Published online 22 February 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 7* ***2249–2263***
doi:10.1093/nar/gkp066

Massive transcriptional start site analysis of human genes in hypoxia cells

Katsuya Tsuchihara[1], Yutaka Suzuki[2,*], Hiroyuki Wakaguri[2], Takuma Irie[2], Kousuke Tanimoto[2], Shin-ichi Hashimoto[3], Kouji Matsushima[3], Junko Mizushima-Sugano[2,4], Riu Yamashita[5], Kenta Nakai[5], David Bentley[6], Hiroyasu Esumi[1] and Sumio Sugano[2]

[1]Cancer Physiology Project, Research Center for Innovative Oncology, National Cancer Center Hospital East: 6-5-1 Kashiwanoha, Kashiwa, Chiba 277-8577, [2]Graduate School of Frontier Sciences, the University of Tokyo: 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, [3]Graduate School of Medicine, the University of Tokyo: 7-3-1 Hongo, Bunkyokyu, Tokyo 113-0033, [4]Laboratory of Molecular Virology, Kitasato Institute for Life Sciences, Kitasato University: 5-9-1 Shirokane, Minatoku, Tokyo 108-8641, [5]Institute of Medical Science, the University of Tokyo: 4-6-1 Shirokenedai, Minatoku, Tokyo 108-8639, Japan and [6]Illumina, 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 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 20-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 mean 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% O_2 and 5% CO_2 at 37°C for 24 h followed by incubation in 21% O_2 or 1% O_2 and 5% CO_2 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′-AAUGAUACGGCGACCACCGAGAUCUACACU CUUUCCCUACACGACGCUCUUCCGAUCUGG-3′) using 250 U T4 RNA ligase (TaKaRa) at 20°C for 3 h. After the DNase I treatment (TaKaRa), polyA-containing RNA was selected using oligo-dT powder (Collaborative). First strand cDNA was synthesized from 10 pmol of random hexamer primer (5′-CAAGCAGAAGACGGCA TACGANNNNNNC-3′) using Super Script II (Invitrogen) by incubating at 12°C for 1 h and at 42°C overnight. Template RNA was degraded by alkarine 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′-AAT GATACGGCGACCACCGAG-3′ and 5′-CAAGCAGA AGACGGCATACGA-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 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 totalled. 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-snorna.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 that 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 fold induction 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 oligo-capped with the RNA oligo (5′-AGCAUCGAGUCGGCCUUGUUGGCCUACUGG-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′-AGCATCGAGTCGGCCTTGTTG-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 instructions 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 records 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

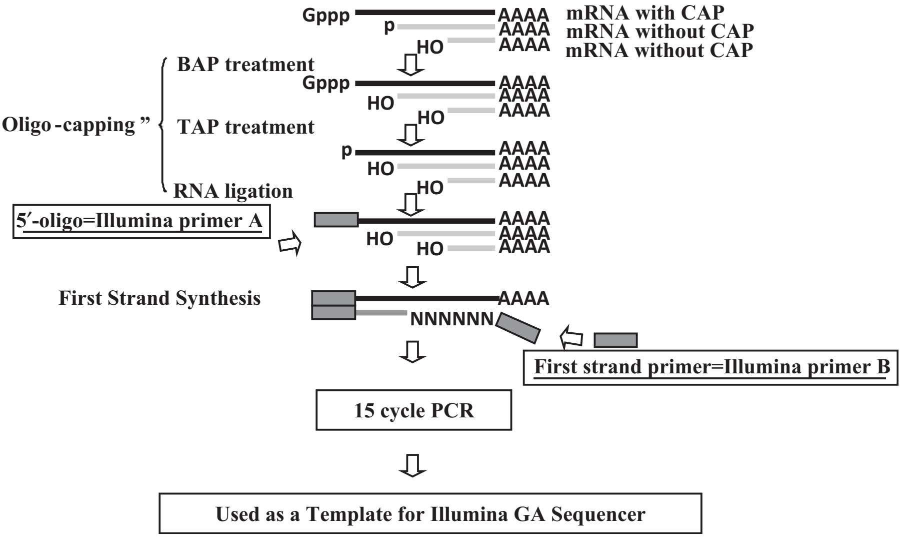


Figure 1. Scheme of 5′-end sequencing using the Illumina GA Sequencer. Adaptors containing necessary sequence for the Illumina GA sequencer are represented as grey boxes. For further information, see Supplementary Data. Gppp: cap structure. AAA: polyA.

as of UCSC Genome Browser (for the version information, see Materials and Methods section). As shown in 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-tags 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

