論文内容の要旨

Characterization of intergenic transcription start sites by multifaceted use of massively paralleled sequencer

（遺伝子間領域における転写開始点に関する多角的トランスクリプトーム解析）

氏　名：　サティラポンサスティ　ヌアンカンヤー

I. Introduction

Since human genome and several other organisms' genomes are completely sequenced, gene regulation is realized to be more immense and more complex than expected. The functional properties are known not only on protein coding sequences but also non protein coding or untranslated regions. Recently, ten of thousands of non protein coding RNAs (ncRNAs) have been identified. Also in our large-scale studies of human transcriptome, large numbers of putative ncRNAs have been observed, especially from intergenic regions. However, solid evidences for the in vivo existence of them are inadequate, and knowledge on the transcriptional mechanism of the ncRNAs remains elusive.

In this study, the intergenic transcription start site (TSS) clusters, identified by TSS Seq analysis of twelve human cell lines and tissues, were characterized. By multifaceted use of Illumina GA sequencer (Fig.1), made it possible to classify the TSSs depending on possible biological relevance.

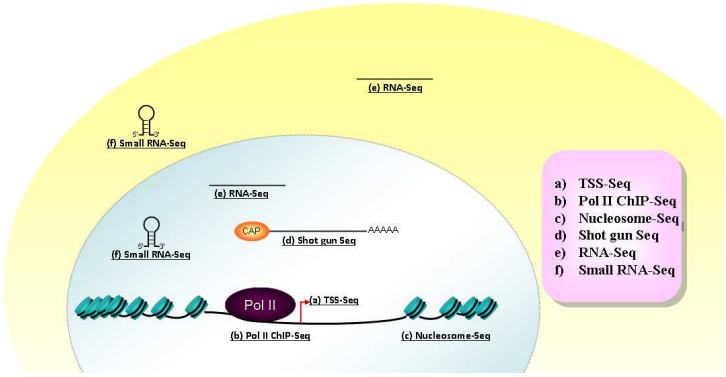


Fig.1. Schematic representation of the multifaceted use of massively paralleled sequencer

II. Materials and Methods

In order to perform genome-wide integrated analysis of the intergenic TSSs, various techniques were combined with the massively paralleled sequencing technology, Illumina GA (Fig.1), namely;

a) TSS Seq; 139,446,730 36 bp-TSS tags, uniquely mapped to the human genome (hg18) without any mismatch from six cell lines (a colon cancer cell line, DLD-1; an embryonic kidney cell line, HEK293; a breast cancer cell line, MCF-7; a normal fibroblast cell, TIG-3; a Burkitt's lymphoma cell line, Ramos; a lung epithelial cell line, BEAS2B) and six normal tissues (brain, heart, kidney, fetal brain, fetal heart, fetal kidney) of humans. Additional, 36,761,810 36 bp-TSS tags from DLD-1, MCF7, HEK293, and TIG3 cells

which were cultured in 1% O_2.

b) ChIP Seq for RNA polymerase II; 10,687,815 tags for IP and 6,506,186 tags for background control from DLD-1 cells

c) Nucleosome Seq; 19,570,149 single-end tags from DLD-1 cells

d) Shotgun Seq; 3,382,901 shotgun tags

e) RNA Seq; 20,094,475 tags and 14,879,174 tags for nuclear and cytoplasmic fraction from DLD-1 cells

f) Small RNA Seq; 604,141 21-25 bp size-selected RNA tags from DLD-1 cells

RT-qPCR; expression profiling of the iTSCS and neighboring RefSeq genes were performed in 20 kinds of human normal tissues (Clontech Human Total RNA Master Panel II; Adrenal gland, Bone marrow, Brain(whole), Fetal Brain, Fetal liver, Heart, Kidney, Liver, Lung (whole), Placenta, Prostate, Salivary gland, Skeletal muscle, Testis, Thymus, Thyroid gland, Trachea, Uterus, Colon w/ mucosa, and spinal cord)

III. Results

Integrative transcriptome analysis of the intergenic TSCs (iTSCs) revealed that not all of the iTSCs have the same characters. Figure 2 represents the scheme of the classification and the used methods. The findings in this study (will be described in more details below) demonstrated that possible biological relevance iTSCs had distinctive features, thus could be discriminated from the other noise-level iTSCs.

