Massive transcriptional start site analysis of human genes in hypoxia cells

Katsuya Tsuchihara1, Yutaka Suzuki2,*, Hiroyuki Wakaguri2, Takuma Irie2, Kousuke Tanimoto2, Shin-ichi Hashimoto3, Kouji Matsushima3, Junko Mizushima-Sugano2,4, Riu Yamashita5, Kenta Nakai5, David Bentley6, Hiroyasu Esumi1 and Sumio Sugano2

1Cancer Physiology Project, Research Center for Innovative Oncology, National Cancer Center Hospital East: 6-5-1 Kashiwanoha, Kashiwa, Chiba 277-8577, 2Graduate School of Frontier Sciences, the University of Tokyo: 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, 3Graduate School of Medicine, the University of Tokyo: 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, 4Laboratory of Molecular Virology, Kitasato Institute for Life Sciences, Kitasato University: 5-9-1 Shirokane, Minato, Tokyo 108-8641, 5Institute of Medical Science, the University of Tokyo: 4-6-1 Shirokane-dai, Minato, Tokyo 108-8639, Japan and 6Illumina, Inc: 25861 Industrial Boulevard Hayward, CA 94545, USA

Received October 3, 2008; Revised January 20, 2009; Accepted January 22, 2009

ABSTRACT

Combining our full-length cDNA method and the massively parallel sequencing technology, we developed a simple method to collect precise positional information of transcriptional start sites (TSSs) together with digital information of the gene-expression levels in a high throughput manner. We applied this method to observe gene-expression changes in a colon cancer cell line cultured in normoxic and hypoxic conditions. We generated more than 100 million 36-base TSS-tag sequences and revealed comprehensive features of hypoxia responsive alterations in the transcriptional landscape of the human genome. The features include presence of inducible ‘hot regions’ in 54 genomic regions, 220 novel hypoxia inducible promoters that may drive non-protein-coding transcripts, 191 hypoxia responsive alternative promoters and detailed views of 120 novel as well as known hypoxia responsive genes. We further analyzed hypoxic response of different cells using additional 60 million TSS-tags and found that the degree of the gene-expression changes were different among cell lines, possibly reflecting cellular robustness against hypoxia. The novel dynamic figure of the human gene transcriptome will deepen our understanding of the transcriptional program of the human genome as well as bringing new insights into the biology of cancer cells in hypoxia.

INTRODUCTION

Aberrantly growing cancer cells in solid tumors frequently encounter a shortage of blood flow, which leads to insufficient oxygen supply. Tumor cells adapt themselves to such hypoxic microenvironment by shifting their ATP production metabolism from oxidative phosphorylation to anaerobic glycolysis, and by enhancing glucose intake. Tumor cells also induce angiogenesis to acquire additional blood supplies. Such adaptations are supposed to be essential in survival as well as malignant transformation of tumor cells in vivo (1,2). During this series of events, transcriptional regulation plays a pivotal role. It has been well documented that hypoxia inhibits proteasomal degradation of α subunits of hypoxia inducible factors (HIF1α and HIF2α). Stabilized subunits translocate from the cytoplasm into the nucleus and form a heterodimer complex with HIF1β. HIF complexes transactivate various downstream genes, such as the genes encoding glycolytic enzymes, glucose transporters, the enzymes eradicating organic acids and VEGF which induces angiogenesis. However, the specific function of each isoform of the subunits remains unclear. Meanwhile, ‘HIF-independent’ regulation of hypoxia-inducible genes has also been documented (3,4). Thus, the current view of hypoxic versatility in transcriptome programs in cancer cells is still far from comprehensive. A bird’s eye view on what range of genes are induced in what manner still remains mostly elusive. Although some genome-wide expression profiles using microarrays have been reported, they represent mere collective information of the fold inductions of the individual genes (5–9). In this regards, we believed that information

*To whom correspondence should be addressed. Tel/Fax: +81 4 7136 3607; Email: ysuzuki@k.u-tokyo.ac.jp

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
about exact positions of transcriptional start sites (TSSs) and absolute levels of the transcriptions starting from them would lead to more comprehensive understandings.

Several methods based on cDNA analysis have been developed for large-scale identification of TSSs (10–13). We have also developed a method to selectively replace the cap structure of the mRNA with a synthetic oligo, which we named the oligo-capping method (11). By sequencing 1.8 million cDNAs isolated from oligo-cap cDNA libraries from various kinds of human cells and tissues (14), we have collected the positional information of the TSS and analyzed putative proximal promoter regions (15,16). We, as well as another research group in RIKEN, have further improved the efficacy of this approach by combining the cap-selection method with the SAGE method (17,18). In these methods, 5’-ends of full-length cDNAs were concatenated, so that 10–15 base-long 5’-end tag sequences could be identified by single-pass sequencing. By intensive analysis of CAGE-tag libraries in humans and mice, the FANTOM consortium reported a first glimpse of the transcription landscape of mammalian genomes (19,20). However, such an overview of the TSSs has been obtained from collective analysis of various cell types and tissues, for each of which the data coverage still remains scarce. Therefore, it does not represent the actual transcriptional landscape in any given cell type. Besides, it has been suggested recently that mammalian genes seem to utilize multiple alternative promoters very frequently, which enable a single locus to encode functionally distinct proteins, thereby serving as a molecular basis for realizing multifaceted use of a limited number of human genes (21,22). Nonetheless, the depth of the analysis has not reached the level of these alternative promoters, whose expression levels are often low and limited to particular cell types or cellular environments.