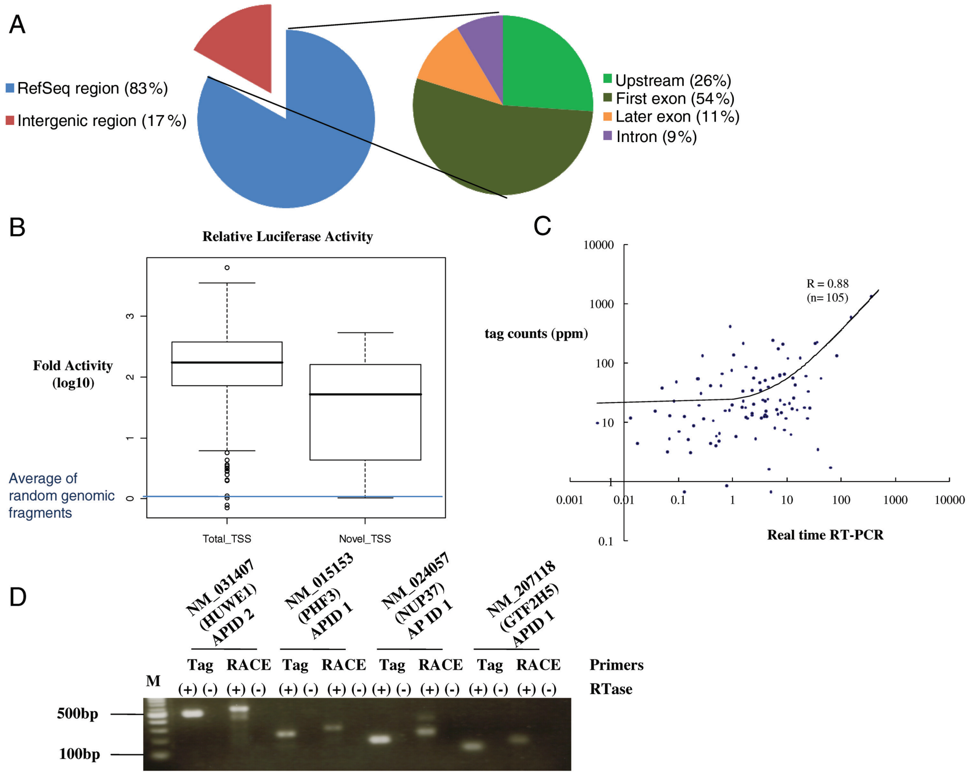


Figure 2. Validation analyses of the TSS-tag library. (**A**) Mapped positions of the TSS-tags relative to the RefSeq genes were evaluated. Population of the TSS-tags mapped at the corresponding positions indicated by the color bars in the margin is shown. The right circle graph shows the composition of the blue section in the left circle graph. (**B**) Distribution of the luciferase activities of the upstream 1 kb regions of the TSSs (n = 351; right). Luciferase activities of upstream regions of the TSS-tags that were not supported by any RefSeq gene models are calculated separately (n = 20; left). Luciferase activities were normalized against the average luciferase activity of randomly isolated 1 kb genomic fragments (n = 251). For further details, see the reference (27). (**C**) Correlation between the TSS-tag counts and the copy number estimated by real-time RT–PCR normalized by individual plasmids (n = 105). Each value is the average of three experiments. Sequences of the used primers and quantitative data are presented in Supplementary Table 10. R: correlation-coefficient calculated by linear regression. Note that, because the graph is written in log scale and the y intersect is not 0, the liner regression line is curved where the x value is small. (**D**) Examples of the real-time RT–PCR and independent oligo-cap RACE analyses. Experimental conditions are shown in the margin. For details, see Materials and methods section. For the primer, 'Tag' indicates the PCR primer targeted to the overlapping region of the TSS-tag and 'RACE' indicates the PCR primer targeted to the cap-replacing oligo. APID: alternative promoter ID. M: molecular marker.

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 condition as well as in xenografted tumor tissues *in vivo* (4). We cultured DLD-1 cells under hypoxic and normoxic 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, 14 001 295 (73%) out of total 19 213 284 TSS-tags were mapped in the RefSeq regions. Of these, 4 310 405 (31%), 7 384 800 (53%) and 1 459 600 (10%) TSS-tags were mapped to the upstream, first exon and intron regions, respectively. Therefore, we estimated at least 84–94% of the TSS-tags represent the real TSSs in this case, too. For the purposes of the following analyses, the TSS-tags were further clustered into 500-bp bins to generate TSS clusters (TSCs) (15). In case of 'hypoxia with non-targeted RNAi' library, a total of 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,

