Gene Annotation Using the Proteome

by

Lulu Wang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2010 by the
Graduate Supervisory Committee:

Douglas Lake, Chair
Yung Chang
Jeffery Touchman

ARIZONA STATE UNIVERSITY

December 2010

ABSTRACT

While the entire human genome has been sequenced, the understanding of its functional elements remains unclear. The Encyclopedia of DNA Elements (ENCODE) project analyzed 1% of the human genome and found that the majority of the human genome is transcribed, including non-protein coding regions. The hypothesis is that some of the “non-coding” sequences are translated into peptides and small proteins. Using mass spectrometry numerous peptides derived from the ENCODE transcriptome were identified. Peptides and small proteins were also found from non-coding regions of the 1% of the human genome that the ENCODE did not find transcripts for. A large portion of these peptides mapped to the intronic regions of known genes, thus it is suspected that they may be undiscovered exons present in alternative spliceoforms of certain genes. Further studies proved the existence of polyadenylated RNAs coding for these peptides. Although their functional significance has not been determined, I anticipate the findings will lead to the discovery of new splice variants of known genes and possibly new transcriptional and translational mechanisms.

ACKNOWLEDGEMENTS

I would like thank firstly my advisor Dr. Douglas Lake. Not only did he provide me the wonderful opportunity to do my research, his being a great example of a scientist has also inspired and motivated me. He has helped me through these years, from project design to experimental setting, from literature search to troubleshooting.

I am in debt to Dr. Yung Chang. Her sharp comments on my experimental results helped clear out a lot of my confusions. She also encouraged me to think independently and creatively, and has given me advice on career choices.

I would like to thank Dr. Jeffery Touchman. He is an expert on my thesis project so he provided me with reasonable explanations when I felt lost. More importantly, his optimism cheered me up when I was discouraged by unsuccessful results.

My labmates helped me a lot both in research and in life. Yvette Ruiz, Shen and Hojoon were the first I knew in the lab, they practically taught me how to do everything. Dr. Antwi Kwasi was among those that initiated the project, and he is the go-to person when I have chemistry questions. I also want to thank Ben, Steve, Paul, Lauren for their help and support.

I am especially grateful to my parents in China. I could never imagine better parents than they are, loving, understanding and supporting unconditionally. I also want to thank my dear husband, Taisong, for together we make a sweet home in the United States.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

All the genetic information of an individual is encrypted in its DNA sequence, hence, whole genome sequencing is the foundation to study genes and their interactions which directly control the growth and development of an organism. It also facilitates the research on infectious diseases (Katalin & Kapranov, 2009), cancer biology (Mardis & Wilson, 2009) and agriculture improvement (Edwards & Batley, 2010). Over the decades genome sequencing technologies have significantly increased throughput and reduced cost (Jiang, Rokhsar, & Harland, 2009). According to the statistical data from NCBI, the genomes of 832 organisms have been completely sequenced and this number will increase to roughly 2000 in the year 2012 (Chain et al., 2009; http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html). The human genome project (HGP), which was launched in 1990, was finished 2 years earlier with a highly accurate sequence of the vast majority of 2.85 billion nucleotides of the human genome (Collins, Lander, Rogers, Waterston, & Consortium, 2004; Lander et al., 2001). The challenge now is how to annotate the functional elements contained in the sequences. Experimental and computational methods have been developed to address the issue, but they are not quite satisfactory. Experimentally cloning all genes at present is both labor intensive and time consuming. Computational recognition programs, though more high-throughput, fail to provide accurate and reliable predictions (Frazer, Elnitski, Church, Dubchak, & Hardison, 2003), especially for eukaryotes because their genes are more complicated with intron-exon structures (Ratsch et al., 2007). In addition, all gene prediction programs are genome-dependent

and need experimental verification (Burge & Karlin, 1997). A thorough Gateway-cloning of all predicted protein-coding open reading frames (ORFs) of *C. elegans* shows that over 50% of the intron-exon structures needed corrections (Reboul et al., 2003). Here, we'll demonstrate the combination of mass spectrometry of tumor cell surface elutions and some theoretical databases of human transcripts may assist gene annotation.

Recent findings by the ENCODE pilot project revealed that at least 93% of the analyzed human genome are transcribed, including non-protein-coding sequences (Birney et al., 2007). The biological functions of non-protein-coding RNAs (ncRNA) are not well known yet, but they may play an important role in gene regulation (Amaral, Dinger, Mercer, & Mattick, 2008). Because the molecular machinery involved in protein synthesis is not fully understood, we propose that some of the ncRNAs may be translated. Using LC-MS/MS we identified numerous peptides encoded by the ENCODE RNAs. Further studies suggest there are polyadenylated RNAs encoding these peptides in some tumor cell lines and normal donor PBMCs. These peptides come from both intergenic and intronic regions of the human genome, suggesting they may be novel genes or undiscovered splice variants of known genes. Conservation of some of these non-exonic peptides suggests that whatever the mechanisms are that these peptides are translated, evolution chose to keep them for a reason. In addition, several non-exonic peptides are derived from different intronic regions of a tumor suppressor gene, suggesting there may be potential biomarkers for cancer diagnosis and even drug targets. In conclusion, with the detectable level of

peptides and small proteins from mass spectrometry and the existence of corresponding mRNA, we believe that non-exonic peptides are a starting point to examine new transcriptional and translational machineries and to identify genes, proteins, regulatory and structural elements.

MATERIALS AND METHODS

Cell lines and Cell Culture

Table 1

Cell Lines Used in the Discovery of ENCODE Peptides

Cell name	ATCC number	Human Cell type
A549	CCL-185	Lung carcinoma
BXPC-3	CRL-1687	Pancreas adnocarcinoma
DU-145	HTB-81	Prostate carcinoma
MCF-7	HTB-22	Breast adenocarcinoma
Panc-1	CRL-1469	Pancreas epitheloid carcinoma
PMBC	NA	Human Peripheral blood mononuclear cells

Tumor cells were grown in DMEM (Cellgro) supplemented with 10% heat-inactivated fetal bovine serum (FBS), 1000 U/ml penicillin-streptomycin, and 2 mM L-glutamine. Peripheral blood mononuclear cells (PBMC) were purified from blood samples of normal donors by density centrifugation with Ficoll Hypaque.