Recently developed massively parallel sequencing technologies have provided a potential means to further improve the throughput of TSS identification. For example, Illumina GA sequencer (23) can sequence 10–30 million sequences per run. Although the read length which this sequencer can generate is short (currently up to 36 bases), it is sufficient to uniquely determine the precise positions of TSS. By combining oligo-capping method with the Illumina GA technology, we developed a simple method to collect information of the TSS together with the digital data of the expression levels of the transcripts. Here we show this approach enabled us to see the genome wide transcriptional landscape in response to hypoxia in a human colorectal cancer cell line.

**MATERIALS AND METHODS**

**Cell culture and RNA interference**

Human cell line, DLD-1 cells, was purchased from American Type Culture Collection (ATCC number: CCL-221). Cells were maintained in Dulbecco’s modified Eagle’s medium (DMEM) (Invitrogen) supplemented with 10% fetal calf serum, 4.5 g/l glucose, and antibiotics. RNA interference was accomplished by transfecting DLD-1 cells with the specific siRNA. HIF1A and EPAS1 (HIF2A)-targeting siRNA pool and non-silencing siRNA pool were purchased from Dharmacon. Short oligo-RNAs were transfected using Dharmafect 1 transfection reagent (Dharmacon) as recommended by the manufacturer. For constructing other TSS-libraries, HEK293, MCF7 and TIG3 cells (ATCC number: CRL-1573, ATCC number: HTB-22 and Japan Cell Resource Bank number: JCRB0506, respectively) were cultured in standard conditions and were subjected to the hypoxic shocks in a similar manner.

**Oligo-capping and massively parallel sequencing by Illumina GA Sequencer**

Six million DLD-1 cells were seeded 24 h before transfection. The cells transfected with HIF-targeting and control siRNA were cultured in 21% O2 and 5% CO2 at 37°C for 24 h followed by incubation in 21% O2 or 1% O2 and 5% CO2 for 24 h. Cells were harvested and RNA was extracted using RNeasy (Qiagen). Two hundred microgram of the obtained total RNA was subjected to oligo-capping with some modifications from the original protocol; namely after the successive treatment of the RNA with 2.5 U BAP (TaKaRa) at 37°C for 1 h and 40 U TAP (Ambion) at 37°C for 1 h, the BAP-TAP-treated RNAs were ligated with 1.2 μg of RNA oligo (5’-AAUGAUACGCGGACCACCCGAUCAUCACUCUUUCUCUACCAACGACCUUUCCCUCUACCAACGACCUUCCGACUUCAUGGG-3’) using 250 U T4 RNA ligase (TaKaRa) at 20°C for 3 h. After the DNase I treatment (TaKaRa), poly-A-containing RNA was selected using oligo-dT powder (Collaborative). First strand cdNA was synthesized from 10 pmol of random hexamer primer (5’-CAAGCAGAAGACGCCGACCGAGAACCACTCGANNNNNNC-3’) using Super Script II (Invitrogen) by incubating at 12°C for 1 h and at 42°C overnight. Template RNA was degraded by alkaline treatment. For PCR, one-fifth of the first strand cDNAs were used as the PCR template. Gene Amp PCR kits (PerkinElmer) were used with the PCR primers 5’-ATTGATACGGCCGACCAAGAAGAAGATCAGNNNNN-3’ and 5’-CAACGCCGACCGCATCGCAATCGANNNNNN-3’ under the following reaction conditions: 15 cycles of 94°C for 1 min, 56°C for 1 min and 72°C for 2 min. The PCR fragments were size fractionated by 12% polyacrylamide gel electrophoresis and the fraction of 150–250 bp was recovered. The quality and quantity of the obtained single-stranded first strand cDNAs were assessed, again, using BioAnalyzer (Agilent). One nanogram of the size fractionated cDNA was used for the sequencing reactions with the Illumina GA. 15 000–20 000 clusters were generated per ‘tile’ and 36 cycles of the sequencing reactions were performed according to the manufacturer’s instructions.

**Data processing**

The obtained sequences were mapped onto human genomic sequences (hg18 as of UCSC Genome Browser; http://genome.ucsc.edu/) using the sequence alignment program Eland. Unmapped or redundantly mapped sequences were removed from the dataset. For uniquely mapped sequences, relative positions to RefSeq genes were calculated based on the respective genomic coordinates.
Genomic coordinates of exons and other information of the RefSeq transcripts are as described in hg18 as of UCSC Genome Browser. GO (as of June 14th, 2007) and KEGG (Release 42) terms were associated with RefSeq genes by using the loc2go (as of June 14th, 2007) using NCBI Entrez Gene database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene). For each RefSeq gene, a RefSeq region was defined as the region from 50 kb upstream of the most upstream 5'-end exon to the most downstream 3'-end exon. TSS-tags were further clustered into 500-bp bins to generate TSS clusters (TSCs). Details and rationalization of the procedure is described in the ref. (15). For the expression analysis at the gene levels, TSS-tag counts of TSCs belonging to the corresponding RefSeq regions were tallotted. For the expression analyses at the alternative promoter level, intergenic and antisense transcripts, the TSS-tags belonging to the corresponding TSCs were counted. In either case, TSS-tag counts were divided by the total number of uniquely and perfectly (with no mismatch) mapped TSS-tag to calculate TSS-tag ppm (parts per million). For the analysis of intergenic TSCs, overlap between the TSCs and miRNA and snoRNA, from miRBase (http://microrna.sanger.ac.uk/sequences/index.shtml) and snoRNABase (http://www-snorRNA.biotoul.fr/index.php), respectively, were examined.

