	1889	1889	92.0%	X	-	148924338	148926195	1858
1432	4	1889	1889	92.0%	X	+	148867228	148869085	1858
1352	79	1874	1889	90.2%	X	-	139617282	139619085	1804
1248	170	1889	1889	90.2%	17	-	8758031	8759716	1686
1054	123	1874	1889	85.9%	X	+	142529941	142531911	1971
865	105	1215	1889	89.7%	4	+	103236029	103237142	1114
848	105	1215	1889	88.7%	13	+	81203705	81204810	1106
843	105	1215	1889	88.7%	8	+	47314369	47315482	1114
797	105	1215	1889	90.8%	11	-	120207696	120209383	1688
769	105	1215	1889	89.0%	16	-	64917797	64918909	1113
747	105	1215	1889	87.8%	6	-	133711455	133713071	1617
697	197	1215	1889	85.6%	X	-	48572046	48573053	1008
696	110	1210	1889	88.6%	8	-	126415906	126417561	1656
669	197	1182	1889	88.7%	9	+	38881804	38883298	1495
650	197	1179	1889	88.2%	7	-	5079020	5080869	1850
627	239	1182	1889	85.2%	11	-	12836862	12837792	931
608	149	1161	1889	82.2%	13	+	112045413	112046425	1013
603	1169	1877	1889	95.5%	15	-	24786928	24787658	731
580	200	1205	1889	87.5%	2	-	113873204	113876352	3149
580	252	1127	1889	85.0%	10	+	24773337	24774486	1150
539	330	1206	1889	87.3%	19	-	34316158	34323026	6869
523	1076	1868	1889	89.2%	12	+	63166551	63167380	830
514	67	945	1889	81.4%	X	+	4442681	4722836	280156
511	67	945	1889	81.3%	X	+	3744346	4217107	472762
509	67	945	1889	81.2%	X	+	31828510	31829389	880
500	200	1157	1889	87.5%	6	-	83463029	83476931	13903
483	231	1185	1889	79.6%	6	-	10121809	10122713	905
482	172	945	1889	83.3%	X	+	30594661	30595436	776
478	172	945	1889	83.0%	X	+	30728974	30729749	776
476	172	945	1889	82.9%	X	+	30974148	30974923	776
471	172	945	1889	83.3%	X	-	31353875	31756103	402229
466	172	945	1889	82.8%	X	-	3246498	3247274	777
462	172	945	1889	82.7%	X	-	3466211	3466986	776
447	172	945	1889	83.2%	X	-	31880375	31881150	776
447	172	945	1889	83.2%	X	-	30295910	30296685	776
447	172	945	1889	83.2%	X	-	3670314	3671089	776

continued

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
447	172	945	1889	83.2%	X	-	3670314	3671089	776
447	172	945	1889	83.2%	X	+	3407017	3407792	776
444	172	945	1889	83.1%	X	+	4216332	4595304	378973
432	172	945	1889	82.8%	X	+	30241827	30242602	776
422	172	945	1889	83.3%	X	-	32134595	32135370	776
422	224	945	1889	82.1%	X	+	5242634	5243343	710
402	172	945	1889	83.3%	X	-	3998446	3999221	776
382	431	1215	1889	84.9%	X	-	75927536	75928252	717
372	116	1068	1889	79.6%	18	-	31679491	31680401	911
369	534	1056	1889	86.1%	12	-	63166034	63166515	482
369	323	945	1889	82.9%	X	+	3903537	4072084	168548
368	67	945	1889	83.7%	X	+	31127180	31128059	880
368	339	1165	1889	82.8%	6	+	107831312	107839108	7797
359	323	945	1889	83.8%	X	+	4071462	4443560	372099
351	1367	1889	1889	86.6%	3	-	31564067	31564579	513
348	654	1215	1889	84.1%	17	+	63696793	63697354	562
334	292	1164	1889	83.8%	17	-	92771395	92772258	864
333	1365	1889	1889	90.5%	3	-	49234097	49239898	5802
325	58	729	1889	78.1%	X	-	36466144	36466809	666
322	313	1583	1889	83.6%	9	+	65212848	65214038	1191
320	1369	1877	1889	89.1%	3	-	45677606	45683486	5881
320	1367	1889	1889	86.1%	3	+	28810058	28810570	513
316	403	945	1889	84.6%	X	+	3744681	3745225	545
275	24	625	1889	89.4%	17	-	63696269	63696870	602
264	770	1186	1889	87.9%	8	-	67617505	67617924	420
254	609	1156	1889	83.0%	11	+	107052406	107052918	513
248	172	652	1889	83.1%	X	-	31354159	31354650	492
242	1394	1874	1889	84.1%	3	-	44890744	44891071	328
216	1621	1878	1889	96.2%	1	-	145409110	145409371	262
216	586	945	1889	82.5%	X	+	3903810	3904159	350
206	249	796	1889	80.5%	11	+	35809702	35810295	594
202	209	660	1889	85.4%	13	+	51440365	51440963	599