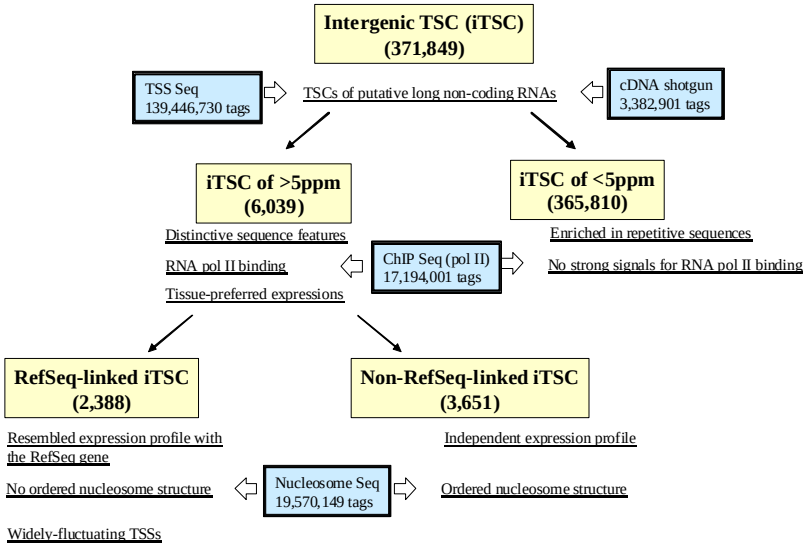


Fig.2. Schematic representation of the integrative transcriptome analysis of the iTSCs

i) **iTSCs of >5ppm are only a minor population of the entire iTSCs**

Base on 140 millions of TSS Seq tags, 371,849 iTSCs (TSS tag clustered into 500 bp bins) were characterized. However, the number of iTSCs with significant number of TSS tags, >5 parts per million (ppm), at least in one cell type was only 6,039.

ii) **iTSCs have strong likelihood of representing TSSs of long ncRNAs**

Overlapping analysis of the iTSCs position with the 5'ends of the cDNA collections; MGC and FLJ, and complete cDNA sequencing of shotgun sequence tags, revealed that 395, 1,617, and 464 iTSCs overlapped with the 5'-ends of the MGC, FLJ, and shotgun cDNAs, respectively.

iii) **Surrounding sequences of iTSCs of >5ppm have several distinctive features**

By analyzing characters of the surrounding sequences of the iTSCs of >5ppm, many characteristic properties were found differently from iTSCs of <5ppm as shown in Table 1.

	>5ppm	<5ppm	statistics	p_value
total	6039	365810		
G+C%	46%	41%	Wilcoxon test	P<1e-100
CpG island	13%	1.7%	Wilcoxon test	P<1e-100
Conservation score	0.21	0.09	Proportion test	P<1e-100
Repetitive element	1256 (21%)	184012 (50%)	Proportion test	P<1e-100

Table 1. Characteristic features observed in the iTSCs of >5ppm and <5ppm

The ChIP Seq analysis also showed diversity of Pol II bindings in the nearby regions of iTSCs of >5ppm and <5ppm. The results showed that 376 (40%) out of 931 iTSCs which were expressed at >5ppm in DLD-1 overlapped the pol II binding sites, while only 3,305 (7%) out of 46,416 iTSCs of <5ppm overlapped them. Interestingly, among these 3,305 cases, although iTSCs were of <5ppm in DLD-1 cells, those in other cell types were of >5ppm in 305 cases.

iv) **Majority of iTSCs of >5ppm expressed as tissue-preferred manner**

By using digital tag count of the TSS-Seq, 3,739 (62 %) of iTSCs were found expressed at >5ppm only in one cell type. RT-qPCR assays were performed in additional 20 kinds of human normal tissues revealed that clear signals of 44 out of 67 iTSCs of >5ppm were observed at least in one tissue.

v) **A quarter of iTSCs of >5ppm are located in the proximal regions of the RefSeq genes, which their transcription is likely to be influenced from open chromatin of the neighboring transcription**

23% (1,382 cases) of the iTSCs of >5ppm were found located within 10 kb downstream on the same strand of the RefSeq genes (RefSeq-linked iTSCs), while this frequency was found only 7% (24,486 cases) in iTSCs of <5ppm. The expression patterns of iTSCs of and their neighboring RefSeq genes were compared by calculating Pearson's correlation of the respective TSS tag counts. The expressions of the RefSeq-linked iTSCs were found generally resembled to the expressions of their nearest RefSeq genes (Fig3A). To further observe the expression correlation in wider cell types, the real time RT-PCR assays were done in 20 normal tissues. Again, general correlation of the expression patterns between the RefSeq-linked iTSCs of >5ppm and the nearby RefSeq genes was observed (Fig3B), on the other hand no such a correlation was found for the non-RefSeq-linked iTSCs of >5ppm.