Acid Elution, 3 kDa Filtration and LC-MS/MS

$1x10^8$ cells of each tumor cell line A549 (lung), BXPC-3 (pancreas), DU145 (prostate), MCF-7 (breast), and Panc-1 (pancreas) were harvested with Cellstripper

(Mediatech Inc.), a non-enzymatic cell disassociation solution. Cells were centrifuged at 1000 rpm for 5 minutes and washed in 45 ml phosphate-buffered saline (PBS) for 3 times. The cells were then resuspended in 1 ml citric acid buffer (pH 3.0) for 15 minutes followed by 2-minute centrifugation. The supernatant was collected, immediately filtered through a 0.45 μm filter and further passed through a 3 kDa Microcon Ultracel filter (Millipore, Bedford, MA). The filtrate was transferred to siliconized tubes (VWR Scientific) and stored at -20 ℃ until they were analyzed by liquid chromatography followed by tandem mass spectrometry (LC-MS/MS). The filtrates were resolved on an Agilent 1100 HPLC-Chip Cube Interface (Agilent Technologies). (Antwi et al., 2009)

Construction of the ENCODE and Non-Exonic Databases

The 1% of the human genome which the ENCODE consortium analyzed and found RNA from was bioinformatically translated into amino acid sequences into 6 frames. Stop codons were used to define the endpoints of each peptide with a minimum length of 8 amino acids. This is the ENCODE database. The same 1% that is neither RefSeq nor ENCODE RNA were translated to form a non-exonic database. Both databases were created by Hojoon Lee. Figure 1 is an illustration of how the ENCODE database was generated.

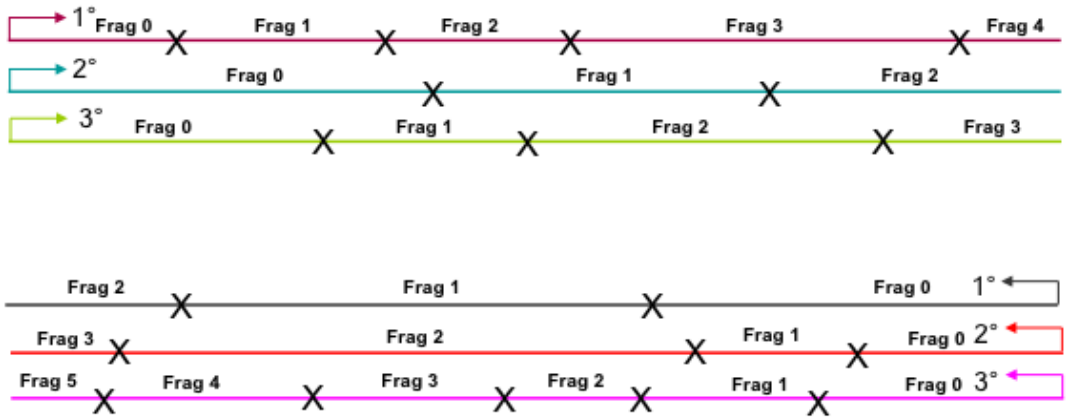


Figure 1. Diagram of how the ENCODE database was bioinformatically constructed. The forward (top) and reverse (bottom) DNA strands were translated into amino acid sequences in 6 frames. Stop codons (X) were used to distinct one peptide from the adjacent ones. Frag 0, Frag 1, Frag 2, etc. represent different transcripts in all frames. The 1% of the human genome (ENCODE) was translated using this method.

Peptide Synthesis

Peptides were chemically synthesized at the proteomics core facility at the Arizona State University on a Milligen 9050 peptide synthesizer (Millipore, Bedford, MA). Mass spectrometric analysis was used to confirm the amino acid sequence. (Antwi, et al., 2009)

mRNA Validation of Peptides

$1x10^7$ cells of each tumor cell line A549 (lung), DU145 (prostate), MCF-7 (breast), and Panc-1 (pancreas) were harvested with Cellstripper (Mediatech Inc.). Total RNA was isolated from each tumor cell line and PBMCs with RNeasy Mini Kit (Qiagen). cDNA was synthesized using oligo-dT_{20} primers and reverse transcriptase III (Invitrogen) and treated with DNase to make sure only polyadenylated RNA were amplified. PCR was performed with primers flanking each peptide in the genomic

region. Transcripts of expected sizes were found for all peptides tested. DNA sequencing confirmed the sequences encoding each peptide. B-actin control was used to exclude the possibility of genomic DNA contamination.

Polyclonal Antibody, Affinity Column

Two peptides, HVTFILSNVII and PSYKPIPIPL, were chemically synthesized with a Cysteine on the N-terminus respectively. 2mg of each peptide was conjugated to KLH, a carrier protein, with the Imject Maleimide Activated mcKLH Kit (Thermo Scientific). Each KLH-conjugated peptide was injected into two rabbits on day 0, boosted 3 times from days 7-21, and anti-sera were collected in 28 days (Antibody China). Anti-sera were tested by ELISA. Anti-Met antibody was purchased from Abcam (ab10728). Anti-ST7 antibody was purchased from ProteinTech Group, Inc (11945-1-AP).

In order to purify the polyclonal antibodies from rabbit anti-sera, peptide PSYKPIPIPL was immobilized on the SulfoLink Resin (Thermo Scientific). Rabbit anti-sera were incubated with the resin and peptide-specific antibodies bound to the peptides and were later eluted. The eluate was tested by ELISA to ensure activity and specificity.

SDS-PAGE and Western Blotting

Cell lysates from A549, DU145, MCF-7, Panc-1 and PBMC were generated using an established protocol (Abmayr, Yao, Parmely, & Workman, 2006). 20 μg of each cell lysate protein was loaded into the lanes of a 12% SDS-PAGE gel and electrophoresed under reducing conditions. The gel was transferred to PVDF

membrane, blocked with 5% nonfat powdered milk and probed with anti-ST7 and anti-PSYKPIPIPL rabbit polyclonal antibodies at concentrations of 0.41 μg/ml (1:500 dilution) and 0.58 μg/ml for 1 hour, respectively. The blot was washed free of primary antibodies followed by the addition of goat anti-rabbit IgG coupled with horseradish peroxidase (HRP) (Jackson Immunolabs, Western Grove, PA) at 1:5000 dilution and incubation for 1 hour. After washing BCIP/NBT substrate (Pierce Chemical, Rockford, IL) was added and the blot was developed for 15 minutes to 1 hour at room temperature.

RESULTS

"Non-Exonic" Peptides Discovered by LC-MS/MS

Acid elutions of each cell line (A549, BXPC-3, DU-145, MCF-7, Panc-1) and PBMC were analyzed using LC-MS/MS and the MS/MS data were analyzed. 241 peptides were identified from searching the non-exonic database with mass spectra. They were denoted "non-exonic" peptides. Table 2 lists 12 such peptides which were further studied.

