**Protein multiplicity can lead to misconduct in western blotting and misinterpretation of immunohistochemical staining results, creating much conflicting data**

Xingde Liu[1,*,#], Yiming Wang[2,#], Wenxiu Yang[3,*], Zhizhong Guan[3,4,*], Wenfeng Yu[4], and D. Joshua Liao[3,4,*]

[#]These authors are co-first authors.

[1]Department of Cardiology Department, Guizhou Medical University Hospital, Guiyang, Guizhou 550004, P.R. China
[2]Department of Psychiatry, Guizhou Medical University Hospital, Guiyang, Guizhou 550004, China.
[3]Department of Pathology, Guizhou Medical University Hospital, Guiyang, Guizhou 550004, P.R. China
[4]Department of Molecular Biology, Guizhou Medical University, Guiyang, Guizhou 550004 P.R. China

*Addresses for correspondence:
Dr. Xingde Liu
Department of Cardiology
Guizhou Medical University Hospital
Guiyang, Guizhou 550004
P.R. China
Email : lxd@gmc.edu.cn

Dr. Wenxiu Yang
Department of Pathology
Guizhou Medical University Hospital
Guiyang, Guizhou 550004
P.R. China
Email: ypq1964@163.com

Dr. Zhizhong Guan
Department of Molecular Biology
Guizhou Medical University
Guiyang, Guizhou 550004
P.R. China
Email: 1457658298@qq.com

Dr. D. Joshua Liao
Pathology Department
Guizhou Medical University Hospital
Guizhou Province 550004
P.R. China
Email: djliao@gmc.edu.cn

**Abstract**

Western blotting (WB) and immunohistochemical staining (IHC) are common techniques for determining tissue protein expression. Both techniques require a primary antibody specific for the protein in question. WB data is band(s) on a membrane while IHC result is a staining on a tissue section. Most human genes are known to produce multiple protein isoforms; in agreement with that, multiple bands are often found on the WB membrane. However, a common but unspoken practice in WB is to cut away the extra band(s) and present for publication only the band of interest, which implies to the readers that only one form of protein is expressed and thus the data interpretation is straightforward. Similarly, few IHC studies discuss whether the antibody used is isoform-specific and whether the positive staining is derived from only one isoform. Currently, there is no reliable technique to determine the isoform-specificity of an antibody, especially for IHC. Therefore, cutting away extra band(s) on the membrane usually is a form of misconduct in WB, and a positive staining in IHC only indicates the presence of protein product(s) of the to-be-interrogated gene, and not necessarily the presence of the isoform of interest. We suggest that data of WB and IHC involving only one antibody should not be published and that relevant reports should discuss whether there may be protein multiplicity and whether the antibody used is isoform-specific. Hopefully, techniques will soon emerge that allow determination of not only the presence of protein products of genes but also the isoforms expressed.

**Key words:**
Western blotting, immunohistochemical staining, Protein isoform, Genome

## Introduction

Western blotting (WB), also called immunoblotting, has been a commonly used technique for the determination and semi-quantitation of protein expression in cells or tissues since H. Towbin et al described it in 1979 (Towbin et al. 1992;Towbin et al. 1979). In WB, protein samples, which usually are the soluble components of cell or tissue lysates and are denatured by boiling, are fractioned using electrophoresis in SDS (sodium dodecyl sulfate, sodium salt) containing polyacrylamide gel (SDS-PAGE) and are then transferred onto a membrane, followed by identification of the protein in question using a specific antibody. The name of WB was given by W. Neal Burnette in 1981 (Burnette 1981), because transferring DNA onto a membrane was described by Edwin Southern in 1975 and thus named as Southern blotting (Southern 2015;Southern 1975), and because later transferring RNA onto a membrane was referred to as Northern blotting.

Immunocytochemical staining or immunohistochemical staining (IHC), a century-old technique (Matos et al. 2010;Gosselin et al. 1986), is also very commonly used for the determination of protein expression in isolated cells or in tissues, respectively. In IHC, a tissue usually has been fixed with formalin, embedded in paraffin, and cut to a thin (usually 5-6 μm) section. After the section has been mounted onto a glass slide, deparaffinized and run through many other steps of tissue processing, the protein of interest in the section will be identified using an antibody specific to it. An enzyme-coupled secondary antibody is used to recognize the primary antibody so that the enzyme can oxidize a chemical applied onto the section and convert the chemical to a colored one (so-called staining).

One strength of WB is that a protein is recognized by a specific antibody as a band at a certain position on the membrane. The position is calibrated as a molecular weight of a protein in kilo-Dalton (kD). If a band appears at a position on the membrane too far away from the position expected from the theoretical molecular mass (TMM) of the protein in question, it is often considered as a nonspecific protein. For instance, p53 protein has a TMM of 53 kD and thus a band at a position far from 53 kD, say 73 kD, should be questioned for its authenticity. The biggest weakness of WB is that it does not provide information on the tissue- and subcellular-locations of the identified protein. This is because all the different proteins from many different cells in the same tissue have been mixed together and then stratified via electrophoresis according to their molecular weights before the primary antibody is applied to identify the

specific protein. In contrast, IHC can localize the protein in the tissue and even inside the cell. However, IHC does not allow us to distinguish nonspecific protein(s) from the true one predicted by the TMM of the to-be-interrogated protein. Because of these strengths and weaknesses of these two techniques, they are usually used together, so that one can correct the weakness of the other. For decades, these two methods have been used as a golden pair in biomedical research and have provided us with a huge amount of useful information about functions and mechanisms of many genes by determination of their protein products. However, with the quick accumulation of knowledge of molecular biology in the past couple of decades, we have gained the basic concept that most genes, especially those in the human genome, are expressed to multiple protein isoforms via many mechanisms. We should therefore revisit our WB procedure and our interpretations of IHC data.