Table 1. Statistics of the TSS-tags generated from DLD-1 cells

	Relative to RefSeq regions		Relative to exons of RefSeq gene models			
	#total mapped tag	#NM_associated tag (%)	Upstream (%)	First exon (%)	Other exon (%)	Intron (%)
Hypoxia non-targeted RNAi	19213284	14001295 (73)	4310405 (31)	7384800 (53)	846490 (6)	1459600 (10)
Hypoxia with HIF1A RNAi	17995370	13758453 (76)	4116100 (30)	6995301 (51)	1303571 (9)	1343481 (10)
Hypoxia with HIF2A RNAi	17047001	14304678 (84)	4547387 (32)	8045603 (56)	830696 (6)	880992 (6)
Normoxia with non-targeted RNAi	17878365	14194520 (79)	3850858 (27)	8449751 (60)	820989 (6)	1072922 (8)
Normoxia with HIF1A RNAi	15190726	12628363 (83)	3554553 (28)	7469575 (59)	822034 (7)	782201 (6)
Normoxia HIF2A RNAi	17175662	14117263 (82)	3702462 (26)	8810503 (62)	685561 (5)	918737 (7)
Total	104500408	83004572 (79)	24081765 (29)	47155533 (57)	5309341 (6)	6457933 (8)

Mapped positions of the TSS-tags were counted relative to RefSeq regions and relative to exons of RefSeq gene models, when mapped inside of the RefSeq regions.

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 representing noise-level transcriptions in the cell (see Supplementary Figure 1; overlap with the 5′EST data is also shown there).

Genome-wide distribution of hypoxia responsive transcripts

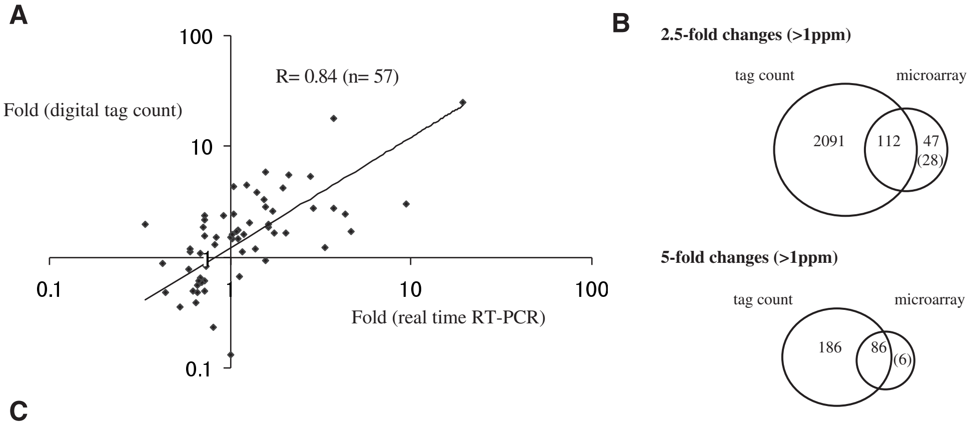
We normalized TSS-tag counts of TSCs to tags per million (ppm). In order to avoid noise level signals and possible experimental errors, we focused on TSCs for which TSS-tag ppm increased by at least 5-fold, having more than 1 ppm TSS-tags. One ppm corresponds to 15–20 independent TSS-tags per TSC depending on the dataset. By the conservative criteria of >1 ppm and >5-fold, most of the intergenic TSCs were removed. Some of the transcripts of previously identified ‘hypoxic responsive genes’ were also removed. We tentatively employed these very conservative criteria, considering that this is the first analysis taking the TSS-tag approach. However, further detailed analyses and re-evaluation of the data should be necessary on very rarely expressed intergenic transcripts, although some of them might be the system noise of the transcription machinery. For the number of ‘hypoxia-induced’ TSCs with different parameters, see Supplementary Table 2.

In order to validate the calculated fold inductions, we performed real-time RT–PCR analysis. For this purpose, RNAs were independently isolated from the DLD-1 cells cultured in similar hypoxia (1% O_2) and normoxia (21% O_2) conditions. From this analysis, again, we observed that the expression information obtained using this method was well-correlated with the results obtained using real-time RT–PCR (Figure 3A). We then performed microarray analysis and compared the obtained data with the digital expression data using independently prepared RNAs. As shown in Figure 3B, the hypoxia responsive genes detected in microarrays were mostly detected so in the digital expression profiling, too. At the same time, we identified additional putative hypoxia responsive transcripts (TSSs) by the new approach possibly owing to the improved sensitivity and coverage of the analysis (see below).