Validation analysis
Real-time RT–PCRs were performed using 7900HT (ABI) following the standard protocol. PCR primer sequences are shown in the Supplementary Table 10. For the RT–PCR, 1 ng of the first strand cDNAs, which were synthesized by random hexamer primer, was used. In the case of plasmids, 1 pg of the DNA, quantified by O.D., was used, instead. The primer sets were first tested by amplifying the plasmid DNA and the primer sets giving less than 35 Ct cycles were used. The absolute copy number of each transcript in the cDNA population was calculated based on the Ct value of the corresponding plasmid (as 2^(-deltaCt)). Based on the quantitative data, correlation with the digital expressions (TSS-tag counts) was calculated by linear regression. Validation analysis of the induction fold was similarly performed without the plasmid control. RNA independently isolated from the DLD-1 cells cultured in similar hypoxia (1% O_2) and normoxia (21% O_2) conditions were used. The samples were normalized according to the total amount of the first strand cDNAs and were subjected to the real-time RT–PCR.

For the individual oligo-cap RACE, similarly isolated total RNAs were olio-capped with the RNA oligo (5'-AGCAUCGAUGUGCCUUGUUGGCCUACUGG-3') by the standard protocol. After the DNaseI treatment, the first strand cDNA was synthesized using random hexamer primers. One nanogram of the first strand cDNA was used for the PCR using the 5'-end primer 5'-AGCATCGAGTCGGCCTTTGTTG-3' and the gene-specific 3'-end primers used for the real-time RT–PCR.

For validation experiments using microarray, RNAs were isolated from the DLD-1 cells cultured in similar hypoxia (1% O_2) and normoxia (21% O_2) conditions. The RNAs were further processed according to the manufacturer’s instruction Using the Agilent Human Gene Expression Array G4112F platform.

For the microarray analysis, for each of the total RNA preparations (from 1% and 21% O_2 conditions), 700 ng of total RNA was used for the labeling according to the manufacturer’s instruction. The normalization was done at the sample preparation step. The following signal intensity processing was performed using GeneSpring (Agilent) with default parameters. The cut-offs used in this study was either 5-fold or 2.5-fold (Figure 2B). The experiments were repeated twice with the labeling dye exchanged.

For the comparison with the previous microarray studies, we retrieved the data from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) [a: GDS2758-61 (7); b: GDS1209 (8); c: GDS2018 (6); d: GDS1779 (9), for the GEO accession numbers and references, respectively]. We examined the original papers and prepared the list of the ‘hypoxia induced genes’ from each of the datasets. Using these datasets, the overlap between the genes identified as ‘hypoxia induced’ by this study and the previous studies was examined.

RESULTS
Construction of a TSS-tag library
By combining the oligo-capping method with a massively parallel sequencing technology, Illumina GA sequencer, we developed a simple method to collect TSS information together with a quantitative analysis of the expression levels of the transcripts (digital expression profile) in an extremely high-throughput manner (Figure 1). First, the primer sequence necessary for the sequencing was directly introduced at the 5'-ends of capped transcripts by replacing the cap structure with a cap-replacing RNA oligo (11). Then, cDNA was synthesized using random hexamers, amplified with 15 cycles of PCR and directly introduced into the sequencer without cloning (for the detailed protocol, see Materials and Methods section). The 36-base long tags corresponding to the 5'-ends of transcripts were generated by the sequencer at the rate of 10–30 million TSS-tags per run. This simple procedure eliminates laborious cloning step and allows us to easily monitor the genome-wide positions of TSSs. Furthermore, the number of TSS-tags corresponds to the number of transcripts within the cell starting from that site, since each transcript has only one cap structure.

Validation of the TSS-tag library
We first validated whether the TSS-tags collected by this method correctly indicate the positions of the TSSs and whether the counts of the TSS-tags represent the expression levels of the transcripts in vivo. For this purpose, we constructed a TSS-tag library from human embryonic kidney 293 (HEK293) cells. In total, we generated 10,401 151 TSS-tags which were uniquely and perfectly (with no mismatch) mapped to the human genome (hg 18; UCSC Genome Browser). We compared the mapped position of the TSS-tags with 18,001 protein-coding RefSeq gene models. Genomic coordinates of exons and other information of the RefSeq transcripts are as described in hg18.
are represented as grey boxes. For further information, see Supplementary Data. Gppp: cap structure. AAA: polyA.