**Most human genes utilize many mechanisms to produce multiple protein isoforms**

Most genes in the human genome can use alternative initiation sites and/or alternative termination sites for transcription to produce different RNA transcripts that differ from each other in the 5' or 3' sequence (Fig 1), usually to meet different functional needs in different developmental, physiological and pathological situations. Reliable estimation on how many genes in the human genome can do so is still lacking. The ENCODE project has estimated that transcripts from 65%, or about two-thirds of the human genes, in most cases of two neighboring genes on the same chromosome, form chimeric RNAs (Birney et al. 2007), as exemplified by the TSNAX-DISC1 chimera illustrated in figure 2. The ENCODE project did not give any information of how those chimeras, containing sequences of two neighboring genes, are formed, but we suspect that the majority of them are formed via cis-splicing of a single RNA molecule produced by reading from the upstream gene to the downstream one during transcription (Peng et al. 2015;Yang et al. 2013;Xie et al. 2016). This conjecture actually infers that transcription of nearly two-thirds of the human genes may be able to use an alternative termination site downstream of the annotated one, although some studies show that probably only 6-10% of drosophila genes do so (Dunn et al. 2013;Jungreis et al. 2011).

Over 95% of the genes in the human genome consist of exons and introns, and, as exemplified by the DISC1 gene (Fig 2), 95% of these exon-intron containing genes undergo alternative cis-splicing to produce different mRNA variants in different situations, and even in

the same cell and under the same condition (Jia et al. 2015). Moreover, translation of one single mRNA may use different start codons or stop codons to produce different protein isoforms that differ from each other in their N- or C-terminal region. More complicatedly, one single mRNA may be translated from different open reading frames (ORF) to different proteins, i.e. proteins that are not related to each other, as described in more detail previously with good examples (Jia et al. 2015).

It is worth mentioning that a single complementary DNA (cDNA) inserted into an expression vector may still be translated to different protein isoform(s) (Sun et al. 2013) or even protein(s) completely different from the cDNA-encoded one. This is because the 5' region of the insert and its nearby vector sequence may constitute a new 5' untranslated region (5'UTR), as illustrated in figure 3, which may change the leader (Malys and McCarthy 2011;Laursen et al. 2005), Kozak (Kozak 2005;Kozak 2006;Kozak 2007;Kozak 2007), or Shine-Dalgarno sequence (Beck et al. 2016;Malys and McCarthy 2011;Sugiura 2014) or may make the 5'UTR leaderless (less than 10 nucleotides) ( (Richman et al. 2014;Moll et al. 2002), depending on the translation system used. Sometimes, the cDNA-vector recombination creates a short upstream ORF (uORF) or a completely new ORF or selects a new start codon in the cDNA for translation (Arsenault et al. 2014;Tholen et al. 2014;Sondo et al. 2014;Song et al. 2003;Janes et al. 2012). In general, we still know too little about translation of RNA to protein, as evidenced by the fact that there are still a large number of annotated genes the protein products of which have not yet been identified by a proteomic approach or any other technique (Jia et al. 2015;Ezkurdia et al. 2014;Kim et al. 2014;Wilhelm et al. 2014;Reddy et al. 2015).

Because of the above-described and some other unmentioned mechanisms, in most cases a gene can produce different protein isoforms, occurring more often in different cells or tissues and in different developmental, physiological and pathological situations but also occurring in the same cell or tissue and in the same situation. Unfortunately, a good estimation on how many genes in the human genome can produce how many protein isoforms is still lacking, due to the unavailability of a reliable technology. One of us has recently shown that two-thirds to three-fourths of the proteins from HEK293 human embryonic kidney cells do not migrate in SDS-PAGE as expected from their TMMs, suggesting that these proteins may be additional isoform(s) with their wild type (WT) protein existing somewhere else in the SDS-PAGE. In other words,

two-thirds or three-fourths of the human genes produce at least one additional isoform besides the WT protein (Zhang et al. 2014).

**In WB, presenting only one band can be a form of misconduct and will likely mislead**

Although in most cases a gene is expressed to multiple protein isoforms, most publications reporting WB data present only one single band on the membrane. Certainly, in some cells or tissues or in some situations, the gene of interest is indeed expressed to only a single form of protein. However, much more often there are actually additional band(s) on the membrane, as one of us has shown for estrogen receptor alpha (Bollig-Fischer et al. 2012;Liao et al. 1998), progesterone receptor (Liao et al. 1998), cyclin E (Liao et al. 2000), CDK4 (Sun et al. 2013), RSK4 (Sun et al. 2013), SMARCA-2 (Yang et al. 2011), etc. A common but unspoken practice is that the extra band or bands are assumed to be nonspecific, usually just because they appear at unexpected positions of the SDS-PAGE, and thus are cut away to simplify the data interpretation. The extra band(s) are still removed even when they are also recognized by an additional antibody that targets a different region of the protein sequence. It goes without saying that presenting only the band of interest implies to the readers that only this form of protein is expressed in the tissue.