Table 2

List of "Non-Exonic" Peptides Discovered by LC-MS/MS

Peptide sequence	AA between stops	Conservation	Class	Cell line
HVTFILSNVII	38	0.349833398	Intronic_Distal	PANC-1 (3X), A549 (1X)
KTKIALSLSLPV	58	0.943750598	Intronic_Distal	PANC-1 (3X)
PSYKPIPIPL	82	0.05095702	Intronic_Distal	PANC-1 (2X)
IVSLEGKPL	22	0.00331329	Intronic_Distal	MCF-7 (4X), BXPC-3 (2X)

APSRLTANSASRVH	36	0.013517592	Intronic_Distal	MCF-7 (1X)
CLPGSSNSPASASRVPGTTGARHHA	96	0.003254897	Intronic_Proximal	DU145 (1X)
EFHNKKIL	15	5.80E-05	Intronic_Proximal	PBMC (1X)
ELPSLPPSLLLRLLSSPSSRVP	51	0.323618631	Intronic_Proximal	DU145 (1X)
FFYCSRVPQGLLLL	50	0.02791535	Intronic_Distal	PANC-1 (1X)
FITLTVRVLPF	41	0.049096215	Intronic_Distal	PANC-1 (8X)
LKWFQNVLT	20	0.005172044	Intronic_Proximal	PANC-1 (1X)
LSPSVCGPASKPFKI	77	0.097504436	Intronic_Distal	PANC-1 (1X)

Peptides ranging from 8 a.a. to 25 a.a. were identified from searching the non-exonic database. Some non-exonic peptides were found in intergenic regions (in between known genes); some were found in intronic regions (in between exons). Distal means the distance between the peptide and the nearest exon is >5kb and proximal means the distance is <5kb. A conservation score of 1.0 is 100% conserved in all species analyzed in the UCSC database. The cell lines in which these peptides were found are also listed. X means the number of times a peptide was identified in each acid elution by LC-MS/MS. The chromosomal and "gene" locations of each peptide were known because the peptide/protein databases were constructed from regions of the genome that the ENCODE consortium analyzed. As Figure 2 demonstrates, 5 non-exonic peptides are located in the intronic region of a gene called ST7, a candidate tumor suppressor gene on human chromosome 7q31.2 (Zenklusen, Conti, & Green, 2001). ST7 has two established isoforms, a and b, as shown in Figure 3 (Frazer, et al., 2003). Furthermore, peptide PSYKPIPIPL was found between exons 1 and 2 of MET (Figure 4), KTKIALSLSLPV was found between exons 11 and 12 of CFTR (Figure 5), both of which are present on human chromosome 7q31.2, along with ST7 (Figure 6). This region has been reported for frequent loss of heterozygosity (LOH) of markers in malignant myeloid disorders as well as breast, prostate, ovarian,

colon, head and neck, gastric, pancreatic, and renal cell carcinomas, and it is suggested there is a tumor suppressor gene (TSG) present (Zenklusen, Weintraub, & Green, 1999).

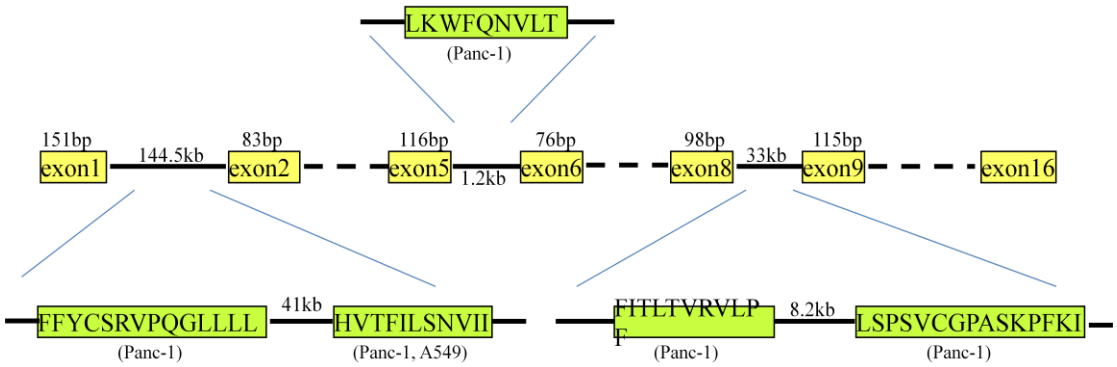


Figure 2. RefSeq Gene ST7, *Homo Sapiens* suppression of tumorigenicity 7. The human ST7 spans 16 exons. Peptides FFYCSRVPQGLLLL and HVTFILSNVII are located between exons 1 and 2 of ST7, both are distant from adjacent exons. Peptide LKWFQNVLT is located between exons 5 and 6, in a relatively small intron. Peptides FITLTVRVLPF and LSPSVCGPASKPFKI are located between exons 8 and 9. "()" indicates from which tumor cell line the peptide was found.

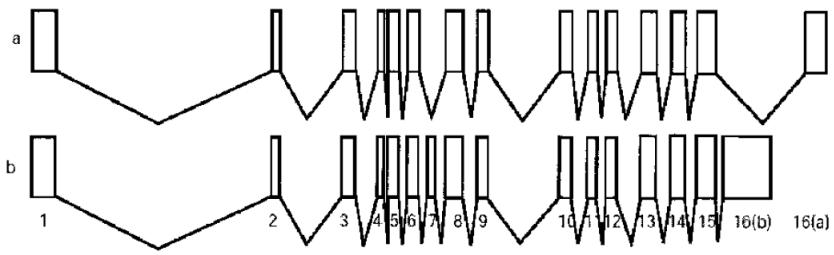


Figure 3. The human ST7 locus on 7q31 has two splice variants (Frazer, K, *et al).* Isoform a is shorter in length than isoform b due to the fact that it is missing the alternatively spliced exon 7 and has a shorter length 3'-end exon. Both isoforms have their own NCBI access numbers, NM_018412 and NM_021908, respectively.

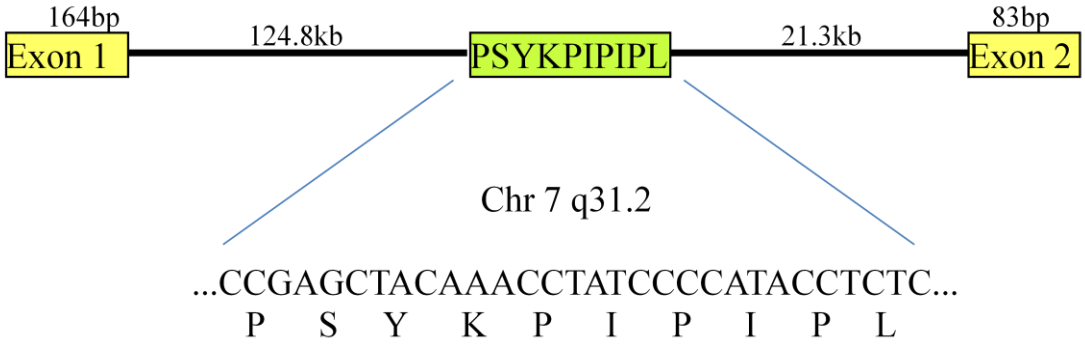


Figure 4. PSYKPIPIPL is located between exons 1 and 2 of MET, *homo sapien* met proto-oncogene (hepatocyte growth factor receptor). The PSYKPIPIPL peptide is proximal to exon 2.