We also compared the results of digital gene-expression data with the previous microarray studies. We first searched for the data focusing on hypoxia responses of human cells in GEO database (28). Then, we examined the original papers and retrieved a list of the genes which were identified as ‘hypoxia responsive genes’ in the corresponding study. We examined overlap of the ‘hypoxia induced genes’ identified from the previous studies and from this study. As shown in Figure 3C, 11 genes, which were reported as hypoxia responsive genes in at 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 intergenic 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 regardless 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



C

NM ID	definition	fold: hypoxia/normoxia (TSS-tag count)	fold: hypoxia/normoxia (realtime RT-PCR)	fold: hypoxia/normoxia (microarray in this study)	reference
NM_001124	adrenomedullin	12.5	6.1	4.1	a, b, d
NM_005165	fructose-bisphosphate aldolase C	25.2	19.2	5.7	a, c
NM_022073	egl nine homolog 3	3.0	62.0	9.1	a, d
NM_181054, NM_001530	hypoxia-inducible factor 1, alpha subunit	0.7	0.6	0.9	b, c
NM_000598, NM_001013398	insulin-like growth factor binding protein 3	5.1	2.0	1.5	a, b, c, d
NM_002317	lysyl oxidase preproprotein	5.4	2.1	2.4	a, b, c, d
NM_002318	lysyl oxidase-like 2 precursor	1.7	not amplified	1.2	a, d
NM_152878,NM_012323	transcription factor MAFF	4.4	not amplified	1.3	a, d
NM_006096	N-myc downstream regulated gene 1	18.8	20.5	7.5	a, d

Figure 3. Validation analyses of the fold induction. (**A**) Correlation between the fold changes observed using digital-expression information (vertical axis) and real-time RT–PCR (horizontal axis). In total, 57 genes out of 63 total glycolysis related genes, from which we obtained meaningful data, were used for the validation. Each value is the average of three experiments. Sequences of the used primers are shown in Supplementary Table 10. R: correlation-coefficient calculated by linear regression. (**B**) Validation experiments by the microarray analysis. Overlap of the microarray results and digital gene-expression profiling is shown in the bottom margin. The numbers in the parentheses indicate the number of genes induced by more than 1.5-fold (upper) and 2.5-fold (lower), although they were not detected as 'induced' by the criteria of 2.5-fold induction (upper) and 5-fold induction (lower). The statistical significance of the overlap by calculating hypergenometric distribution was $P < 5E–67$ for 2.5-fold change and $P < 3E–15$ for 5-fold change. (**C**) Comparison with data from previous studies. For the comparison with previous microarray studies, the overlap between the genes identified as 'hypoxia induced' by this study and the previous studies was evaluated. The 'hypoxia responsive genes' in the previous studies were as of those described in the corresponding papers.

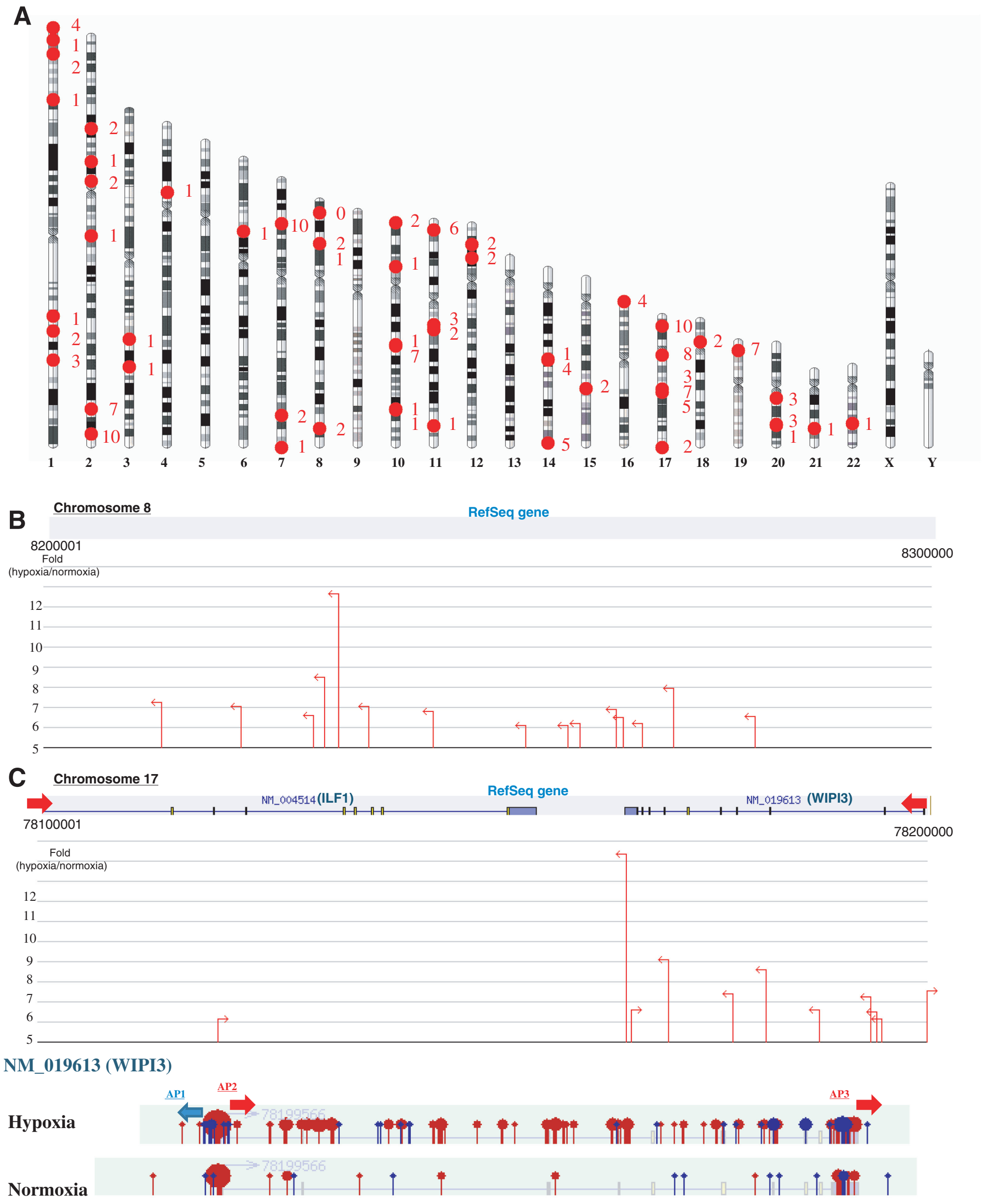
the presence of multiple hypoxia-induced transcriptions start in a single gene. Collective transcriptional induction events as represented in Figure 4 should not be extremely rare.