Figure S3:

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
1302	1	1310	1310	100.0%	12	+	6643657	6647536	3880
1180	27	1310	1310	95.8%	X	-	39646186	39647466	1281
1109	27	1310	1310	93.3%	X	-	46298948	46300220	1273
1077	27	1310	1310	91.8%	6	-	80662524	80663799	1276
1022	29	1310	1310	91.4%	15	-	64820668	64821939	1272
1021	27	1307	1310	90.0%	13	+	29880844	29882111	1268
1014	83	1310	1310	91.9%	6	+	166477759	166478989	1231
1014	29	1310	1310	91.3%	5	+	173940236	173941515	1280
1006	27	1310	1310	89.9%	5	-	159377299	159378576	1278
1001	27	1295	1310	90.4%	19	+	47061817	47063085	1269
998	27	1305	1310	91.1%	6	+	84101825	84103323	1499
992	27	1305	1310	90.1%	4	+	88128172	88129436	1265
989	27	1295	1310	90.6%	16	-	28250820	28252048	1229
977	96	1310	1310	90.7%	11	-	88141146	88142398	1253
956	27	1307	1310	89.0%	1	-	117256257	117257528	1272
950	31	1302	1310	88.7%	1	-	120004458	120139869	135412
927	83	1272	1310	89.5%	20	+	13373312	13374485	1174
920	27	1307	1310	88.6%	7	+	9654538	9655816	1279
905	27	1212	1310	91.0%	12	+	63148933	63150512	1580
903	96	1310	1310	88.3%	10	-	57426899	57428117	1219
878	96	1295	1310	88.9%	8	-	101562781	101563948	1168
876	83	1310	1310	87.6%	2	+	38512533	38513736	1204
874	144	1278	1310	90.3%	6	-	135940134	135941268	1135
836	53	1293	1310	85.3%	1	-	120076188	120139823	63636
832	28	1295	1310	84.7%	1	+	215044007	215045249	1243
831	27	1286	1310	85.6%	22	-	41069312	41070526	1215
817	97	1295	1310	86.2%	15	+	44355703	44648084	292382
801	141	1255	1310	89.1%	1	-	92580209	92581324	1116
793	151	1223	1310	87.1%	10	-	93426422	93427491	1070
788	134	1267	1310	87.2%	18	-	3977496	3978584	1089
769	83	1261	1310	85.0%	3	-	143221964	143223135	1172
758	130	1278	1310	85.8%	1	+	94767621	94768751	1131
743	127	1302	1310	87.8%	1	-	52172598	52173794	1197
741	160	1302	1310	85.1%	9	+	103738007	103739132	1126
740	149	1309	1310	83.6%	6	+	70455808	70456947	1140
724	163	1310	1310	85.2%	12	+	7719019	7720174	1156
721	150	1267	1310	82.8%	2	-	3735418	3736530	1113
719	157	1302	1310	86.5%	1	-	32867509	32868662	1154