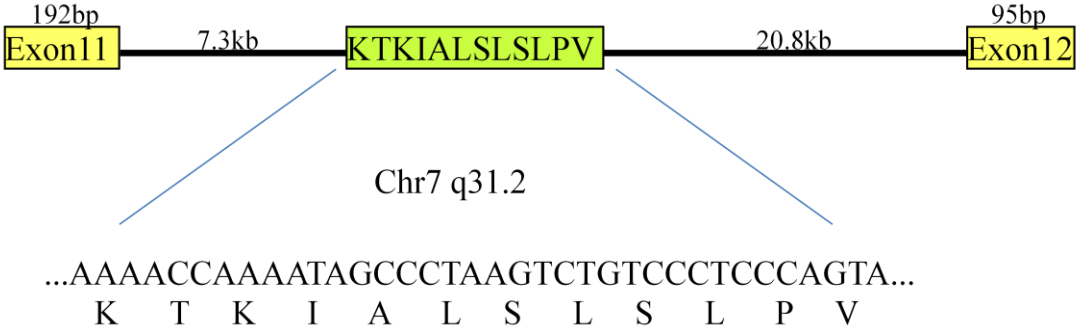


Figure 5. KTKIALSLSLPV is located between exons 11 and 12 of CFTR, homo sapien cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7).

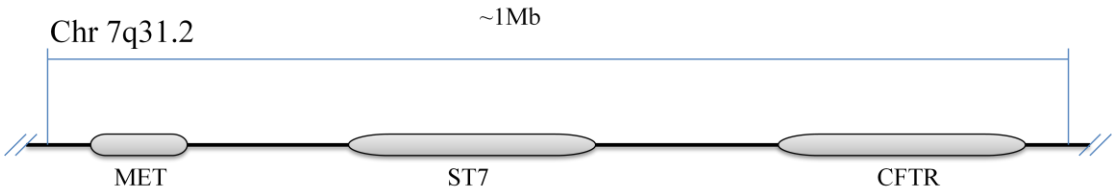
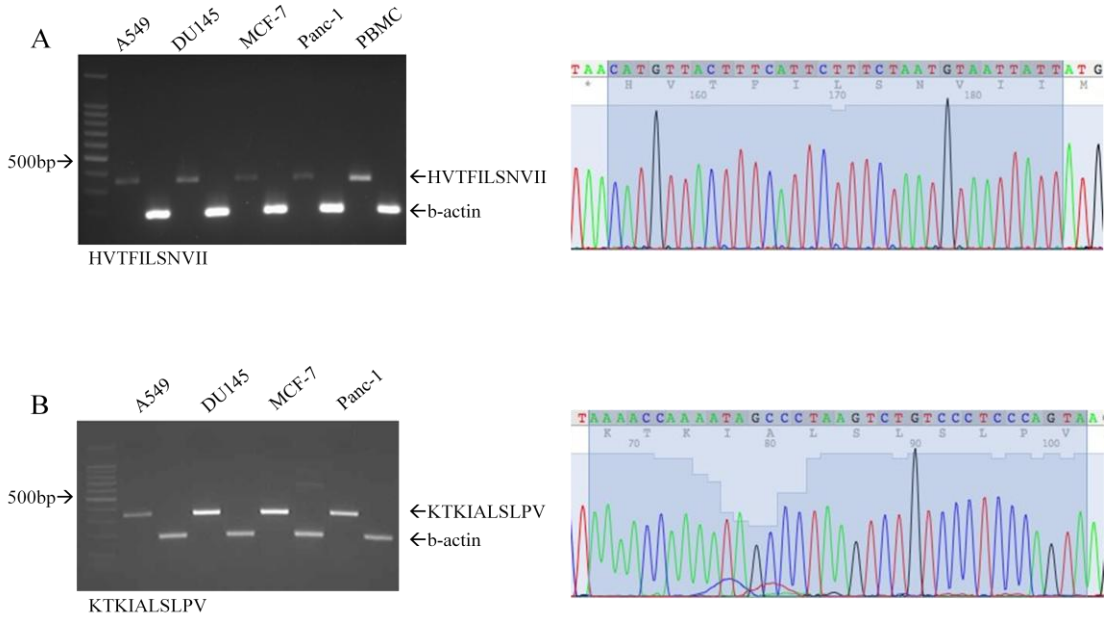


Figure 6. Schematic representation of the positions and relative sizes of genes within the 7q31.2 on chromosome 7. A number of non-exonic peptides were found from intronic regions of genes on chromosome 7q31.

mRNA Validation of Non-Exonic Peptides

PCR was performed with primers flanking the genomic region containing each

peptide to test for transcripts in 4 tumor cell lines A549 (lung), DU145 (prostate), MCF-7 (breast), Panc-1 (pancreas) and PBMC. B-actin control primers were used to exclude the possibility of any genomic DNA contamination. The expected PCR product of b-actin control is 189 bp, with a second band at 650 bp if there is any genomic contamination. Amplicons of expected sizes were observed by agarose gel electrophoresis. Each band was purified and sequenced. Sequencing validated that mRNA encoding these peptides were present in the tumor cell lines and in PBMC. The gel electrophoresis pictures of the PCR and sequencing trace files of peptides HVTFILSNVII, KTKIALSLPV, PSYKPIPIPL, IVSLEGKPL were included in Figure 7.