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 snoRNABase (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 3×10^5 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



D

	Promoter ID	direction	fold: hypoxia/normoxia (expected from TSS-tag count)	fold: hypoxia/normoxia (real time RT-PCR)
NM_019613	AP1	RV	>7.5	5.4
	AP2	FW	3.6	2.9
	AP3	FW	3.7*	6.8

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.

genes. For this, we used the 220 hypoxia-induced intergenic TSCs with TSS-tags of >10 ppm. We found that, in 28 cases, the nearest genes were also up-regulated by >2.5-fold, while down-regulation by >2.5-fold was observed only in five cases. The TSS-tag count level of the intergenic TSCs correlated with that of the nearest protein-coding genes (upper panels in Supplementary Figure 2). When other intergenic TSCs were also considered as the nearest TSCs, this correlation became even more significant. A similar tendency for co-elevation of proximal transcriptions was also observed for hypoxia-responsive TSCs which were mapped to antisense positions of RefSeq genes. Among 124 hypoxia-induced antisense TSCs with >10 ppm TSS-tags, the corresponding protein-coding transcripts were also up- and down-regulated by more than 2.5-fold in 24 cases and two cases, respectively. Again, the TSS-tag counts of putative non-coding transcripts and the corresponding antisense protein-coding transcripts were at similar levels (lower panels in Supplementary Figure 2).