Figure 2A, 8,647,513 (83%) of the TSS-tags were mapped within the RefSeq regions. Among them, 2,255,507 (26%) and 4,647,102 (54%) of the TSS-tags were mapped upstream and inside regions of the first exons, respectively. 739,319 (9%) of the TSS-tags were mapped to intronic regions of the RefSeq gene models, which may correspond to the TSSs of unknown alternative promoters, because there should be rare chance that they are derived from broken-down products of the mRNAs [for further discussion, see (15)]. Also, these numbers resemble the results from our previous analysis using 1.8 million 5'-ESTs (15). We observed no significant difference in the size of representative mRNAs between the TSS-library and the HEK293 oligo-cap cDNA library, which was constructed using random primers (data now shown).

Many of the TSS-tags which were mapped outside of the RefSeq regions overlapped with cDNAs in our cDNA collection (14). In particular, at least 1374 TSS-tag sites (mapped positions of the TSS-tags) overlapped with the 5' ends of the 5'-ESTs [also see Supplementary Figure 1 for further details]. Of these, 80 TSS-tag sites overlapped with our completely sequenced cDNAs. For the latter cases, average length of the representative cDNAs was 2323 bp. Of these, 55 (70%) cDNAs were spliced and, in 53 (66%) cDNAs, the longest protein-coding region was less than 100 amino acids (300 bp). Therefore, many of those intergenic TSS-tag should represent so-called mRNA-like non-protein-coding transcripts (14,24,25). We further compared the TSS-tag sites with the 5'-end data from the RNA-Seq analysis (26) and the CAGE analysis (19), which have been the only two studies that produced comparable amount of the TSS information. Among the TSS-tag sites in our dataset, 1456 sites overlapped with the ‘5’ extension’ data of the RNA-Seq analysis, of which 1105 sites also overlapped with the CAGE data. Although biological functions of many of those transcripts still remain elusive, the TSS-tags correctly represented the TSSs of previously identified transcripts.

We also wished to directly demonstrate the correct identification of the TSSs by luciferase reporter gene assays of the upstream regions of the TSS-tag sites and by real-time (quantitative) RT–PCR assays. In our previous study, we reported systematic luciferase assays in HEK293 cells (27). Among our TSS-tag dataset, luciferase data was available for the upstream regions (1 kb-upstream) of 359 TSS-tag sites. As shown in Figure 2B, distribution of the promoter activities for the 359 TSS-tag sites was clearly distinct from that of randomly isolated genomic fragments. Especially we observed clear promoter activities even for 14 TSS-tag sites with which no 5' exons of the RefSeq gene models overlapped and for six additional TSS-tag sites which were located more than 50 kb apart from any of the RefSeq genes.

We then validated whether real-time RT–PCR primers targeted at the TSS-tag sites with no RefSeq gene support could detect transcripts, and to what extent the quantitative data are correlated with the TSS-tag counts. For the purpose of quantifications, we selected TSS-tag sites which overlapped with the 5'-ends of cDNA clones in our cDNA collection. We performed real-time RT–PCR using immediately downstream sequences of the TSS-tags for the 5'-end PCR primers (Figure 2C and D). We observed clear real-time RT–PCR signals for 80 TSS-tag sites within the RefSeq regions but outside of the 5'-ends of RefSeq gene models, and for 25 TSS-tag sites mapped outside of the RefSeq regions (overall success rate was 78%). We also performed independent oligo-cap RACE analysis and, for 21 TSS-tag sites (out of 25 cases attempted), we confirmed amplification of the cDNA fragments of the expected lengths. We further quantified the absolute expression levels of those 105 (80 + 25) TSS-tag sites by using the individually isolated and quantified cDNA plasmids as controls. As shown in Figure 2C, we observed that the correlation of the absolute expression

Figure 1. Scheme of 5'-end sequencing using the Illumina GA Sequencer. Adaptors containing necessary sequence for the Illumina GA sequencer are represented as grey boxes. For further information, see Supplementary Data. Gppp: cap structure. AAA: polyA.
levels calculated by the TSS-tag counts and the real-time RT–PCR are generally well-correlated, although we also observed deviations in some cases (also see Discussion section).

Based on these results, we concluded that our TSS-tag library analysis should be reliable and useful for identifying both the TSS positions and their corresponding expression levels.

Application of the TSS-tag library for the analysis of hypoxia responses in a colon cancer cell line

Taking advantage of this new method, we wished to reveal the dynamic nature of the human gene transcriptome in a focused cell type with particular environmental perturbations. We performed genome-wide analysis of the alterations in both promoter usage and expression levels of the transcripts invoked by hypoxia. In a human colon cancer cell line, DLD-1 cells, expression of a well known hypoxia-induced gene, VEGF, is induced under hypoxic culture conditions with and without transfection of siRNAs targeted to HIF1A or HIF2A. This experimental design is the same as the previous studies of other groups (5,7). We generated 15–19 million 36-base TSS-tags per condition (Table 1). A summary of the sequence quality is shown in Supplementary Table 1.