In many other cases, although there is only a single band on the membrane, it is not because the gene is expressed only to a single form of protein, but, rather, it is because the antibody used can recognize only this protein isoform. Indeed, probably all antibody producing companies have received numerous complaints for selling antibodies which are “not specific enough”. To avoid such allegations, companies try hard to select and market those antibodies that only recognize a single band on WB, while researchers also choose such “more specific” antibodies, making them dominant in the literature. This alliance between antibody suppliers and researchers will likely serve as a “natural selection” to extinguish those antibodies that can react to more isoforms, and thus have a profound but adverse impact on biomedical research. In our meditation, many of those antibodies that give rise to multiple bands on WB may not be bad ones and may provide us with a better global picture of the gene in question.

**IHC data is often partly misinterpreted, as it does not address protein multiplicity**

Realizing the importance of antibody specificity, many researchers use WB data to endorse the specificity of the antibody used in IHC. However, in most cases WB detects denatured

proteins as it is often carried out using SDS-PAGE after boiling of the protein samples, whereas in IHC proteins usually remain in a conformation close to a native status, due to a quick fixation by formalin. Because of this difference between the two techniques, the same protein may or may not have the same conformation in SDS-PAGE and in the paraffin section that is usually used in IHC, and thus the same antibody may or may not recognize the same protein isoform(s) in WB and in IHC. In other words, WB data may not be able to endorse the antibody specificity for IHC. Sometimes, WB is performed using a native polyacrylamide gel in the absence of SDS and without boiling of the protein sample. This so-called "native WB" usually gives rise to more bands on the membrane than denatured WB, according to our experience, making the resultant data more difficult to interpret. Actually, often, even though the primary antibody has already detected multiple bands in WB, it is still used for IHC. During preparation of this perspective article, we randomly inquired of peers, most of them in the US, whether, when performing IHC, they knew how many protein isoforms their target genes might produce and what similarity and disparity were among these isoforms in terms of protein sequence. Unfortunately, most of them did not have such information albeit they kept publishing IHC data. Some researchers may know that the to-be-interrogated gene can produce multiple proteins, but they assume, without any solid proof, that the antibody used recognizes solely the isoform of their interest, usually the WT form, and thus comfortably ascribe all the observed functions of the gene solely to it without discussing any possible contribution from other isoforms.

In our opinion, a good technical approach is still needed to determine isoform-specificity of antibodies for IHC. Use of the same synthetic peptide that was used to immunize animals for antibody development to pre-block the to-be-stained section is a relatively good test for the antibody specificity. However, this approach still has weaknesses, in part because the short peptide may have different conformations, and thus may pre-block different isoform(s), when it is applied onto the section and when it is administered into an animal to immunize it. Therefore, it is still impossible to completely interpret correctly IHC data produced by most commercially available antibodies. In a well-conducted IHC with all proper controls, properly unmasked antigens, and optimal concentrations of the primary and secondary antibodies, a positive staining can be interpreted as the presence of protein product(s) of the gene in question, and any conclusion beyond this may mislead. Only for those well-characterized protein isoforms of those well-studied genes with those primary antibodies well-confirmed for isoform-specificity, should

a positive staining mean the presence of the isoform of interest. Unfortunately, there are very few such genes among the 20,000 genes in the human genome (Jia et al. 2015;Zhang et al. 2014), while for the vast majority of the human genes such primary antibodies are lacking. This unsolved technical bottleneck greatly diminishes the power of IHC in the exploration of functions and mechanisms of genes, meaning that many studies involving IHC may partly misinterpret, usually overgeneralizing, the data by leaving out the discussion of protein multiplicity.

**Conflicting data on functions of genes are omnipresent, partly due to protein multiplicity**

It had been known decades ago that some genes produced more than one protein isoform, but then these genes were considered exceptions from the majority in the human genome. The technique of cloning and amplifying DNA in bacteria emerged in 1973 (Cohen et al. 1973;Cohen 2013). Because a new technology for the development of monoclonal antibodies emerged in 1975 (Alkan 2004;Liu 2014;Kohler and Milstein 1975), even before the emergence of WB, the uses of IHC and WB have been swiftly spread to every newly cloned gene in the past 40 years or so. Owing to advances in these technologies, our knowledge about functions and mechanisms of genes has quickly accumulated, especially since DNA sequencing technology had been dramatically improved and widely used over the last two decades. However, most of those genes that have been well characterized for their functions without much conflict in the literature are those that are expressed to a dominant protein, usually annotated as the WT form, in terms of both abundance (expression level) and function. For example, the human insulin gene (Gene ID: 3630) only has one form of protein listed in the NCBI (National Center for Biotechnology Information of the US), albeit it is expressed to four mRNA variants that differ from each other at the 5'UTR. Likely, the insulin gene is mainly controlled at the levels of transcription and translation, but is not controlled via protein multiplicity. Even if there may exist some other protein isoform(s) that have not yet been identified, their expression levels are likely low and their functions may be easily overridden by the WT insulin. For this type of gene, WB and IHC have provided us with useful detail about their functions and mechanisms with relatively little confusion. However, it is unimpeachable that perplexity about functions of more and more genes had also started to soar two decades ago, and there now exist, ubiquitously, contradictory data and ensuing bafflement, on genes' functions in the biomedical literature. For most genes, there is

a plethora of data supporting one function but concomitantly there also is plentiful data opposing it. Because of these omnipresent pros and cons, saying "on one hand… but on the other hand…" has become a standard and safe way of describing functions of genes.