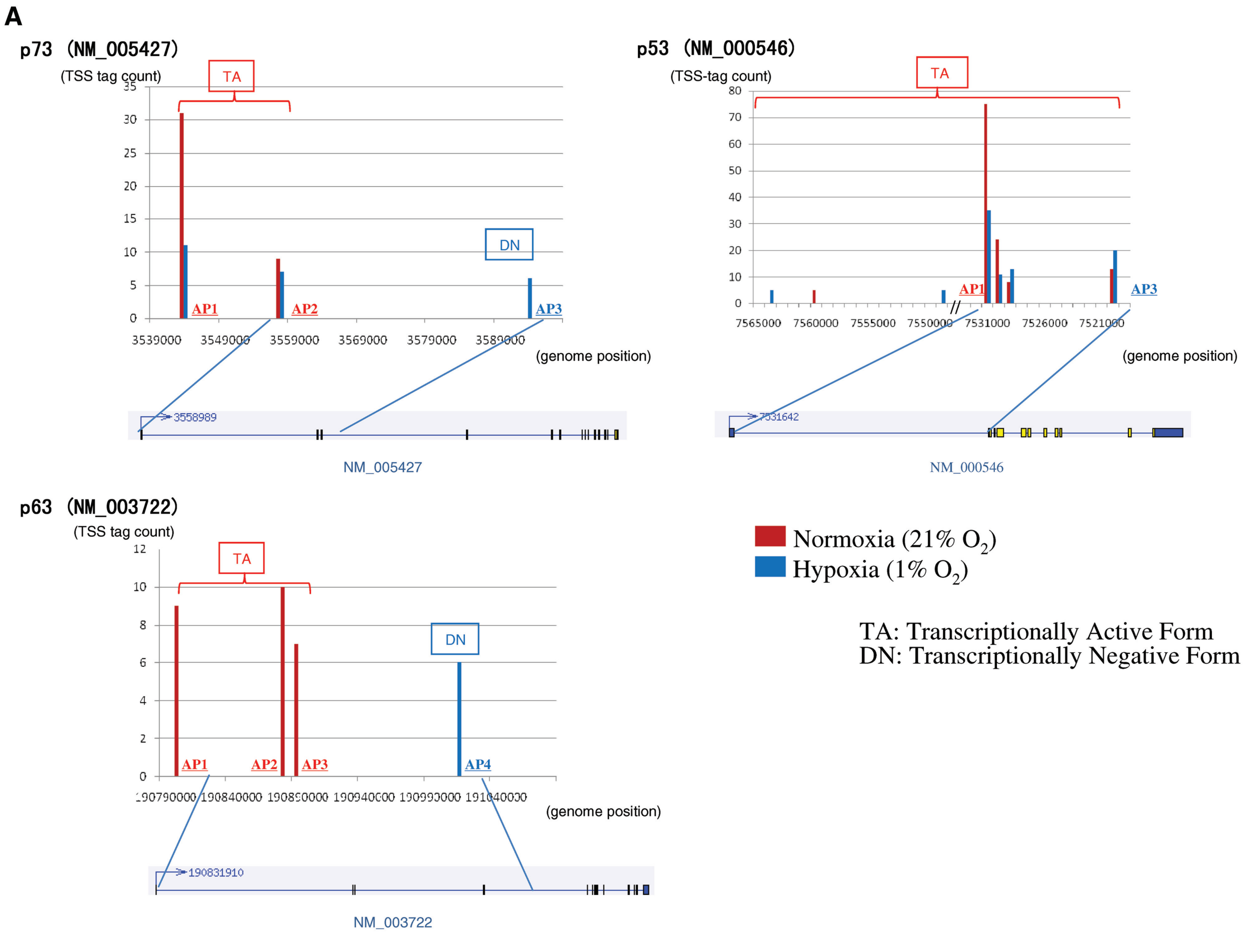
Alternative hypoxia responsive promoters

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 TAp63 down-regulates and DNp63 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; also see Supplementary Figure 3 and Supplementary Table 11). 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



B

	Promoter ID	Transcript type	fold: hypoxia/normoxia (TSS-tag count)	fold: hypoxia/normoxia (realtime RT-PCR)
p73	AP1	TA	0.4	0.6
	AP2	TA	0.8	1.0
	AP3	DN	>6	3.2
p63	AP1-3	TA	<0.1	0.06
	AP4	DN	>6	13.3
p53	AP1	TA	0.5	0.7
	AP3	TA	1.5	1.3

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.