continued

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
716	27	1295	1310	86.0%	21	+	30594575	30595774	1200
715	28	1256	1310	87.5%	19	+	11695989	11698404	2416
682	29	1295	1310	86.0%	8	-	97080762	97081959	1198
651	112	1309	1310	85.6%	6	-	58293686	58294939	1254
637	162	1295	1310	86.5%	X	+	135160061	135160988	928
633	31	1309	1310	86.6%	6	-	57686690	57688014	1325
619	27	1169	1310	83.1%	Y	+	21489384	21490516	1133
613	27	1199	1310	83.6%	1	-	189101395	189102494	1100
603	150	1295	1310	84.8%	3	+	138496714	138497841	1128
586	96	961	1310	85.7%	1	-	120004807	120039471	34665
508	153	1043	1310	80.2%	3	-	89096660	89097542	883
497	122	1128	1310	84.4%	6	-	5817636	5818586	951
497	122	1196	1310	82.6%	13	+	21932525	21933541	1017
474	412	1307	1310	87.1%	1	-	120101175	120102090	916
460	96	766	1310	85.4%	1	-	119977082	119977731	650
443	617	1302	1310	88.1%	5	+	35081321	35082293	973
410	749	1295	1310	88.6%	10	-	15135073	15135612	540
408	812	1310	1310	90.9%	2	+	188280277	188280774	498
362	861	1308	1310	91.4%	4	-	15494147	15494597	451
338	197	1289	1310	80.8%	Y	+	23023315	23024408	1094
324	280	1157	1310	81.8%	4	-	131424463	131425328	866
316	151	662	1310	82.1%	15	+	44646939	44647442	504
308	440	1178	1310	81.8%	3	+	179930237	179930972	736
290	471	894	1310	84.4%	1	-	120076588	120077008	421
289	904	1302	1310	86.4%	1	-	120138597	120138996	400
278	980	1310	1310	92.2%	13	-	45697946	45698277	332
275	29	493	1310	83.8%	13	-	99842870	99843315	446
237	1002	1310	1310	90.4%	X	-	86680390	86680694	305
234	1017	1310	1310	90.1%	X	-	15388930	15389224	295
208	30	1295	1310	83.1%	3	+	141443152	141443571	420

Figure S4:

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
1229	1	1232	1232	100.0%	7	+	59002017	59003250	1234
1228	1	1231	1232	100.0%	4	-	91518178	91519410	1233
1227	1	1232	1232	99.9%	X	-	77439008	77440241	1234
1227	1	1231	1232	99.9%	2	-	28788949	28790179	1231
1227	1	1231	1232	99.9%	11	-	20234836	20236066	1231
1227	1	1231	1232	99.9%	5	+	96290683	96291913	1231
1226	1	1231	1232	99.9%	4	-	57383644	57384876	1233
1226	1	1231	1232	99.9%	15	+	12504594	12505826	1233
1223	1	1232	1232	99.7%	11	+	26686583	26687816	1234
1220	1	1231	1232	99.6%	3	+	142625525	142626757	1233
1219	1	1230	1232	99.6%	6	-	83824193	83825424	1232
1219	1	1232	1232	99.6%	Y_random	+	25801587	25802820	1234
1218	1	1231	1232	99.6%	8	-	95239645	95240877	1233
1217	12	1229	1232	100.0%	5	-	106291541	106292760	1220
1216	1	1231	1232	99.5%	2	-	11330476	11331708	1233
1215	1	1232	1232	99.4%	3	+	139312293	139313526	1234
1214	1	1231	1232	99.4%	8	+	60029049	60030281	1233
1212	1	1231	1232	99.3%	14	-	12113265	12114497	1233
1211	1	1232	1232	99.2%	5	-	13042251	13043484	1234
1211	11	1232	1232	99.6%	3	-	121932931	121934154	1224
1209	8	1231	1232	99.5%	8	-	77929236	77930461	1226
1201	1	1232	1232	98.9%	6	-	128151984	128153301	1318
1199	1	1232	1232	98.7%	1	-	15258698	15259928	1231
1199	1	1232	1232	98.8%	15	+	4835070	4836303	1234
1198	1	1231	1232	98.8%	7	-	107909138	107910370	1233
1198	11	1231	1232	99.1%	3	+	52416717	52417939	1223
1197	1	1231	1232	98.5%	17	+	14846013	14847240	1228
1194	1	1231	1232	98.6%	3	+	78721130	78722381	1252
1192	1	1232	1232	98.6%	11	-	63125610	63126863	1254
1191	1	1232	1232	98.5%	7	-	49380385	49381623	1239
1190	1	1230	1232	98.5%	15	+	93026976	93028228	1253
1186	12	1232	1232	98.5%	7	+	96756177	96757388	1212
1186	1	1231	1232	98.3%	6	+	91378827	91380059	1233
1185	1	1232	1232	98.2%	2	+	104131681	104132914	1234
1185	1	1230	1232	98.0%	14	+	49028253	49029473	1221
1184	1	1232	1232	98.2%	3	-	61265599	61266833	1235
1180	1	1232	1232	98.1%	14	-	103769503	103770713	1211
1175	1	1232	1232	97.8%	5	+	28865060	28866310	1251
1174	8	1231	1232	98.5%	13	-	99171850	99173081	1232
1165	1	1232	1232	97.7%	12	-	82472997	82474258	1262
1164	1	1232	1232	97.7%	X	+	148379862	148381113	1252
1152	11	1229	1232	97.6%	12	-	71021259	71022518	1260
1152	1	1231	1232	97.6%	19	+	40286822	40288065	1244
1151	1	1232	1232	97.2%	9	-	75541256	75542516	1261
1150	1	1231	1232	97.4%	2	+	149568861	149570099	1239
1143	1	1232	1232	96.7%	X	+	117274269	117275527	1259
1142	1	1232	1232	96.7%	4	+	83750749	83752009	1261
1140	1	1232	1232	96.8%	18	+	54143914	54145183	1270
1139	1	1232	1232	97.8%	2	-	43373802	43374968	1167
1136	12	1232	1232	96.9%	2	+	151199705	151200986	1282
1134	1	1228	1232	96.8%	8	-	47339840	47341082	1243
1133	1	1228	1232	97.1%	9	+	51430799	51432057	1259
1132	1	1232	1232	96.3%	6	+	84762184	84763444	1261
1131	1	1228	1232	97.0%	15	+	84696700	84697947	1248
1124	1	1232	1232	96.3%	1	-	182257100	182258364	1265
1123	1	1228	1232	96.3%	1	+	188784337	188785590	1254
1121	1	1232	1232	96.2%	X	-	85817186	85818448	1263
1121	1	1232	1232	96.0%	4	+	82839842	82841098	1257
1120	1	1232	1232	96.6%	13	-	49900113	49901357	1245
1117	12	1232	1232	96.2%	11	+	99539401	99540650	1250
1116	12	1232	1232	96.3%	8	-	42693822	42695062	1241
1115	1	1226	1232	96.5%	2	-	146580255	146581491	1237
1113	1	1230	1232	95.8%	19	-	25870102	25871347	1246
1112	11	1232	1232	96.4%	X	-	112572299	112573538	1240
1105	1	1232	1232	95.9%	6	+	22653394	22654619	1226
1103	11	1232	1232	96.3%	X	-	12902949	12904185	1237
1101	16	1232	1232	95.6%	11	-	48547853	48549098	1246
1096	1	1228	1232	95.9%	11	-	17873595	17874843	1249
1096	9	1232	1232	96.0%	10	+	39731630	39732841	1212
1094	14	1228	1232	95.4%	9	-	45643814	45645055	1242
1093	12	1232	1232	95.6%	1	-	13775764	13776999	1236
1093	1	1229	1232	95.1%	13	+	94063686	94064902	1217
1091	11	1232	1232	95.2%	9	-	109732688	109733935	1248
1091	11	1231	1232	95.4%	12	-	5596224	5597470	1247
1086	11	1232	1232	95.0%	10	-	13610255	13611499	1245
1085	1	1231	1232	94.5%	9	+	113439125	113440557	1433
1081	17	1228	1232	95.1%	9	-	51810182	51811420	1239
1075	17	1231	1232	95.3%	12	-	13759277	13760505	1229