Overall mapping patterns were similar to the case of the HEK293 library (Table 1). For example, in the case of the ‘hypoxia with non-targeted RNAi’ library, 19,213,284 TSS-tags constituted 2,610,785 unique TSS-tags. These unique TSS-tags were further clustered into 1,428,455 TSCs. Of these TSCs,
Supplementary Figure 1; overlap with the 5
senting noise-level transcriptions in the cell (see
within these TSC were usually one or two, possibly repre-
two TSC groups, are high, many of the TSS-tag counts
Although the numbers of the TSCs, especially in the latter
rest were mapped to intergenic regions (709 997; 50%) or
477 936 (33%) were mapped to the RefSeq regions. The
Statistics of the TSS-tags generated from DLD-1 cells
2254
scripts (TSSs) by the new approach possibly owing to
identified additional putative hypoxia responsive tran-
criteria, considering that this is the first analysis taking
intergenic TSCs were removed. Some of the transcripts of
conservative criteria of >1 ppm TSS-tags. One ppm corresponds to 15–20 indepen-
tag ppm increased by at least 5-fold, having more than
hypoxia-induced TSCs. Of 6366 RefSeq regions
progression. We further searched for RefSeq regions
these regions and chromatin remodeling accompanied by
Supplementary Figure 6). We also noticed distal regions
less of their exon–intron structure (lower panel; also see
in the genic regions shown in Figure 4C occurred regard-
transcription was activated on hypoxia
induced TSCs particularly clustered. We identified 54
genomic regions in which hypoxia-
induced TSCs particularly clustered. We identified 54
genomic regions in which seven or more hypoxia-induced
TSCs clustered in a 100-kb window (Figure 4A). In
these regions, transcription was activated on hypoxia
from both inter (Figure 4B) and intra (Figures 4C and
4D) genic regions. Furthermore, transcription activation
in the genic regions shown in Figure 4C occurred regard-
least two of the previous studies (6–9), were detected so
in our tag-based approach.
Using the digital-expression data, we identified 9870
hypoxia-induced TSCs in total. Among them, 6366
(64%) were mapped to RefSeq regions on the sense
strand (for the full list of the induced TSCs, see
Supplementary Table 12). The rest were mapped to inter-
genic regions or anti-sense of the RefSeq regions. Gene-
rich chromosomes 17 and 19 had the largest number of
both genic and intergenic hypoxia-induced TSCs per genic
and intergenic base of the chromosomes (Supplementary
Table 3).
We found some genomic regions in which hypoxia-
induced TSCs particularly clustered. We identified 54
genomic regions in which seven or more hypoxia-induced
TSCs clustered in a 100-kb window (Figure 4A). In
these regions, transcription was activated on hypoxia
from both inter (Figure 4B) and intra (Figures 4C and
4D) genic regions. Furthermore, transcription activation
in the genic regions shown in Figure 4C occurred regard-
less of their exon–intron structure (lower panel; also see
Supplementary Figure 6). We also noticed distal regions
of the chromosomes frequently have such ‘hot regions’
(Figure 4A; also see Supplementary Table 4). There
might be cross-talk between transcription activation in
these regions and chromatin remodeling accompanied by
telomere elongation (29), which is a hallmark of cancer
progression. We further searched for RefSeq regions
with multiple induced TSCs. Of 6366 RefSeq regions
that contained at least one hypoxia-induced TSC, 131
regions had five or more hypoxia-induced TSCs, reflecting
477 936 (33%) were mapped to the RefSeq regions. The
rest were mapped to intergenic regions (709 997; 50%) or
to anti-sense regions of the RefSeq genes (240 522; 17%).
Although the numbers of the TSCs, especially in the latter
two TSC groups, are high, many of the TSS-tag counts
within these TSC were usually one or two, possibly repre-
sent-
the presence of multiple hypoxia-induced transcriptions start in a single gene. Collective transcriptional induction events as represented in Figure 4 should not be extremely rare.

**Putative hypoxia responsive non-protein-coding transcripts**

Among the 9870 hypoxia-induced TSCs, 3504 were located at least 50 kb away from any protein-coding genes in the same strand, thus they seemed driving non-protein-coding transcripts (25,30) (also see the legend for Supplementary Figure 1). We first searched for hypoxia-induced TSCs located in the proximal regions of previously reported intergenic miRNAs using miRBase (31). We found only two such cases; TSCs 7 kb upstream of hsa-mir-612 and 1 kb upstream of hsa-mir-675 were up-regulated by 18-fold and 8.7-fold, respectively. The latter TSC actually corresponded to the TSS of the H19 non-coding RNA, which is consistent with the recent finding that H19 RNA is induced in hepatocellular carcinoma cells upon hypoxia (32,33). Similarly, we examined overlap of the intergenic TSCs with another class of non-coding RNAs, namely snoRNAs (34). We searched snoRNAdbase (35) and identified five TSCs which were located within 2 kb of regions which contained altogether nine snoRNAs (see Supplementary Table 5A). Most of them were reported to be involved in maturation of ribosomal RNAs, indicating a possibility that general translational machinery might be altered in response to hypoxia.

Although TSS-tag numbers were low for most of the newly found hypoxia-induced intergenic putative non-protein coding transcripts (Supplementary Figure 1), there were still a number of cases where expression and induction levels were at similar levels to the above two cases (33 ppm and 175 ppm for hsa-mir-612 and 675, respectively). There were 220 TSCs with TSS-tags of >10 ppm (10 ppm corresponds about three copies per cell, assuming 3 x 10^5 transcripts within a cell; also see Supplementary Table 5B). Indeed, among those 220 TSCs, four overlapped with our completely sequenced cDNAs (Supplementary Table 5C), whose average length was 1974 bp and, for all of which the longest potential open reading frame was less than 150 amino acids (450 bp). It is also noteworthy that the number 220 is in the similar range of the number of hypoxia-induced protein coding genes (see below).