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

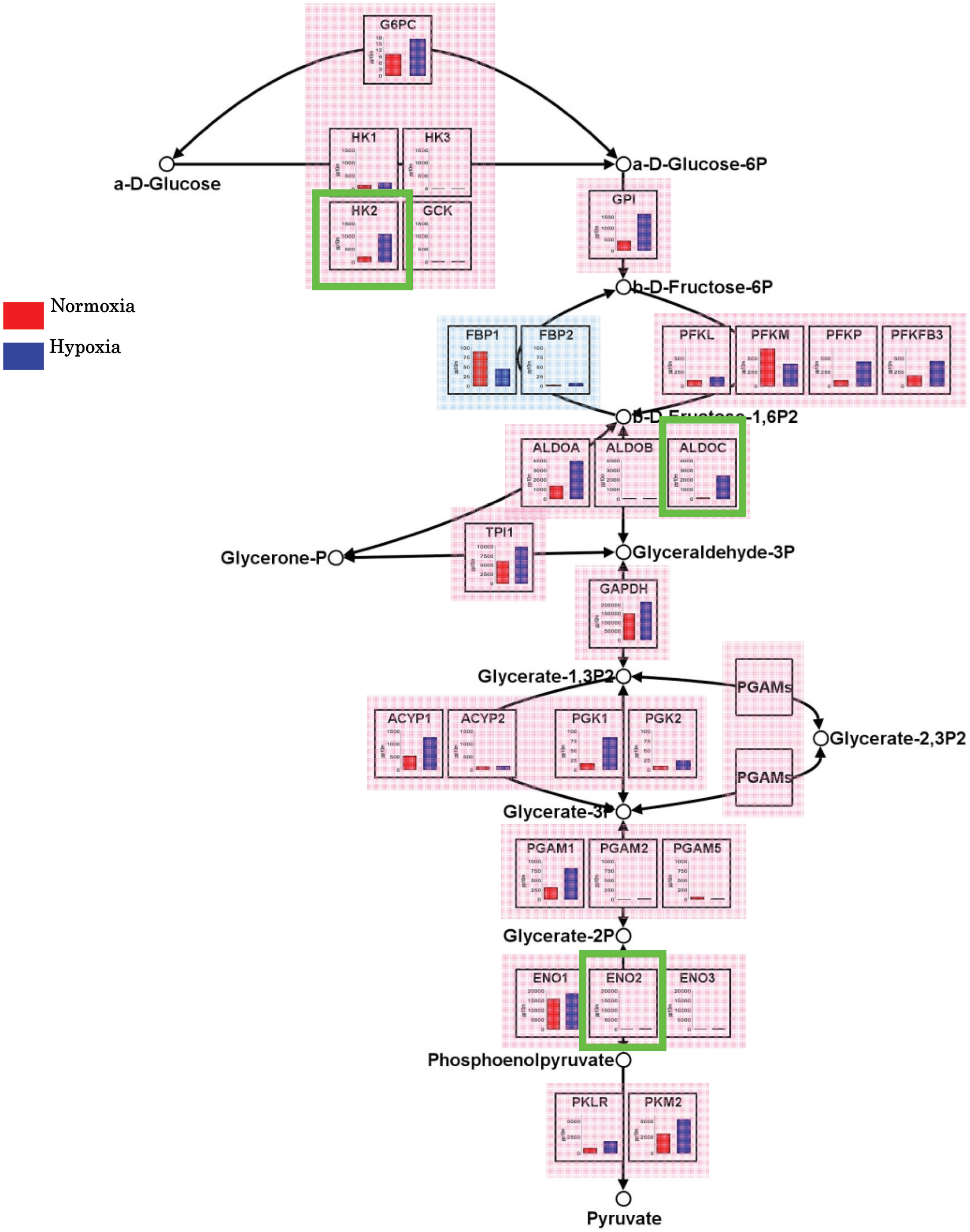


Figure 6. Hypoxia-invoked response of the glycolysis gene network. Expression of glycolysis-enhancing enzymes (masked with pale pink) was up-regulated while that of glycolysis-suppressing enzymes (masked with pale blue) was down-regulated. Human genes assigned to the glycolysis pathway map of the KEGG database were selected, and TSS-tag numbers of the corresponding genes were evaluated. Red bars and blue bars represent TSS-tag ppm in normoxia and hypoxia respectively. Fold induction for each of the genes is shown in Supplementary Table 11. Genes included in the list of the 'hypoxia-induced' 120 genes are highlighted by green boxes. Note that, since we used the stricter criteria for selecting hypoxia-induced genes (>10 p.p.m., >5-fold induction), many of the genes belonging to this pathway are not directly included in the list, though they showed inductions at least to some extent.

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.

Table 2. Statistics of the TSS-tags generated from other cell lines

	Relative to RefSeq regions		Relative to exons of RefSeq gene models			
	#total mapped tag	#NM_associated tag (%)	Upstream (%)	First exon (%)	Other exon (%)	Intron (%)
MCF7 1% O2	7950745	7259512 (91)	2221013 (31)	3859864 (53)	589298 (8)	589337 (8)
MCF7 21% O2	14189873	12955252 (91)	3828159 (30)	6974356 (54)	989360 (8)	1163377 (9)
HEK293 1% O2	10886858	10233645 (94)	3216794 (31)	5764173 (56)	786235 (8)	466443 (5)
HEK293 21% O2	8303754	7766894 (94)	2343996 (30)	4516688 (58)	593494 (8)	312716 (4)
TIG3 1% O2	9043423	8273656 (91)	1993830 (24)	4977677 (60)	799727 (10)	502422 (6)
TIG3 21% O2	9501473	8686047 (91)	2159571 (25)	5300938 (61)	657096 (8)	568442 (7)

As is the case in Table 1, mapped positions of the TSS-tags were counted relative to RefSeq regions and relative to exons of RefSeq gene models, when mapped inside of the RefSeq regions.

Table 3. The number of 'hypoxia responsive' genes identified from indicated cell lines; overlap with the 'hypoxia responsive' genes in DLD-1 is shown in the second line (Panel A); for the 'glycolysis pathway' genes (shown in Figure 6), average fold change (first line) and the TSS-tag counts in p.p.m. in hypoxic conditions (second line) were calculated (Panel B)

Panel A				
	MCF7	HEK293	TIG3	
>5-fold induction (>10 p.p.m.)	86	24	9	
Overlap with DLD1	27	3	4	
Panel B				
	DLD-1	MCF7	HEK293	TIG3
Average fold change	3.6	3.5 (*P* = 0.92)	1.2 (*P* = 7.6e-6)	2.0 (*P* = 0.11)
Average p.p.m. in 1% O2	922.9	657.8 (*P* = 0.33)	209.9 (*P* = 7.6e-7)	542.7 (*P* = 3.8e-3)
Panel C				
	MCF7	HEK293	TIG3	
'Hot region'	37	8	9	
Overlap with DLD1	3	1	0	

Statistical significances of the difference compared with the cases in DLD-1, which were calculated by paired Wilcoxon test, are shown in the parentheses.