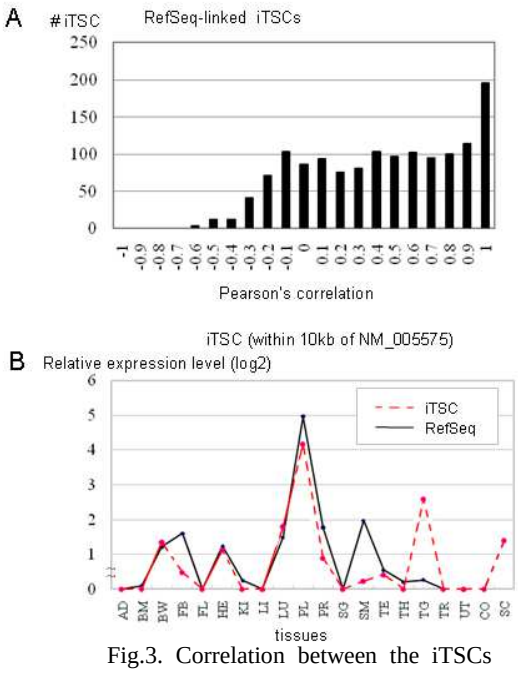


Fig.3. Correlation between the iTSCs and the nearest RefSeq genes

By nucleosome Seq analysis, formation of the nucleosome-free regions were found just upstream of the TSSs the non-RefSeq-linked TSCs of >5ppm and periodic positioning of nucleosome immediately downstream of the TSS. On the contrary, no such an orderly nucleosome structure was observed in surrounding areas of the RefSeq-linked iTSCs >5ppm either the iTSCs of <5ppm.

vi) **Biological relevance of the majority of the non-RefSeq-linked iTSCs is not serving as miRNA precursor**

The overlapping analysis of iTSCs with previously reports small RNAs; miRNAs and snoRNAs, and small RNA Seq data was performed. However, for only 22 miRNAs, 11 snoRNAs, and 2 small RNA Seq, complete cDNA sequences directly showed that the mature forms of the small RNAs were actually located within the transcripted regions following the non-RefSeq-linked iTSCs.

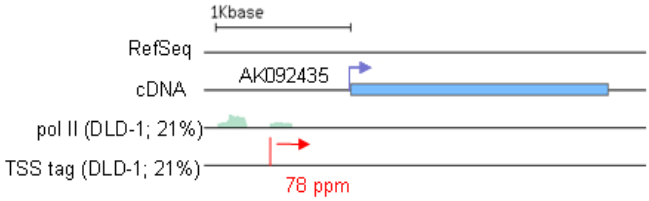
vii) **Large numbers of iTSC transcripts are exported into cytoplasm**

Sub-cellular localization examined by RNA-Seq of nuclear and cytoplasmic fractions in DLD-1 cells, showed that 59% of non-RefSeq-linked iTSCs overlapped at least 3 RNA Seq tags from the cytoplasmic fraction. Of these, 99 iTSCs (17%) were particularly enriched in cytoplasm ($p<0.01$).

viii) **Numbers of non-RefSeq-linked iTSCs showed differential expression in hypoxic response**

By comparing digital tag count of the TSS-Seq of normoxic state and hypoxic state of DLD-1, MCF7, HEK293, and TIG3 cells found that in total, 508 and 365 iTSCs of >5ppm showed expression induction or repression by more than 5 fold in response to the hypoxic shock in at least one cell type.

For further study on the biological roles of iTSCs involve in hypoxic response, the differential expression pattern of iTSCs and RefSeq genes were compared by calculating Pearson's correlation of the respective TSS tag counts. Then particular Gene Ontology (GO) terms of RefSeq genes which showed associated expression with hypoxic induced/repressed iTSCs were examined. Interestingly, while various GO terms were associated, 53 iTSCs were found correlated to the term of "hemophilic cell adhesion". Therefore, one of the biological roles of iTSCs in response to hypoxia might relate with cell-cell interaction control.



Distance to the nearest RefSeq	>10kb
cDNA (longest ORF)	AK092435 (73 a.a.)
Associated GO term	Cell adhesion
Sub-cellular localization (RNA Seq)	Cytoplasm (p= 0.0)
Maximum fold change by hypoxia (1%/21% O_2)	9.6 (in MCF-7)

Fig.4.Example of hypoxic induction iTSCs.

IV. Discussion

In this study, I described the attempt of characterizing the iTSCs by multifaceted use of massively paralleled sequencer. Integrated analysis of various types of transcriptome data revealed the diversity of iTSC properties, particularly for non-RefSeq-linked iTSCs of >5ppm. Non-RefSeq-linked were found to be determinedly and independently regulated, which should be essential to realize biologically relevant transcriptions. This study also demonstrated that integrative transcriptome data was a useful starting point for further study on biological roles of the intergenic transcripts.