(continuing to the next page)

continued

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
1072	12	1232	1232	95.5%	12	-	72947095	72948317	1223
1069	1	1232	1232	94.5%	19	+	16352237	16353466	1230
1068	12	1188	1232	95.9%	8	-	37435491	37436726	1236
1068	15	1232	1232	95.2%	10	+	121606393	121607622	1230
1055	16	1232	1232	93.8%	2	-	162807788	162809032	1245
1050	11	1232	1232	93.6%	16	+	64911059	64912307	1249
1047	1	1232	1232	94.4%	18	+	56131890	56133147	1258
1044	12	1227	1232	94.5%	10	+	21008986	21010231	1246
1038	8	1232	1232	94.1%	2	-	161079989	161081236	1248
1034	11	1232	1232	93.1%	14	+	56868866	56870107	1242
1034	13	1229	1232	93.4%	10	+	43541000	43542237	1238
1033	11	1232	1232	93.5%	7	-	80133834	80135067	1234
1033	12	1232	1232	93.1%	12	-	108443078	108444315	1238
1031	17	1232	1232	94.1%	15	-	43048495	43049742	1248
1030	1	1232	1232	93.0%	X	-	101082734	101084001	1268
1027	11	1225	1232	93.2%	7	+	34187798	34189037	1240
1023	1	1232	1232	98.6%	10	-	33528121	33529354	1234
1021	11	1232	1232	93.7%	4	+	41115634	41116824	1191
1019	11	1232	1232	92.1%	7	+	32664019	32665265	1247
1019	27	1221	1232	93.7%	6	+	86234857	86236081	1225
1017	11	1232	1232	92.0%	7	+	33389794	33391040	1247
1016	1	1228	1232	93.6%	19	-	46748576	46749820	1245
1016	11	1217	1232	93.4%	6	+	18094694	18095929	1236
1013	12	1232	1232	92.0%	18	+	9748681	9749925	1245
1012	11	1230	1232	92.6%	11	-	109020300	109021548	1249
1007	8	1232	1232	92.4%	9	+	37493479	37494710	1232
1004	12	1229	1232	91.9%	9	-	49993948	49995189	1242
1004	14	1232	1232	92.2%	7	+	32529172	32529475	730304
1004	1	1164	1232	94.8%	16	+	50111136	50112301	1166
1002	12	1221	1232	92.4%	12	+	50241204	50242434	1231
999	13	1232	1232	92.6%	8	+	131157788	131159026	1239
998	13	1200	1232	93.0%	4	-	128545466	128546678	1213
998	1	1232	1232	93.1%	12	-	50941553	50942788	1236
997	11	1227	1232	91.6%	2	+	11341812	11343063	1252
996	14	1229	1232	92.1%	7	+	33258235	33257703	319469
995	138	1220	1232	96.2%	13	-	98227915	98229017	1103
993	11	1228	1232	91.3%	4	+	118659579	118660823	1245
993	44	1227	1232	92.5%	2	+	65115467	65116681	1215
991	11	1228	1232	91.3%	8	+	78817831	78819076	1246
990	12	1232	1232	91.7%	X	-	149837446	149838651	1206
990	11	1230	1232	92.6%	11	-	33175554	33176777	1224
989	143	1232	1232	95.6%	17	-	25463357	25464472	1116
989	26	1231	1232	91.6%	13	-	45281701	45282930	1230
987	17	1230	1232	91.1%	6	-	13005395	13006631	1237
985	11	1232	1232	93.6%	14	+	48846836	48848086	1251
984	72	1229	1232	93.1%	9	-	83859391	83860601	1211
983	11	1229	1232	92.4%	7	+	33330738	33804602	473865
981	12	1232	1232	92.8%	6	+	94770485	94771715	1231
975	44	1231	1232	91.7%	4	-	63323191	63324420	1230
974	11	1220	1232	91.0%	7	+	32602089	32957914	355826
974	24	1218	1232	91.8%	10	+	126750178	126751385	1208
972	18	1221	1232	92.1%	13	+	69484500	69485716	1217
967	11	1232	1232	91.8%	1	-	184322480	184323706	1227
966	1	1217	1232	91.3%	13	+	12031775	12032998	1224
965	11	1219	1232	93.1%	3	-	69370165	69371417	1253
965	11	1221	1232	90.6%	16	+	49424879	49426119	1241
965	11	1141	1232	93.7%	12	+	15734862	15736013	1152
964	12	1232	1232	90.8%	14	+	41370492	41371733	1242
963	11	1232	1232	89.8%	7	-	74927459	74928707	1249
963	8	1232	1232	93.0%	8	+	83457533	83458668	1136