In order to further characterize these hypoxia-responsive TSCs, we analyzed the correlation of their fold inductions against the most proximal protein-coding genes (see below).
Figure 4. Hypoxia-induced TSCs for putatively non-protein-coding RNAs. (A) Genomic positions of the regions in which activated TSSs highly concentrated (red circle). Number of RefSeq genes overlapping the corresponding 100 kb region is shown in the left margin. Examples of regions in which large numbers of transcription initiation sites were induced by hypoxia even in intergenic regions [(B): a 100 kb region in Chromosome 8] and inside genic regions [(C): a 100 kb region in Chromosome 17]. The vertical axis represents fold induction of the TSS-tag counts. TSSs of the genic region of NM_019613 (WDR45-like protein gene) are shown in the bottom margin. The direction of the transcription of the RefSeq gene is represented by a red arrow. Radius of each circle represents the number of TSS-tags. Colour of each circle indicates the direction of the transcription (red: same direction with the RefSeq gene; blue: opposite direction of the RefSeq gene). Putative alternative promoters on which confirmation analysis by real-time RT-PCR is shown in (D) are indicated in red and blue letters (AP1-3). AP: Alternative Promoter. (D) Real-time RT-PCR analysis of the putative alternative promoters, AP1-3, shown in (C). Fold inductions calculated by TSS-tag counts and real-time RT-PCR are shown in the third and fourth column, respectively. Primer sequences are shown in Supplementary Table 10. Note that we used first strand single-strand cDNA as template, so that the PCR amplification should be strand-sensitive. (*) Also note that fold inductions estimated for AP3 by TSS-tag counts were the sum of the upstream promoters. As the AP3 were located inside of the last exon, it was impossible to design PCR primers which discriminate the transcript products of AP3 from those of other upstream promoters. Results of the independent oligo-cap RACE analysis for each of the APs are also shown in Supplementary Figure 6.
Among 6366 hypoxia-induced TSCs, which were located within RefSeq genes, 441 TCS had TSS-tags >10 ppm. Among them, 191 had expression levels of particular individual alternative promoters significantly altered (>5-fold when the individual TSCs were evaluated) while the total gene-expression levels were not changed (<5-fold when the TSCs belonging the corresponding genes were totalled). A list of the promoters is shown in Supplementary Table 6. Figure 5 shows a typical but biologically interesting case in which alternative promoters were employed differentially between hypoxic and normoxic conditions. In the p53 tumor suppressor gene family, usage of alternative promoters has been reported for the p73 and p63 genes. In these genes, the upstream promoters encode functional transcriptional activator (TA) proteins and the downstream promoters encode non-functional silencers (DN) (36). We observed clear differential usages of these alternative promoters. In particular, in both of the p63 and p73 genes, the upstream (TA) promoters were down-regulated by hypoxia, while the downstream (DN) promoters were up-regulated. These results were consistent with the previous report that TA-p63 down-regulates and DN-p63 up-regulates VEGF expression (36). We observed that TSS-tag counts of the VEGF gene were induced by 9-fold by hypoxia. In addition to its pivotal roles in regulating cell cycle and apoptosis, p53 is also reported to be involved in modulating the balance between the respiratory and glycolysis pathways by controlling the expression levels of several downstream effectors, including COX complex mitochondrial respiratory genes (37). In this study, 60% reduction of TSS-tag counts for the COX complex assembly gene, the SCO2 gene, was observed. In DLD-1 cells, we observed neither differential usage of the alternative promoters nor overall gene-expression change in the p53 gene itself. It has been reported that the protein-coding sequence of p53 is mutated and the p53 protein product is non-functional in DLD-1 cells. p63 and p73, which share a well-conserved DNA-binding domain with p53, may complementarily regulate downstream target genes of p53. A drastic shift of the p63 and p73 usage from TA to DN supposedly contributes to adaptation of cancer cells to hypoxia.

**HIF cascade in hypoxia responsive genes**

There were 120 protein-coding genes whose expression levels were induced >5-fold and >10 ppm by hypoxia. These are the putative 'hypoxia-induced genes' selected using the strict criteria (a list of the genes and their annotations are shown in Supplementary Table 7; note that some of the previously identified hypoxia induced genes are not included there because the either fold induction was below 5 or expression level was below 10 ppm). We examined whether any of the gene groups were enriched in these 'hypoxia induced genes' for particular Gene Ontology categories (38) or KEGG (39) pathways. We found that 'glycolysis' related genes were particularly enriched (P < 0.002 and P < 0.0008 for GO and KEGG categories, respectively, by calculating hypergeometric distributions). We also examined the fold induction of all of the genes belonging to this gene category. We found that distribution of the fold inductions were statistically significantly deviated compared to other gene groups (Figure 6; P < 0.06 and P < 0.002 for GO and KEGG categories, respectively, by Wilcoxon signed rank test; see also Supplementary Figure 3 and Supplementary Table 8). Interestingly, while the genes encoding enzymes which enhance glycolysis were ubiquitously up-regulated under hypoxia, only FBP, which codes for glycolysis-suppressing fructose-1,6-bisphosphatase, was strikingly down-regulated. On the other hand, genes encoding the enzymes involved in the Complex I of oxidative phosphorylation in mitochondria were down-regulated (Supplementary Figure 4). Although it is a well-known fact that the glycolysis pathway is activated in response to hypoxia, shifting metabolism from oxygen-requiring oxidative phosphorylation to oxygen-independent glycolysis to obtain ATP (40), this is the first report to quantitatively measure gene expression changes (or system-perturbation of a particular gene network) in terms of the absolute copy number for each gene component.