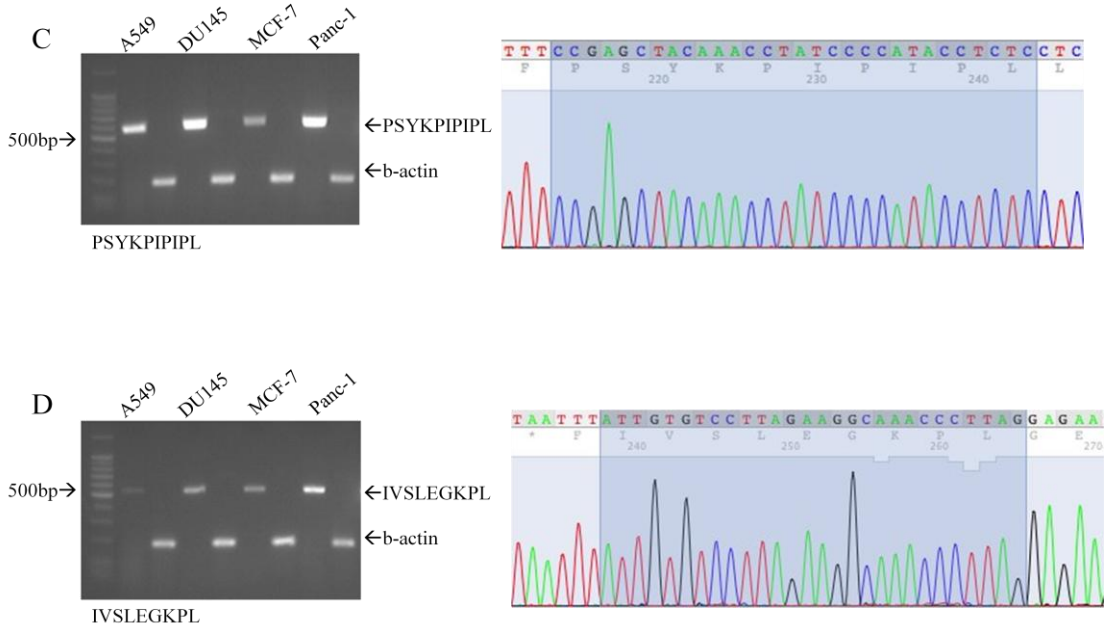


Figure 7. mRNA validation of non-exonic peptides. A. HVTFILSNVII. B. KTKIALSLPV. C. PSYKPIPIPL. D. IVSLEGKPL**.** mRNA was harvested and treated with DNase from 4 tumor cell lines A549 (lung), DU145 (prostate), MCF-7 (breast), Panc-1 (pancreas) and PBMC. cDNA was synthesized using oligo-dT primers to ensure only poly(A) mRNA was replicated. Primers flanking the intronic region containing the peptide were used to amplify the transcript containing each sequence in the five cell types. Primers complementary to the 3' end of exon 3 in b-actin and the 5' end of exon 4 in b-actin were used to amplify the oligo-dT primed cDNA as a positive control and to asses if there was any genomic contamination. The expected PCR product is 189 bp, with a second band at 650 bp if there is any genomic contamination. For every cDNA template from cell lines, 2 PCRs were run. The first lane contains the PCR product of the oligo-dT primed cDNA and the primers flanking the peptide. The second contains the PCR product of the oligo-dT primed cDNA and b-actin primers. The PCR product of each peptide was sequenced, confirming the presence of codons

for each peptide within each intronic region.

PCR with Exonic Primers

We hypothesize that translation of "introns" might indicate that undiscovered splice variants exist for some genes, which means there might be undiscovered exons. If these undiscovered exons are expressed in frame with adjacent known exons, we can test our hypothesis. Thus, exonic primers were designed to test this hypothesis. Figure 8 shows the locations of exonic primers in relation to the intronic primers flanking the peptides that we used to validate the existence of mRNA for the peptides discovered by LC-MS/MS.

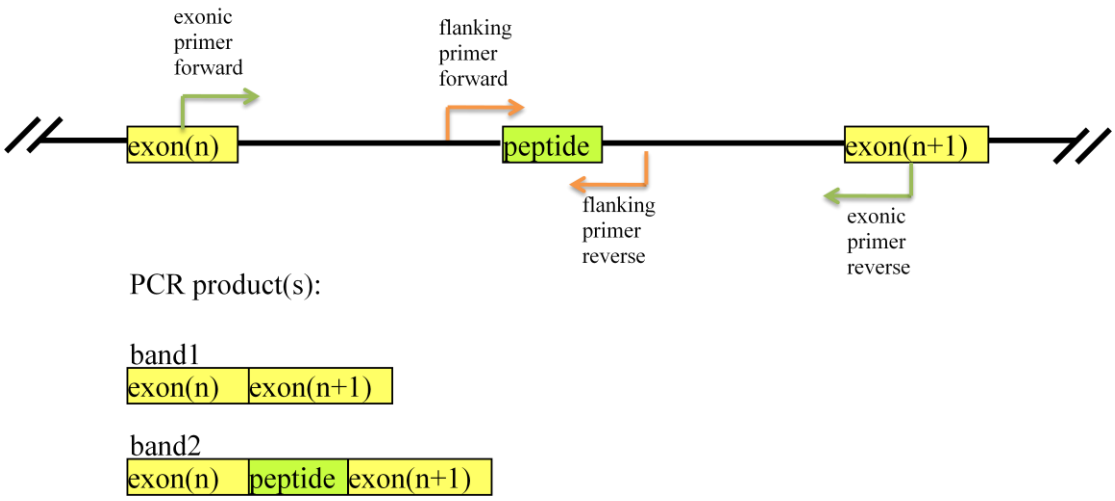


Figure 8. Design of exonic and flanking primers and possible outcomes. As described in the methods section, primers flanking the peptides in the intronic regions were designed to find the peptide in primary transcripts. Furthermore, primers for adjacent exons were designed with the goal of finding additional product that contains any of the non-exonic peptides. Band 1 is the direct product of 2 exons, exon (n) and exon (n+1), spliced together as this is the known splicing product. If any peptide is from an alternative exon, an additional and larger band should be detected. The exception

would be if exon (n+1) was not spliced into the putative exon that the peptide is from. In Figure 9, PCR with exonic primers on HVTFILSNVII have 3 distinct bands. Bands a and b were purified and cloned into pCR2.1 TOPO vector (Invitrogen), and transformed into *E. coli* which were spread on LB plates with 50 ug/ml carbenicillin. Colonies were picked and expanded followed by PCR with M13 primers which would add an extra 201bp to each insert. Sequencing results revealed an extra exon, denoted exon 1a, which had previously been reported (Kimura et al., 2006). However, none of the PCR products (a, b or c) contained the peptide sequences. For the rest of the peptide candidates, a single band was obtained in each PCR product. Figure 10 is an example. The result indicates that the putative exon from which the peptide was detected is not spliced into the known exons, exon (n) or exon (n+1). It is inferred that the peptide is originated from an independent gene. This is the case for all the "non-exonic" peptides studied.

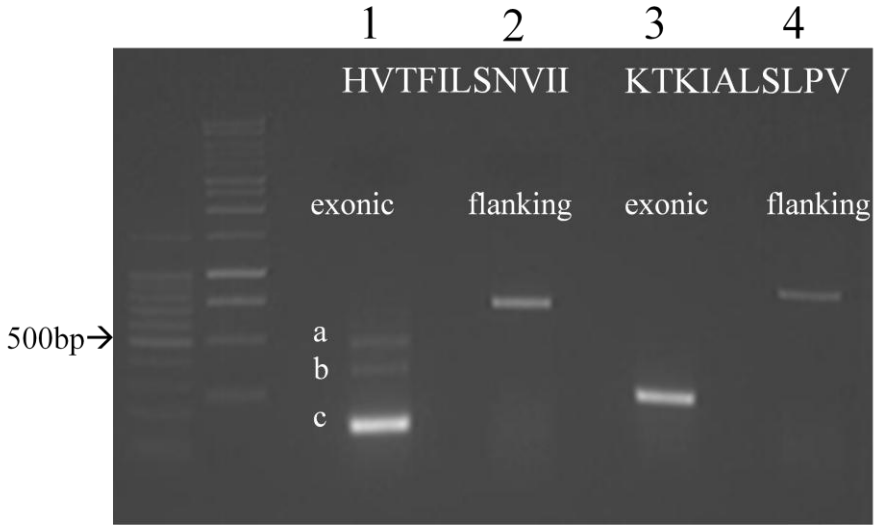


Figure 9. Gel electrophoresis of PCR product with exonic and flanking primers on peptides HVTFILSNVII and KTKIALSLPV. mRNA from Panc-1 cell line was

harvested and treated with DNase. Lanes 1 and 3 are PCR products with exonic primers. Lanes 2 and 4 are PCR products with flanking primers. 3 bands showed up in lane 1. Subsequent cloning confirmed a small exon 1a between exons 1 and 2 in ST7, which had been discovered previously (Kimura, et al., 2006).