continued

SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
960	1	1229	1232	90.7%	18	+	48786308	48787563	1256
960	13	1230	1232	91.4%	1	+	23486340	23487568	1229
959	12	1136	1232	94.9%	10	+	22511098	22512186	1089
957	179	1232	1232	95.8%	15	+	18093635	18094718	1084
948	11	1205	1232	91.3%	X	+	44824945	44826165	1221
947	33	1221	1232	91.4%	2	-	16900259	16901476	1218
947	99	1232	1232	92.6%	10	+	10332897	10334049	1153
944	28	1232	1232	90.7%	2	+	45664969	45666172	1204
941	11	1102	1232	94.3%	10	+	44725885	44726976	1092
938	17	1163	1232	92.0%	X	-	96276606	96277775	1170
937	12	1232	1232	90.8%	9	-	100525192	100526422	1231
931	55	1224	1232	91.8%	14	+	33710909	33712109	1201
929	11	1206	1232	91.3%	1	+	48938517	48939746	1230
926	8	1160	1232	91.8%	7	+	37019268	37020416	1149
926	14	1016	1232	97.1%	11	+	3851680	3852683	1004
917	12	1204	1232	89.6%	14	-	59497894	59499104	1211
916	16	1232	1232	91.1%	17	+	41482930	41484148	1219
912	1	1102	1232	93.2%	8	-	41939125	41940431	1307
910	1	1177	1232	89.2%	12	-	62454539	62455711	1173
910	20	1232	1232	88.2%	16	+	74961454	74962697	1244
909	11	1229	1232	89.4%	8	+	66256808	66258048	1241
901	1	1012	1232	95.5%	15	-	10490319	10491329	1011
894	46	1231	1232	89.8%	3	-	6450412	6451593	1182
889	93	1220	1232	91.0%	7	+	33922179	33923319	1141
888	93	1229	1232	91.1%	7	+	32956774	33331982	375209
887	12	1221	1232	90.9%	18	+	23375092	23376308	1217
882	30	1097	1232	92.7%	13	-	20796205	20797272	1068
876	11	1063	1232	92.0%	8	-	100089492	100090532	1041
875	12	1232	1232	90.1%	3	-	46888450	46889695	1246
874	76	1232	1232	94.4%	1	+	12555683	12556670	988
866	233	1232	1232	95.0%	8	-	94212069	94213034	966
866	354	1232	1232	99.4%	17	-	61920156	61921036	881
859	11	1102	1232	89.9%	3	-	8615221	8616544	1324
858	132	1190	1232	91.5%	X	-	109773969	109775034	1066
856	37	1221	1232	89.9%	11	-	79184594	79185749	1156
852	8	948	1232	96.1%	11	+	4463604	4464521	918
844	93	1232	1232	90.9%	7	+	33803455	34098320	294866
824	12	1228	1232	86.8%	14	-	88748989	88750201	1213
809	94	1145	1232	89.8%	7	+	32844085	33331901	487817
797	1	884	1232	96.1%	13	+	17912730	17913627	898
787	350	1232	1232	94.9%	8	-	89183006	89183920	915
784	243	1220	1232	91.4%	13	-	78512060	78513047	988
774	302	1230	1232	92.2%	18	-	50379549	50380470	922
774	12	895	1232	94.2%	15	-	8878309	8879210	902
766	288	1229	1232	91.7%	5	-	108574042	108575004	963
761	43	1221	1232	86.1%	6	+	92511763	92512942	1180
753	8	956	1232	91.5%	X	-	138308903	138309820	918
730	424	1232	1232	95.4%	15	-	75685046	75686278	1233
723	349	1229	1232	92.1%	7	+	32529510	32845230	315721
708	133	1209	1232	85.0%	4	-	9843609	9844706	1098
676	432	1229	1232	93.6%	X	-	136210426	136211641	1216
675	356	1190	1232	91.8%	X	-	141699213	141700068	856
658	11	1177	1232	93.1%	4	-	84569399	84570133	735
631	12	824	1232	90.0%	12	+	102443894	102444706	813
566	73	640	1232	100.0%	6	-	125112656	125113455	800
562	160	1232	1232	90.6%	7	+	32844155	33145105	300951
561	536	1221	1232	92.4%	9	+	52972292	52972985	694
548	659	1231	1232	98.0%	1	-	102178725	102179298	574
484	1	508	1232	97.7%	1	+	102179330	102179837	508