The reasons for why there are only as few as 20,000 genes in the human genome are multiple, including mRNA and protein multiplicities that allow the genome to be much smaller than our previous expectation but in the meantime require that most genes have many different and even opposite functions, as one of us has previously expounded (Lou et al. 2014;Yuan et al. 2012;Jia et al. 2015;Wang et al. 2016). For instance, the short isoform of the Bcl-X gene, i.e. Bcl-xS, functions to induce cell death whereas the longer isoform, i.e. Bcl-xL, functions to sustain cell survival (Yuan et al. 2012). In contrast, a shorter isoform of the FANCL gene enhances cell survival while a longer one prods cell death (Yuan et al. 2012;Zhang et al. 2010). Even different subcellular locations of a protein may have opposite functions. For example, in its location at the inner membrane of mitochondria, cytochrome c functions to power the cell and thus to sustain the cell's life by participating in ATP production, but, when it relocates to the cytosol in a stressed situation, it triggers stress-induced cell death that is widely mistaken as apoptosis (Liao and Dickson 2003;Liao 2005;Liu et al. 2013;Zhang et al. 2014;Wang et al. 2016).

In our rumination, the ubiquitous existence of conflicting data, and the ensuing bewilderment, on genes' functions may be largely attributed to our misconduct of WB and our overgeneralized interpretation on IHC data. In some studies some isoforms are detected while in other studies other isoforms with different or even opposite functions are detected. Conclusions should be drawn on the isoforms detected and should not be drawn onto the gene, but, unfortunately, the "shouldn't" has occurred in too many publications. Indeed, many researchers regard the functions of the isoform(s) detected in their systems as the functions of the gene. Because for many genes their WT proteins are not absolutely dominant over other isoforms, in terms of the abundance and function, the heterogeneous expression of, and their proportions among, different isoforms in different cells or tissues or in different situations greatly contribute to the ubiquitously existing pros and cons.

**Concluding remarks**

It may be fine to report WB data as a narrow band in publications to save space, but information should be provided about whether there are additional bands on the membrane and

whether the antibody used is isoform-specific. Unfortunately, cutting away of the additional band(s) on the membrane and presenting only a single band without providing such information have become a common practice in WB, and thus is a form of misconduct. A positive IHC staining only indicates the presence of protein product(s) of the gene in question, but which isoform(s) are expressed remains unknown in most cases. Until we are able to determine isoform-specificity of antibodies for IHC, the role of IHC in exploration of functions and mechanisms of genes remains limited. This is because most genes whose functions remain to be determined are those that have multiple protein isoforms but lack an absolutely dominant one. When summarizing a research report, results from IHC should be interpreted with extra caution and information about protein multiplicity and about the isoform-specificity of the antibody used should be provided and discussed. In general, journals should not publish those WB and IHC results using only one antibody or several antibodies that target to the same region or similar regions of the protein sequence. Also importantly, researchers should equip themselves with the knowledge of the protein multiplicity of the to-be-interrogated gene in particular, and the knowledge of antibody epitope in general. Hopefully, the whole biomedical fraternity, i.e. both antibody suppliers and researchers, will realize the value of some antibodies that recognize multiple bands on WB membranes, and hopefully there will soon be new techniques available to overcome some technical bottlenecks and to allow determination of not only whether a gene is expressed but also which isoforms are expressed.

**Acknowledgements:**

We would like to thank Dr. Fred Bogott at the Austin Medical Center, Austin of Minnesota for his excellent English editing of the manuscript. The work is partly supported by a grant from Chinese National Science Foundation (#81160299) to WX Yang.

**References**

Alkan SS. 2004. Monoclonal antibodies: the story of a discovery that revolutionized science and medicine. Nat Rev Immunol 4:153-156.

Arsenault J, Cuijpers SA, Niranjan D, Davletov B. 2014. Unexpected transcellular protein crossover occurs during canonical DNA transfection. J Cell Biochem 115:2047-2054.

Beck HJ, Fleming IM, Janssen GR. 2016. 5'-Terminal AUGs in Escherichia coli mRNAs with Shine-Dalgarno Sequences: Identification and Analysis of Their Roles in Non-Canonical Translation Initiation. PLoS One 11:e0160144.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van CS, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816.

Bollig-Fischer A, Thakur A, Sun Y, Wu J-S, Liao DJ. 2012. The predominant proteins that react to the MC-20 estrogen receptor alpha antibody differ in molecular weight between the mammary gland and uterus in the mouse and rat. Int J Biomed Sci 8:51-63.

Burnette WN. 1981. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate--polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal Biochem 112:195-203.

Cohen SN. 2013. DNA cloning: a personal view after 40 years. Proc Natl Acad Sci U S A 110:15521-15529.

Cohen SN, Chang AC, Boyer HW, Helling RB. 1973. Construction of biologically functional bacterial plasmids in vitro. Proc Natl Acad Sci U S A 70:3240-3244.

Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife 2:e01179-doi: 10.7554/eLife.01179.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, Tress ML. 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet 23:5866-5878.