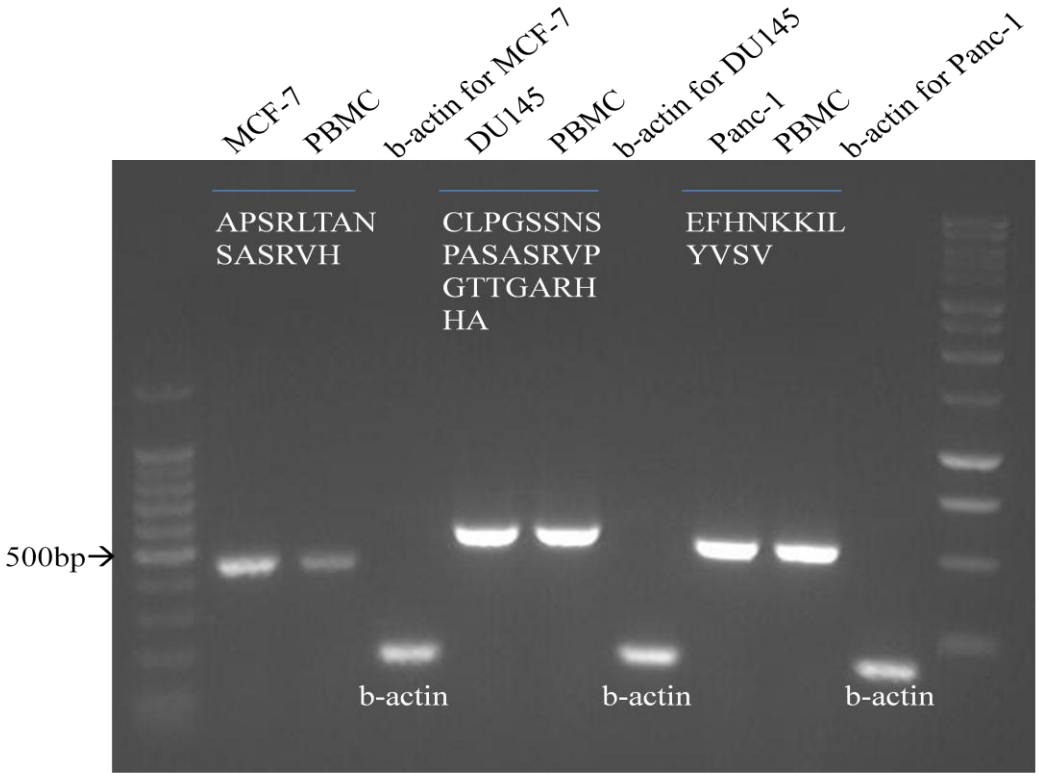


Figure 10. Gel electrophoresis of PCR products with exonic primers of 3 peptides. mRNA was harvested and treated with DNase. For each peptide, PCR (single amino acid code in white at the top of the gel) with exonic primers were tested in PBMC and the tumor cell line for which the peptide was discovered by LC-MS/MS. For each reaction only one band was obtained that corresponded to the product of two known adjacent exons spliced together.

"Baby-step" PCR

In a strategy to find the entire transcript containing each peptide, "baby-step"

primers were designed as in Figure 11. The rationale for this was that trans-splicing might be occurring due to cryptic splice sites in the introns. Primers were designed on the intronic region upstream and downstream of peptides HVTFILSNVII and KTKIALSLPV, as shown in Figure 11. Expected products are around 1 kb for each primer pair, and the products overlap to make a complete connection of the entire transcripts. Oligo-dT primers were used to make cDNA from Panc-1 cell line to ensure only poly(A) RNA were amplified. Surprisingly, all primer pairs had amplification of the right size, which means almost 6 kb of mRNA were found containing the peptides. 2761 bp upstream and 3004 bp downstream of HVTFILSNVII were established (5799 bp total). Similarly, 2765 bp upstream and 2932 bp downstream of KTKIALSLPV were established (5734 bp total).

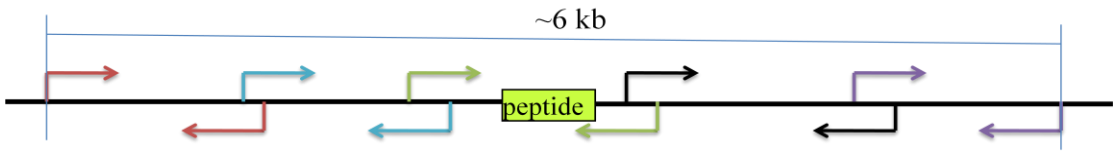


Figure 11. Design of "baby-step" primers covers ~6 kb of genomic sequence. Green primers were the initial flanking primers for detection of peptides. Although the peptide/protein database we built was based on ENCODE RNA, these results may lead to the conclusion that the peptides discovered by LC-MS/MS may have come from transcripts elsewhere in the genome and are not related to the genes from whose intronic regions they were discovered.

Immunoprecipitation with Antibody

Rabbit polyclonal antibodies were generated to two peptides, PSYKPIPIPL and HVTTFILSNVII. Due to the high hydrophobicity of HVTTFILSNVII, coupling

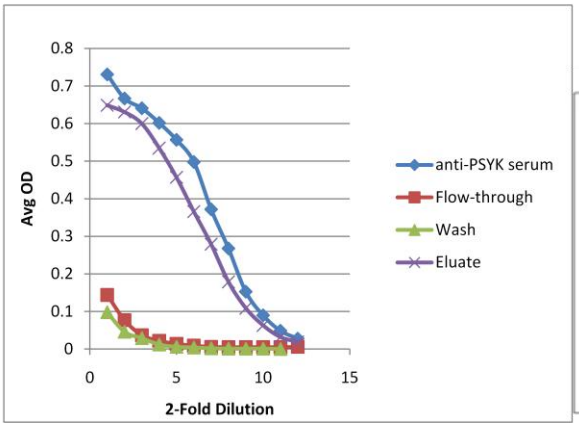
to KLH carrier was inefficient. For the same reason, HVTTFILSNVII was also a poor immunogen, and elicited only low titer antibodies. The rabbit sera against PSYKPIPIPL proved to be high titer, 1:25,600. PSYKPIPIPL peptide was coupled to a solid support via an N-terminal cysteine. Peptide-specific antibodies were purified from the high titer anti-PSYKPIPIPL rabbit antiserum using the peptide affinity column (Figure 12).