Figure S5

Papers published since the last report by the PI (Dr. Liao) who is supported by the DOD grant:

1. Yang M., Wu J., Wu S-H., Bi A-D., and **Liao D. J.** (2012) Splicing of mouse p53 pre-mRNA does not always follow the “first come, first served” principle and may be influenced by cisplatin treatment and serum starvation. **Mol Biol Rep**, 39:9247–9256
2. Yang, X-R., Wang, X-P., Yao, H-L., Deng, J-X., Jiang, Q-Y., Guo, Y-F., Lan, G-Q., **Liao, D. J.**, Jiang, H-S. (2012) Mitochondrial DNA Polymorphisms are associated with the longevity in the Guangxi Bama Population of China. **Mol. Biol. Report**. 39: 9123–9131.
3. Yuan, C-F., Xu, N-Z., and **Liao, D.J.** (2012) Switch of FANCL, a key FA-BRCA component, between tumor suppressor and promoter by alternative splicing. (Invited commentary), **Cell Cycle**, 11(18): 3-4.
4. Sun, Y., Cao, S-S., Yang, M., Wu, S-H., Wang, Z., Lin, X-K., Song, X-R., and **Liao, D.J.** (2012) Basic anatomy and tumor biology of the RPS6KA6 gene that encodes the p90 ribosomal S6 kinase-4. **Oncogene**, doi: 10.1038/onc.2012.200
5. Sun, Y., Li, Y., Luo, D-Z., and **Liao, D. J.** (2012) Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. **Plos One**, 7(8):e41659. doi:10.1371/journal.pone.0041659.
6. Sun, Y. Luo, D-Z., and **Liao, D.J.** (2012) Cyclin D1 protein plays different roles in modulating chemoresponse in MCF7 and MDA-MB231 cells. **J. Carcinogenesis**, DOI: 10.4103/1477-3163.100401.
7. Melissa J.L. Bonorden, Michael E. Grossmann, Sarah A. Ewing, Olga P. Rogozina, Amitaba Ray, Katai J. Nkhata, **Joshua Liao**, Joseph Grande and M. P. Cleary (2012). Growth and Progression of TRAMP Prostate Tumors in Relationship to Diet and Obesity. **Prostate Cancer**, doi:10.1155/2012/543970.
8. Mizuno NK, Rogozina OP, Seppanen CM, **Liao JD M D Ph**, Cleary MP, Grossmann ME (2013) Combination of intermittent calorie restriction and eicosapentaenoic acid for inhibition of mammary tumors. **Cancer Prev. Res.** (Phila). [Epub ahead of print]
9. Yuan, C-F, Liu Y-M., Yang M., and **DJ. Liao.** (2013) New methods as alternative or corrective measures for the pitfalls and artifacts of reverse transcription and polymerase chain reactions (RT-PCR) in cloning chimeric or antisense-accompanied RNA. **RNA Biology**, E-publ ahead.