Gosselin EJ, Cate CC, Pettengill OS, Sorenson GD. 1986. Immunocytochemistry: its evolution and criteria for its application in the study of epon-embedded cells and tissue. Am J Anat 175:135-160.

Janes S, Schmidt U, Ashour GK, Ney N, Concilio S, Zekri M, Caspari T. 2012. Heat induction of a novel Rad9 variant from a cryptic translation initiation site reduces mitotic commitment. J Cell Sci 125:4487-4497.

Jia Y, Chen L, Ma Y, Zhang J, Xu N, Liao DJ. 2015. To Know How a Gene Works, We Need to Redefine It First but then, More Importantly, to Let the Cell Itself Decide How to Transcribe and Process Its RNAs. Int J Biol Sci 11:1413-1423.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res 21:2096-2113.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. 2014. A draft map of the human proteome. Nature 509:575-581.

Kohler G, Milstein C. 1975. Continuous cultures of fused cells secreting antibody of predefined specificity. Nature 256:495-497.

Kozak M. 2005. A second look at cellular mRNA sequences said to function as internal ribosome entry sites. Nucleic Acids Res 33:6593-6602.

Kozak M. 2006. Rethinking some mechanisms invoked to explain translational regulation in eukaryotes. Gene 382:1-11.

Kozak M. 2007. Lessons (not) learned from mistakes about translation. Gene 403:194-203.

Kozak M. 2007. Some thoughts about translational regulation: forward and backward glances. J Cell Biochem 102:280-290.

Laursen BS, Sorensen HP, Mortensen KK, Sperling-Petersen HU. 2005. Initiation of protein synthesis in bacteria. Microbiol Mol Biol Rev 69:101-123.

Liao DJ. 2005. The scavenger cell hypothesis of apoptosis: apoptosis redefined as a process by which a cell in living tissue is destroyed by phagocytosis. Med Hypotheses 65:23-28.

Liao DJ, Dickson RB. 2003. Cell death in MMTV-c-myc transgenic mouse mammary tumors may not be typical apoptosis. Lab Invest 83:1437-1449.

Liao DJ, Natarajan G, Deming SL, Jamerson MH, Johnson M, Chepko G, Dickson RB. 2000. Cell cycle basis for the onset and progression of c-Myc-induced, TGFalpha-enhanced mouse mammary gland carcinogenesis. Oncogene 19:1307-1317.

Liao DZ, Pantazis CG, Hou X, Li SA. 1998. Promotion of estrogen-induced mammary gland carcinogenesis by androgen in the male Noble rat: probable mediation by steroid receptors. Carcinogenesis 19:2173-2180.

Liu B, Xu N, Man Y, Shen H, Avital I, Stojadinovic A, Liao DJ. 2013. Apoptosis in Living Animals Is Assisted by Scavenger Cells and Thus May Not Mainly Go through the Cytochrome C-Caspase Pathway. J Cancer 4:716-723.

Liu JK. 2014. The history of monoclonal antibody development - Progress, remaining challenges and future innovations. Ann Med Surg (Lond) 3:113-116.

Lou X, Zhang J, Liu S, Xu N, Liao DJ. 2014. The other side of the coin: The tumor-suppressive aspect of oncogenes and the oncogenic aspect of tumor-suppressive genes, such as those along the CCND-CDK4/6-RB axis. Cell Cycle 13:1677-1693.

Malys N, McCarthy JE. 2011. Translation initiation: variations in the mechanism can be anticipated. Cell Mol Life Sci 68:991-1003.

Matos LL, Trufelli DC, de Matos MG, da Silva Pinhal MA. 2010. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. Biomark Insights 5:9-20.

Moll I, Grill S, Gualerzi CO, Blasi U. 2002. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Mol Microbiol 43:239-246.

Peng Z, Yuan C, Zellmer L, Liu S, Xu N, Liao DJ. 2015. Hypothesis: Artifacts, Including Spurious Chimeric RNAs with a Short Homologous Sequence, Caused by Consecutive Reverse Transcriptions and Endogenous Random Primers. J Cancer 6:555-567.

Reddy PJ, Ray S, Srivastava S. 2015. The quest of the human proteome and the missing proteins: digging deeper. OMICS 19:276-282.

Richman TR, Rackham O, Filipovska A. 2014. Mitochondria: Unusual features of the mammalian mitoribosome. Int J Biochem Cell Biol 53:115-120.

Sondo E, Scudieri P, Tomati V, Caci E, Mazzone A, Farrugia G, Ravazzolo R, Galietta LJ. 2014. Non-canonical translation start sites in the TMEM16A chloride channel. Biochim Biophys Acta 1838:89-97.

Song L, Mandecki W, Goldman E. 2003. Expression of non-open reading frames isolated from phage display due to translation reinitiation. FASEB J 17:1674-1681.

Southern E. 2015. The early days of blotting. Methods Mol Biol 1312:1-3.

Southern EM. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98:503-517.

Sugiura M. 2014. Plastid mRNA translation. Methods Mol Biol 1132:73-91.

Sun Y, Cao S, Yang M, Wu S, Wang Z, Lin X, Song X, Liao DJ. 2013. Basic anatomy and tumor biology of the RPS6KA6 gene that encodes the p90 ribosomal S6 kinase-4. Oncogene 32:1794-1810.