Then anti-PSYKPIPIPL antibodies were used to make an affinity column using hydrazide chemistry to link oxidized carbohydrate groups on the antibody to a solid support (Thermo Scientific). Once the "anti-PSYK" antibody column was constructed and tested using PSYKPIPIPL linked to bovine serum albumin (BSA), cell lysates from Panc-1 cell line from which PSYKPIPIPL was originally discovered was run through the column. The eluate was run on a SDS-PAGE gel, but no band was observed. Western blotting with commercial anti-ST7 antibody only detected the native form of ST7 (63 kDa) in the 4 tumor cell lines tested, A549, DU145, MCF-7 and Panc-1 (Figure 13 A). Western blotting with affinity-purified anti-PSYKPIPIPL antibody did not show any specific band on the membrane (Figure 13 B), which may be due to the low abundance of the peptide in the cell lysate samples, or the resolution limit of western blotting.

Coating	ASU001						ASU001-KLH					
Sera dilution / R No.	1:1000	1:5000	1:10000	1:50000	1:100000	NC	1:1000	1:5000	1:10000	1:50000	1:100000	NC
1#	0.103	0.071	0.073	0.068	0.06	0.063	3.481	3.125	3.383	3.125	3.078	0.098
2#	0.19	0.084	0.083	0.081	0.057	0.052	3.134	3.133	3.26	3.133	2.993	0.108

A

Note
1.Coatings:ASU001 ASU001-KLH 1ug/well;
2.1st antibody: rabbit sera (8/17/09)
3.2nd antibody: Goat anti Rabbit IgG (H+L);
4.NC: negative control,5%PBS-Milk as 1 st.



C

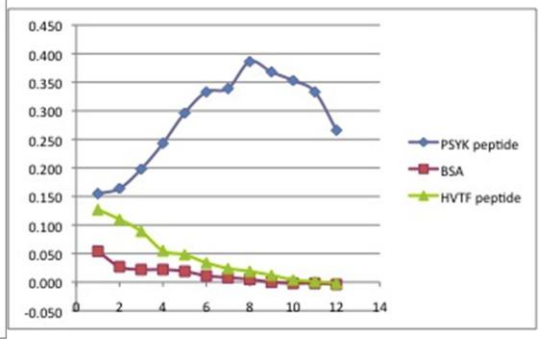


Figure 12. ELISA on rabbit anti-sera. A. Anti-HVTFILSNVII-KLH anti-sera had only KLH-specific antibodies (Ab China). B. Anti-PSYKPIPIPL antibodies were successfully purified using peptide column. ELISA on anti serum and flow-through both started at 1:100 dilution and ELISA on eluted antibodies started at 1µg/ml and 2-fold dilution was performed. C. Anti-PSYKPIPIPL antibodies were tested for cross-specificity for HVTFILSNVII (starting at 1µg/ml and 2-fold dilution), and showed no such activity.

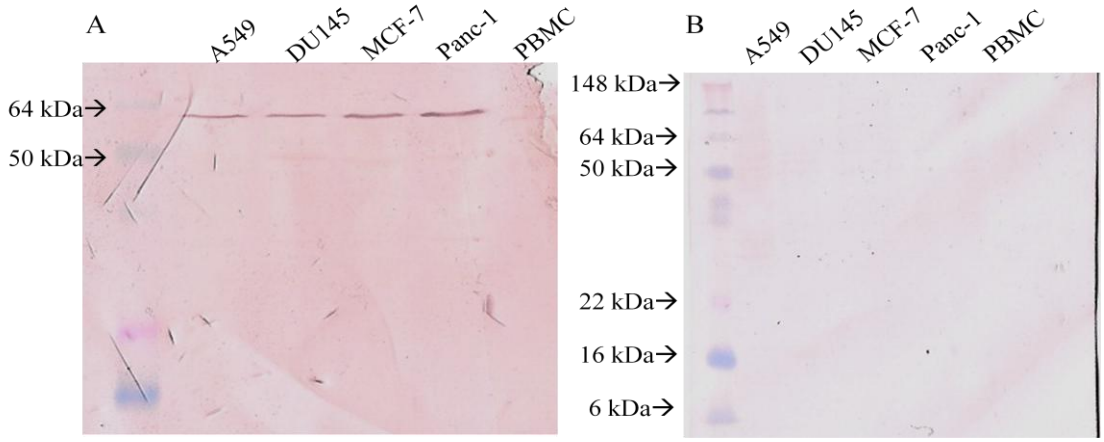


Figure 13. Western blotting with anti-ST7 and anti-PSYKPIPIPL antibodies. A. Western blotting for the detection of possible splicing variants of ST7 with commercial anti-ST7 antibody. Anti-ST7 rabbit polyclonal antibody was incubated with PVDF membrane blotted with cell lysate made from $1x10^7$ cells from 4 human tumor cell lines: A549, DU145, MCF-7, Panc-1 and healthy donor PBMC. 20 μg of cell lysate protein was added to each lane followed by SDS-PAGE as described in Methods section. The protein ladder used was SeeBlue Plus 2 Pre-Stained Standard (Invitrogen, CA). The known size of ST7 is 63 kDa which was detected. B. Western blotting with affinity-purified anti-PSYKPIPIPL antibody. No specific band appeared on the membrane suggesting that the amount of the product including peptide PSYKPIPIPL may be below the detection limit of western blotting, or the size of the product is too small to be resolved on a western blotting.

DISCUSSION

The ENCODE pilot project has provided with us a rich source of functional information on the human genome. One remarkable discovery is the generation of numerous intercalated transcripts spanning the majority of the genome. However, the

biological functions of these transcripts are not known. The ENCODE consortium proposed that some of them were regulatory elements. Our studies suggest that some of the ENCODE RNA is translated into protein and peptides. As is known today, current gene annotation and gene prediction is not 100% accurate. We suggest the peptides we found in tumor cell lines serve some function, whether as peptides, or are fragments of undescribed splice variants. Although the functions of the peptides are unknown at the moment, a recent publication in *Science* indicates that translation of peptides impart a phenotype in Drosophila. Dr. Kageyama and his collaborators demonstrated that during Drosophila embryogenesis, peptides (11-32 amino acids) provided a strict temporal control of the transcriptional program of epidermal morphogenesis by modifying a transcription factor (Kondo et al., 2010). Other studies indicate that a fraction of ncRNAs with short open reading frames (ORFs) potentially encodes peptides and they may play key roles in cell signaling and other processes (Amaral, et al., 2008; Frith et al., 2006).