We then compared changes in TSS-tag counts by transfecting siRNAs targeting HIF1A and HIF2A to evaluate dependency of hypoxia-induced gene-expression levels on HIF transcription factors. Expression of both HIF1A and HIF2A was suppressed by about 70% according to the TSS-tag counts and real-time RT-PCR analysis (Supplementary Table 8; also see Supplementary Figure 7). Among the 120 hypoxia-induced genes, 15 genes were identified with mRNA levels reduced by 80% by RNAi of HIF1A. Meanwhile, HIF2A RNAi caused reduction of mRNA levels of 36 genes. We also examined the sequences of the regions proximal to their TSSs (1 kb upstream to 200 bases downstream) and found clear consensus sequences of the HIF1 and HIF2-binding sites (41) in 11 (79%) and 31 (86%) cases, respectively (a list of the genes is shown in Supplementary Table 9). Although further compilation of the experimental data is obviously essential...
before concluding they are actually direct binding sites of the HIF1 and HIF2, they should be the first targets for exploring the transcriptional network mediated by HIF1 and HIF2.

Interestingly, all of the 15 ‘HIF1A-dependent’ genes were also suppressed by HIF2A RNAi. On the other hand, only one-third (12 out of 36 genes) of ‘HIF2A-dependent’ genes were suppressed by HIF1A reduction. Under hypoxia, the total number of TSS-tags corresponding to HIF1A was increased by 1.5-fold, and HIF2A RNAi reduced the HIF1A-expression level by 60%. Meanwhile, HIF2A expression was not significantly increased by hypoxia, and HIF1A RNAi did not reduce the HIF2A-expression level. These results suggest that HIF2A may regulate hypoxia-induced HIF1A expression. Thus, the effect of hypoxia-activated HIF2A appears to be transmitted to downstream hypoxia-responsive genes not only directly but also

---

**Figure 5.** Hypoxia-induced TSCs in p53 family genes. (A) Count of TSS-tags mapped at the corresponding genomic regions. Red and blue solid bars represent the TSS-tag counts from normoxia and hypoxia, respectively. Exon–intron structures of the RefSeq transcripts are shown in the bottom margins. The genome regions depicted in the bar graphs are magnifications of the regions indicated by thin blue lines. Whether the transcripts from the corresponding promoters should encode the transcriptionally active (TA) or negative (DN) forms is shown in the margin. AP: alternative promoter. (B) Validation of the results shown in (A) by real-time RT–PCR analysis. Promoter ID is as of those represented in the bar graphs in (A). For p63, PCR primers were set in exon 3, so that TA-type transcripts are selectively amplified. Primer sequences are shown in Supplementary Table 10.
indirectly via HIF1A in DLD-1 cells. Previous studies have assumed pivotal roles for HIF1A; while the roles of HIF2A remain mostly uncharacterized, perhaps because of its low expression level (42). The total TSS-tag number of HIF1A was about 4-fold larger than that of HIF2A in this study. In contrast to previous estimates, the high sensitivity of our method may have revealed that the hitherto-supposed ‘minor’ HIF2A plays a dominant role in the hypoxia response.
In order to further investigate the biological relevance of the cellular responses to hypoxia observed in DLD-1, we performed similar analysis using three different cell types; MCF7, HEK293 and TIG3 cells. These cells are breast cancer epithelial cells, non-cancerous immortalized embryonic kidney epithelial cells and normal (primary) embryonic lung fibroblasts, respectively. We constructed a series of TSS-tag libraries from these cells cultured under 21% and 1% O2. From each of the libraries, 8–15 million TSS-tags were generated. Overall qualities of the TSS-libraries were similar to those of the DLD-1 libraries (Table 2).

Using these new TSS-libraries, we examined gene-expression changes invoked by hypoxia in the different cells. The numbers of ‘hypoxia responsive’ genes (as is the case of the DLD-1: >5-fold induction; >10 ppm) significantly differed between the cell types (Table 3A). Eighty-six genes were ‘hypoxic induced’ in MCF7 cells, while far less genes were induced in HEK293 cells and TIG3 cells. Many of the hypoxia responsive genes in MCF7 cells overlapped with those in DLD-1 cells, while the overlaps in the other cell lines were very scarce. Particularly, we focused on the ‘glycolysis pathway’ (Figure 6) and observed significant difference in the expression changes between the cell types in this pathway. For the genes belonging to the glycolysis pathway, gene-expression changes in HEK293 cells and TIG3 cells were smaller than in DLD-1 cells in terms of fold inductions as well as absolute gene-expression levels, while those of MCF7 cells were almost at the level of DLD-1 cells (Table 3B).

We could also identify ‘hot regions’ in these cell lines (Table 3C). However, the number of ‘hot regions’ was different between the cell types. Particularly, there were far more ‘hot regions’ in DLD-1 and MCF7 cells than in TIG3 and HEK293 cells. Interestingly, three of the ‘hot regions’ overlapped between DLD-1 cells and MCF7 cells (Supplementary Table 4A), and should thus be prioritized for further functional characterizations.