Sun Y, Lou X, Yang M, Yuan C, Ma L, Xie BK, Wu JM, Yang W, Shen SX, Xu N, Liao DJ. 2013. Cyclin-dependent kinase 4 may be expressed as multiple proteins and have functions that are independent of binding to CCND and RB and occur at the S and G 2/M phases of the cell cycle. Cell Cycle 12:3512-3525.

Tholen M, Hillebrand LE, Tholen S, Sedelmeier O, Arnold SJ, Reinheckel T. 2014. Out-of-frame start codons prevent translation of truncated nucleo-cytosolic cathepsin L in vivo. Nat Commun 5:4931.

Towbin H, Staehelin T, Gordon J. 1979. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A 76:4350-4354.

Towbin H, Staehelin T, Gordon J. 1992. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. 1979. Biotechnology 24:145-149.

Wang G, Chen L, Yu B, Zellmer L, Xu N, Liao DJ. 2016. Learning about the Importance of Mutation Prevention from Curable Cancers and Benign Tumors. J Cancer 7:436-445.

Wilhelm M, Schlegl J, Hahne H, Moghaddas GA, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. 2014. Mass-spectrometry-based draft of the human proteome. Nature 509:582-587.

Xie B, Yang W, Ouyang Y, Chen L, Jiang H, Liao Y, Liao DJ. 2016. Two RNAs or DNAs May Artificially Fuse Together at a Short Homologous Sequence (SHS) during Reverse Transcription or Polymerase Chain Reactions, and Thus Reporting an SHS-Containing Chimeric RNA Requires Extra Caution. PLoS One 11:e0154855.

Yang M, Sun Y, Ma L, Wang C, Wu JM, Bi A, Liao DJ. 2011. Complex alternative splicing of the smarca2 gene suggests the importance of smarca2-B variants. J Cancer 2:386-400.

Yang W, Wu JM, Bi AD, Ou-Yang YC, Shen HH, Chirn GW, Zhou JH, Weiss E, Holman EP, Liao DJ. 2013. Possible Formation of Mitochondrial-RNA Containing Chimeric or Trimeric RNA Implies a Post-Transcriptional and Post-Splicing Mechanism for RNA Fusion. PLoS One 8:e77016-doi: 10.1371/journal.pone.0077016.

Yuan C, Xu N, Liao J. 2012. Switch of FANCL, a key FA-BRCA component, between tumor suppressor and promoter by alternative splicing. Cell Cycle 11:3355-3356.

Zhang J, Lou X, Shen H, Zellmer L, Sun Y, Liu S, Xu N, Liao DJ. 2014. Isoforms of wild type proteins often appear as low molecular weight bands on SDS-PAGE. Biotechnol J 9:1044-1054.

Zhang J, Lou XM, Jin LY, Zhou RJ, Liu SQ, Xu NZ, Liao DJ. 2014. Necrosis, and then stress induced necrosis-like cell death, but not apoptosis, should be the preferred cell death mode for chemotherapy: clearance of a few misconceptions. Oncoscience 1:407-422.

Zhang J, Zhao D, Park HK, Wang H, Dyer RB, Liu W, Klee GG, McNiven MA, Tindall DJ, Molina JR, Fei P. 2010. FAVL elevation in human tumors disrupts Fanconi anemia pathway signaling and promotes genomic instability and tumor growth. J Clin Invest 120:1524-1534.

**Figure legends**

Fig. 1: Illustration of alternative transcription, splicing and translation of a gene. **A**: A gene may have two additional transcription initiation sites (the 2nd and 3rd arrows) besides the common one (the 1st arrow) and two additional transcription termination sites (black dots) besides the common one. The common transcript, usually annotated as the wide type one, consists of five exons. **B**: The three initiation sites and three termination sites may together result in nine RNA transcripts (long arrows). **C**: The nine transcripts may be cis-spliced differently to produce many different mature mRNAs and even non-coding RNAs, several of which are illustrated as examples, such as the one with only exons 1, 4 and 5. Translation of some of these mRNAs may be initiated at an alternative start codon (ATG or CTG) or may be terminated at an alternative stop codon (TAA or TGA), resulting in different protein isoforms. "T" but not "U" is used herein so as to be consistent with the NCBI (National Center for Biotechnology Information of the US) database that presents mRNAs as DNA sequences. These alternative mechanisms at the levels of transcription, splicing and translation allow a gene to produce many different protein isoforms.

Fig. 2: Human TSNAX-DISC1 chimeras as examples of chimeric RNAs formed by two neighboring genes on the same chromosome. **A** and **B**: Images copied from the NCBI database show the DISC1 gene (long red arrow in **A**) with the TSNAX gene (a grey arrow) at its 5' side and the TSNAX-DISC1 chimeric gene (the long red arrow in **B**) on human chromosome 1. There are many other coding or non-coding genes in this chromosomal region indicated by grey arrows. The arrows pointing to the opposite directions indicate that the genes are harbored by, i.e. are transcribed from, the opposite strands of the DNA double helix. **C**: An image copied from the NCBI database shows eight TSNAX-DISC1 chimeric RNAs, one TSNAX RNA, and 23 DISC1 RNAs (in the red circles) besides RNAs of other genes. In each of these RNAs, as illustrated in an enlarged area (from the purple circle), the bars indicate exons while the horizontal lines indicate introns, with the size of the bars and the length of the lines in proportion to their lengths in the number of nucleotides.