Due to the 3 kDa filtration of acid elutions methods, the peptides we discovered are only 8 to 25 in length. We were able to PCR-amplify polyadenylated mRNA that contained the peptide sequences we detected by LC-MS/MS and confirm the presence of the peptide by DNA sequence analysis. Initially, we hypothesized the intronic peptides came from different splice variants. To prove this hypothesis, PCR was performed using primers in exons that flanked the intronic peptide. The results suggestsed the peptide had no link to known exons, because we never observed a product containing flanking exons plus an exon containing the target peptide. This

negative result, along with the large size of the intron, lead us to hypothesize that the peptides may come from independent genes, which means there could be a gene within a gene. Supporting this hypothesis is the data indicating that Western blotting using commercial antibodies against ST7 and MET did not detect any other spliceoform of either ST7 or MET. However, the inconclusive result from 5' RACE and 3' RACE did not solve the puzzle, and Northern blots were not successful. Moreover, “baby-step” PCR revealed there were nearly 6 kb continuous mRNA present between both ST7 and MET, which is already larger than any known exon. One possibility is that pre-spliceosome was amplified which explains why I had a 6kb-mRNA. Despite the difficulties encountered in the latter stages of this project, we showed that peptides are produced from large introns in 5 different human tumor cell lines and human PBMC. We were able to detect the peptides directly from cell lysate by mass spectrometry. We validated polyadenylated RNAs coding for these peptides. Our results from both “baby-step” PCRs and western blots suggest that these peptides may come from independent genes that locate in the intronic regions of other genes. However, there are many questions that remain to be answered.

Another question to answer is how the mRNAs got translated. All peptide sequences and contexts were examined. There were no close start codons upstream in frame with the peptides. No traditional splicing donor and acceptor sites were found. More interestingly and mysteriously, some of the peptides had stop codons in frame not far upstream, especially the HVTFILSNVII peptide, which has a stop codon right ahead of it. The stop codon got skipped for mechanisms unknown to us yet. We hope

with the inventions of new technologies we can examine these non-exonic peptides further.

Though many questions await to be answered, our findings have significant implications. Conservation of many of those non-exonic peptides (Table 2) suggests that evolution selected to preserve the ability to transcribe and translate the "non-coding" regions of DNA. The most difficult future of this project will be to elucidate the biological functions of those small peptides. We anticipate our preliminary results will lead to the study and findings of novel transcriptional and translational mechanisms. The elevated transcription and translation activity of the genome may give us biomarkers for early diagnosis and potential drug targets.

REFERENCES

Abmayr, S. M., Yao, T., Parmely, T., & Workman, J. L. (2006). Preparation of nuclear and cytoplasmic extracts from mammalian cells. *Curr Protoc Mol Biol, Chapter 12,* Unit 12 11. doi: 10.1002/0471142727.mb1201s75

Amaral, P. P., Dinger, M. E., Mercer, T. R., & Mattick, J. S. (2008). The eukaryotic genome as an RNA machine. *Science, 319*(5871), 1787-1789. doi: 319/5871/1787 [pii]

10.1126/science.1155472

Antwi, K., Hostetter, G., Demeure, M. J., Katchman, B. A., Decker, G. A., Ruiz, Y., . . . Lake, D. F. (2009). Analysis of the Plasma Peptidome from Pancreas Cancer Patients Connects a Peptide in Plasma to Overexpression of the Parent Protein in Tumors. *Journal of Proteome Research, 8*(10), 4722-4731. doi: Doi 10.1021/Pr900414f

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., . . . Consortium, E. P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447*(7146), 799-816. doi: Doi 10.1038/Nature05874

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol, 268*(1), 78-94. doi: S0022-2836(97)90951-7 [pii]

10.1006/jmbi.1997.0951

Chain, P. S., Grafham, D. V., Fulton, R. S., Fitzgerald, M. G., Hostetler, J., Muzny, D., . . . Detter, J. C. (2009). Genomics. Genome project standards in a new era of sequencing. *Science, 326*(5950), 236-237. doi: 326/5950/236 [pii]

10.1126/science.1180614

Collins, F. S., Lander, E. S., Rogers, J., Waterston, R. H., & Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945. doi: Doi 10.1038/Nature03001

Edwards, D., & Batley, J. (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J, 8*(1), 2-9. doi: PBI459 [pii]

10.1111/j.1467-7652.2009.00459.x

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., & Hardison, R. C. (2003). Cross-species sequence comparisons: a review of methods and available resources. *Genome Res, 13*(1), 1-12. doi: 10.1101/gr.222003

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., . . . Grimmond, S. M. (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genet, 2*(4), e52. doi: 10.1371/journal.pgen.0020052

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html.

Jiang, Z., Rokhsar, D. S., & Harland, R. M. (2009). Old can be new again: HAPPY whole genome sequencing, mapping and assembly. *Int J Biol Sci, 5*(4), 298-303.

Katalin, F., & Kapranov, P. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature, 457*(7232), 1028-1032. doi: nature07759 [pii]

10.1038/nature07759

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., . . . Sugano, S. (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res, 16*(1), 55-65. doi: gr.4039406 [pii]

10.1101/gr.4039406

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., . . . Kageyama, Y. (2010). Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis. *Science, 329*(5989), 336-339. doi: DOI 10.1126/science.1188158

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Conso, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: a review. *Hum Mol Genet, 18*(R2), R163-168. doi: ddp396 [pii]

10.1093/hmg/ddp396

Ratsch, G., Sonnenburg, S., Srinivasan, J., Witte, H., Muller, K. R., Sommer, R. J., & Scholkopf, B. (2007). Improving the Caenorhabditis elegans genome annotation using machine learning. *PLoS Comput Biol, 3*(2), e20. doi: 06-PLCB-RA-0040R2 [pii]

10.1371/journal.pcbi.0030020

Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., . . . Vidal, M. (2003). C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet, 34*(1), 35-41. doi: 10.1038/ng1140

ng1140 [pii]

Zenklusen, J. C., Conti, C. J., & Green, E. D. (2001). Mutational and functional analyses reveal that ST7 is a highly conserved tumor-suppressor gene on human chromosome 7q31. *Nat Genet, 27*(4), 392-398. doi: 10.1038/86891

86891 [pii]

Zenklusen, J. C., Weintraub, L. A., & Green, E. D. (1999). Construction of a high-resolution physical map of the approximate 1-Mb region of human chromosome 7q31.1-q31.2 harboring a putative tumor suppressor gene. *Neoplasia, 1*(1), 16-22.