**DISCUSSION**

We have described a simple method to massively collect positional information of TSSs together with digital information of the expression levels of the transcripts. By this approach, time, costs and efforts necessary for laborious cDNA cloning and sequencing steps could be greatly reduced. Most part of the technical difficulties to construct
a full-length cDNA library or a 5′ SAGE or CAGE library could be skipped. Although other cap selection methods, such as the cap-trapper (10) and Smart system (43), can be also applied for massively parallel sequencing systems, our oligo-capping method has a clear advantage. Among those similar methods, only oligo-capping includes a step to replace the cap structure with synthetic oligo, in which sequence necessary for massively parallel sequencing can be embedded. Therefore, the protocol presented here should be applicable for any other massive sequencing technologies.

This approach has several advantages compared to the current expression profiling methods. Compared with microarray-based or real-time PCR-based approaches, our method does not need any probes or PCR primers, which should be designed based on presumed transcript sequences, and thus prevent the detection of novel transcripts with these previous methods. Also, while the previous methods are designed to detect relative change in expression of the same transcript between two states, absolute quantification of the transcripts could be enabled only by our method. Compared with the recent RNA-Seq method (26,44,45), our method has two major advantages and one clear disadvantage. Advantages are: (i) exact positional information of the TSS can be obtained; (ii) thorough of the expression analysis is better because our method does not sequence internal part of transcripts. A disadvantage is that our method cannot detect the splicing pattern of the exons.

A series of validation analyses showed that the data from our new method is quite reliable (Figures 2 and 3). However, in some cases, we also noticed that there were some discrepancies (Figures 2B and 3A). Because we did not use redundantly mapped TSS-tags, we may have incorrectly assigned small number of TSS-tag counts, when a real TSS is located within repetitive sequence elements. Conversely, the expression level could have been overestimated, when a small population (but a large number) of TSS-tags deviated from a huge TSS cluster by sequence errors, which would be mapped elsewhere otherwise, were uniquely mapped at the corresponding gene region. Careful evaluation is crucial especially when the redundantly mapped tags would be rescued (46). In either case, confirmation analysis on individual genes should be essential, as was the case with microarray analysis in its early days.

Taking advantages of our new method, we revealed genome-wide changes of the transcriptional landscape in response to hypoxia for the first time. (All the sequence data and the cluster data will be made freely available from our web site (DBTSS: http://dbtss.hgc.jp/) and from NCBI Short Read Archives (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi/) under the accession number of SRA003625. Visualization of some of the results for each gene is also available there (for example, see Supplementary Figure 5). In our analysis, we identified ‘hot regions’ where hypoxia-induced promoters are enriched in particular genomic regions, as well as ‘hot’ genes which have many hypoxia-responsive alternative promoters. It is possible that hypoxia-invoked chromosomal changes came to allow access of transcriptional factors in a somewhat global manner. Consistently, some of the transcriptions from proximal regions, occasionally including transcriptions of putative non-protein-coding transcripts, seemed to be under similar regulation (Supplementary Figure 3), with the extreme cases being the above-mentioned ‘hot regions’. These observations could be explained if the surrounding chromosome context, which shapes the transcriptional landscape proximally, is shared between the TSCs of non-protein-coding transcripts and the TSCs of RefSeq transcripts.

It is noteworthy that some of such ‘hot regions’ were also identified from different cell types of distinct cancer origin, though number and frequencies of them were different. It should be important to further analyze cells of other mammalian species under hypoxia to see whether these ‘hot regions’ or ‘hot’ genes are evolutionarily conserved. Genome-wide high-throughput methods to monitor DNA binding of proteins (42), DNA and histone modifications (47,48) and DNase I hyper sensitive sites (49) or combination of them (50), will be needed for directly analyzing chromosomal structural changes.

It should be also noteworthy that the gene-expression changes were somewhat similar between DLD-1 and MCF7 cells, though they were distinct from HEK293 and TIG3 cells. Both DLD-1 and MCF7 cells were derived from solid tumour, which may have originally grown in hypoxic conditions. The enhanced gene-expression changes observed in DLD-1 and MCF7 should explain the distinct biology of the cells in response to hypoxia.

Indeed, various new types of analyses have been enabled by the hereby described method, in which detection of TSS positions and digital-expression information can be obtained without a prior knowledge of transcript structures. Although this is only the first step towards monitoring dynamic behavior of the human transcriptome, our new method and its application will supply a unique tool for thorough understanding of the dynamic nature of the transcriptional program encoded by the human genome.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS
We are grateful to K. Abe, E. Sekimori, K. Imamura and H. Onozuka for technical supports. We are also thankful to Dr Sierro, Dr M. Frith and Mr F. Sathirapongsasuti for critical reading of the manuscript.

FUNDING
This work was supported by grants from the Ministry of Economy, Trade and Industry (METI) of Japan, the Japan Key Technology Center project of METI of JAPAN, Grants for the Third-Term Comprehensive 10-Year Strategy for Cancer Control and Grants-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare and a Grant-in-Aid for Scientific Research on

Conflict of interest statement. None declared.

REFERENCES