Fig 3: Illustration of how the sequence around the joining site between the vector and the 5' end of the cDNA insert may unexpectedly impact the translation of the cDNA. When a cDNA (long black arrow) is inserted into an expression vector (grey bars), the 5' end of the cDNA and its

nearby vector sequence may together create a short uORF or may alter the translation-regulatory element, such as the Kozak or Shine-Dalgarno sequence that may or may not overlap with the uORF. As a result, the translation efficiency may be affected or the translation machinery may skip the annotated ATG and select a different start codon, such as a downstream “atg” or “ctg”.

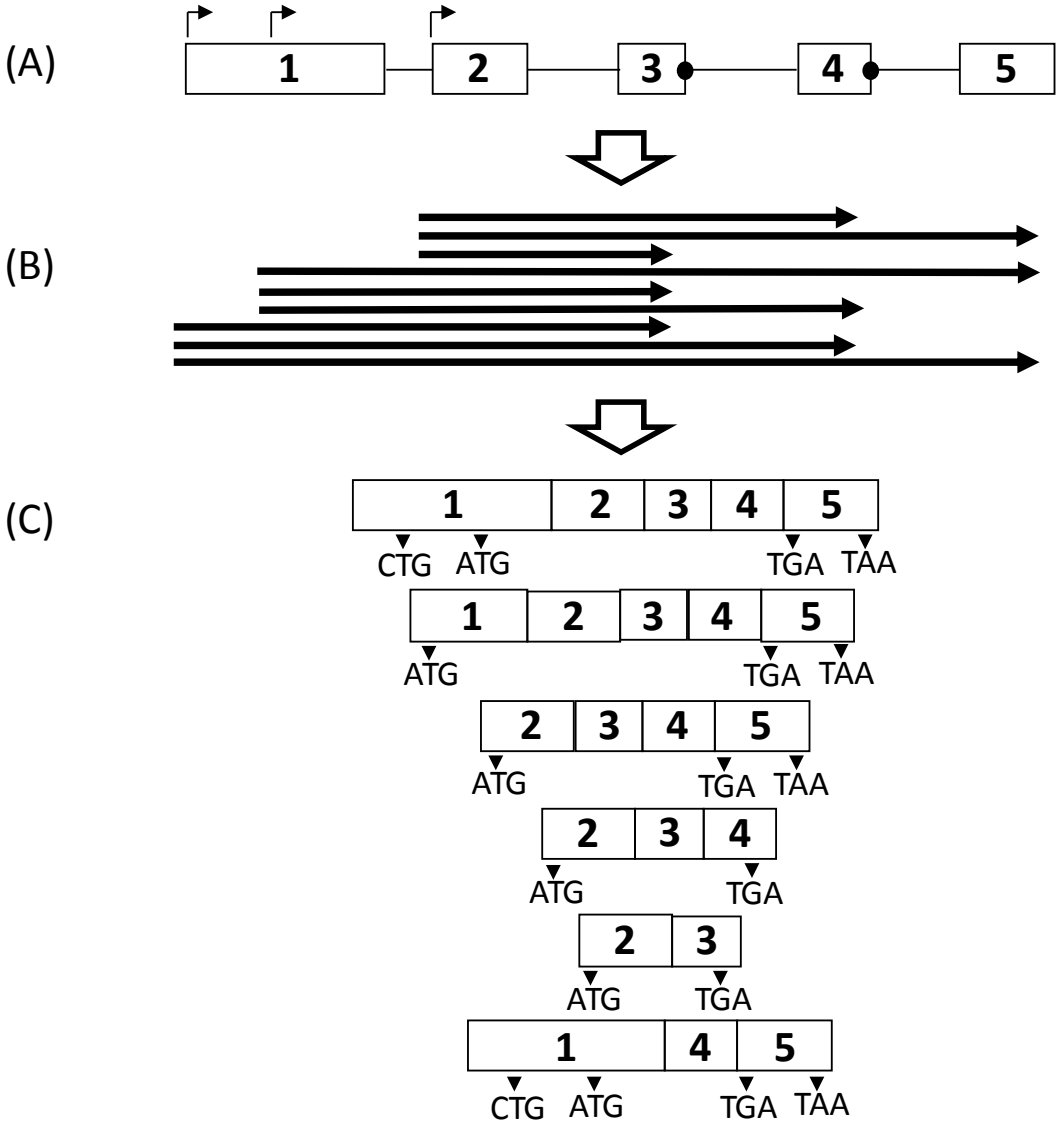
(A)
1
2
3
4
5
(B)
(C)
1
2
3
4
5
CTG
ATG
TGA
TAA
1
2
3
4
5
ATG
TGA
TAA
2
3
4
5
ATG
TGA
TAA
2
3
4
ATG
TGA
2
3
ATG
TGA
1
4
5
CTG
ATG
TGA
TAA

Figure 1

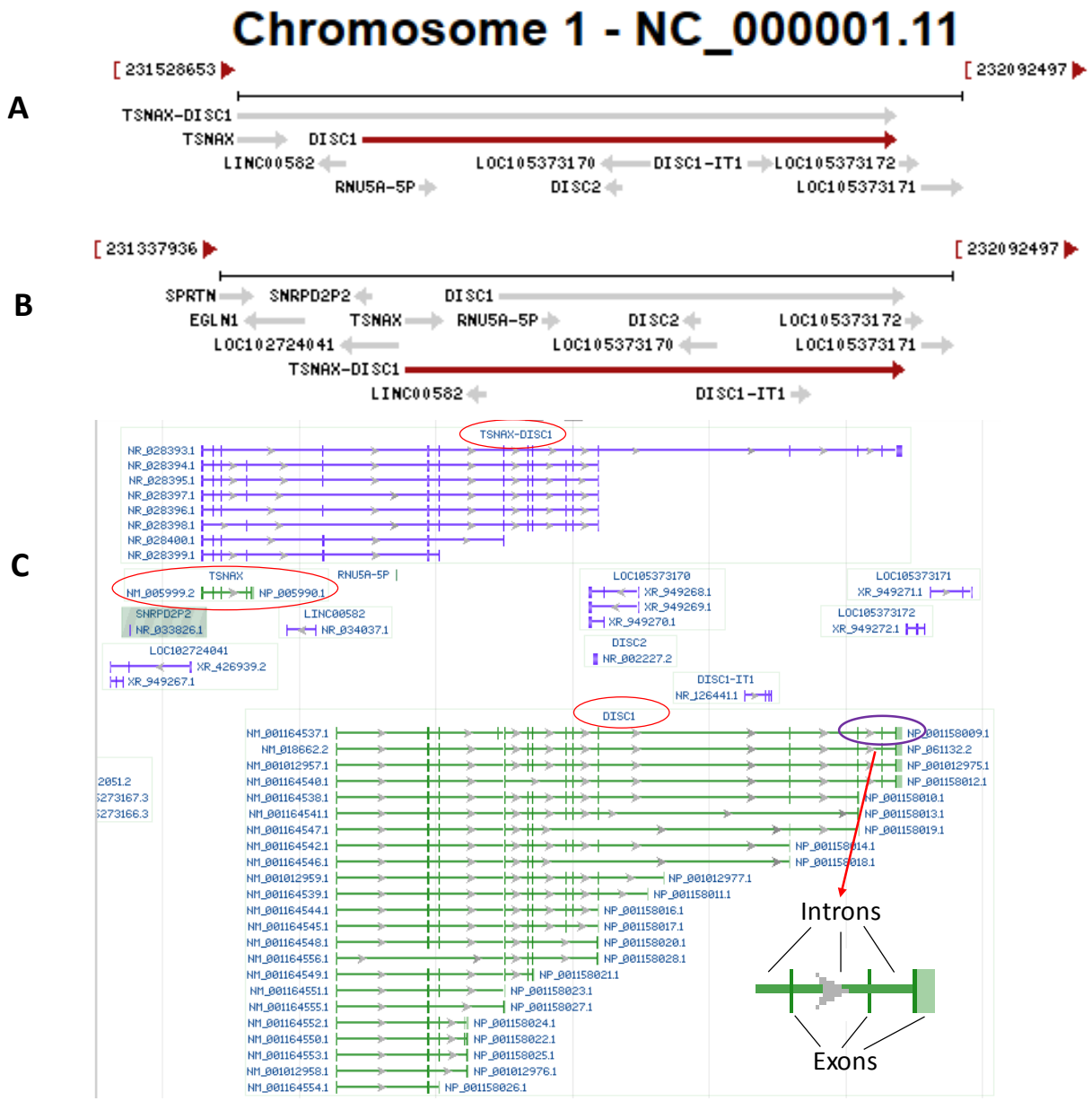
Chromosome 1 - NC_000001.11
A
[ 231528653
[ 232092497
TSNAX-DISC1
TSNAX
DISC1
LINC00582
LOC105373170
DISC1-IT1
LOC105373172
RNU5A-5P
DISC2
LOC105373171
B
[ 231337936
[ 232092497
SPRTN
SNRPD2P2
DISC1
EGLN1
TSNAX
RNU5A-5P
DISC2
LOC105373172
LOC102724041
LOC105373170
LOC105373171
TSNAX-DISC1
LINC00582
DISC1-IT1
C
TSNAX-DISC1
NR_028393.1
NR_028394.1
NR_028395.1
NR_028397.1
NR_028396.1
NR_028398.1
NR_028400.1
NR_028399.1
TSNAX
NM_005999.2
NP_005990.1
RNU5A-5P
SNRPD2P2
NR_033826.1
LINC00582
NR_034037.1
LOC102724041
XR_426939.2
XR_949267.1
LOC105373170
XR_949268.1
XR_949269.1
XR_949270.1
LOC105373171
XR_949271.1
LOC105373172
XR_949272.1
DISC2
NR_002227.2
DISC1-IT1
NR_126441.1
DISC1
Introns
Exons

Figure 2

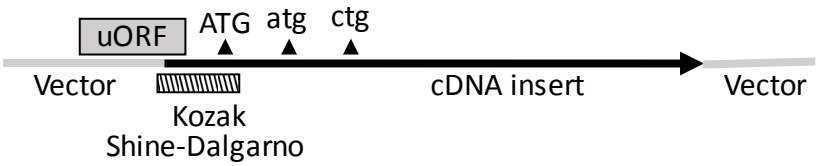
uORF
ATG
atg
ctg
Vector
cDNA insert
Vector
Kozak
Shine-Dalgarno

Figure 3