Hypoxia responses in different cell lines

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% O_2. 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) throughput 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

Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology-Japan. Funding for open access charge: Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology Japan.

Conflict of interest statement. None declared.

REFERENCES

1. Harris,A.L. (2002) Hypoxia–a key regulatory factor in tumour growth. *Nat. Rev. Cancer*, **2**, 38–47.
2. Keith,B. and Simon,M.C. (2007) Hypoxia-inducible factors, stem cells, and cancer. *Cell*, **129**, 465–472.
3. O'Reilly,S.M., Leonard,M.O., Kieran,N., Comerford,K.M., Cummins,E., Pouliot,M., Lee,S.B. and Taylor,C.T. (2006) Hypoxia induces epithelial amphiregulin gene expression in a CREB-dependent manner. *Am. J. Physiol. Cell Physiol.*, **290**, C592–C600.
4. Mizukami,Y., Kohgo,Y. and Chung,D.C. (2007) Hypoxia inducible factor-1 independent pathways in tumor angiogenesis. *Clin. Cancer Res.*, **13**, 5670–5674.
5. Chi,J.T., Wang,Z., Nuyten,D.S., Rodriguez,E.H., Schaner,M.E., Salim,A., Wang,Y., Kristensen,G.B., Helland,A., Borresen-Dale,A.L. *et al.* (2006) Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS Med.*, **3**, e47.
6. Wang,V., Davis,D.A., Haque,M., Huang,L.E. and Yarchoan,R. (2005) Differential gene up-regulation by hypoxia-inducible factor-1alpha and hypoxia-inducible factor-2alpha in HEK293T cells. *Cancer Res.*, **65**, 3299–3306.
7. Elvidge,G.P., Glenny,L., Appelhoff,R.J., Ratcliffe,P.J., Ragoussis,J. and Gleadle,J.M. (2006) Concordant regulation of gene expression by hypoxia and 2-oxoglutarate-dependent dioxygenase inhibition: the role of HIF-1alpha, HIF-2alpha, and other pathways. *J. Biol. Chem.*, **281**, 15215–15226.
8. Detwiller,K.Y., Fernando,N.T., Segal,N.H., Ryeom,S.W., D'Amore,P.A. and Yoon,S.S. (2005) Analysis of hypoxia-related gene expression in sarcomas and effect of hypoxia on RNA interference of vascular endothelial cell growth factor A. *Cancer Res.*, **65**, 5881–5889.
9. Mense,S.M., Sengupta,A., Zhou,M., Lan,C., Bentsman,G., Volsky,D.J. and Zhang,L. (2006) Gene expression profiling reveals the profound upregulation of hypoxia-responsive genes in primary human astrocytes. *Physiol Genomics*, **25**, 435–449.
10. Carninci,P. and Hayashizaki,Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol.*, **303**, 19–44.
11. Suzuki,Y. and Sugano,S. (2003) Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.*, **221**, 73–91.
12. Kato,S., Ohtoko,K., Ohtake,H. and Kimura,T. (2005) Vector-capping: a simple method for preparing a high-quality full-length cDNA library. *DNA Res.*, **12**, 53–62.
13. Edery,I., Chu,L.L., Sonenberg,N. and Pelletier,J. (1995) An efficient strategy to isolate full-length cDNAs based on an mRNA cap retention procedure (CAPture). *Mol. Cell Biol.*, **15**, 3363–3371.
14. Ota,T., Suzuki,Y., Nishikawa,T., Otsuki,T., Sugiyama,T., Irie,R., Wakamatsu,A., Hayashi,K., Sato,H., Nagai,K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
15. Kimura,K., Wakamatsu,A., Suzuki,Y., Ota,T., Nishikawa,T., Yamashita,R., Yamamoto,J., Sekine,M., Tsuritani,K., Wakaguri,H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.*, **16**, 55–65.
16. Wakaguri,H., Yamashita,R., Suzuki,Y., Sugano,S. and Nakai,K. (2008) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res.*, **36**, D97–D101.
17. Hashimoto,S., Suzuki,Y., Kasai,Y., Morohoshi,K., Yamada,T., Sese,J., Morishita,S., Sugano,S. and Matsushima,K. (2004) 5′-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.*, **22**, 1146–1149.
18. Shiraki,T., Kondo,S., Katayama,S., Waki,K., Kasukawa,T., Kawaji,H., Kodzius,R., Watahiki,A., Nakamura,M., Arakawa,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.
19. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
20. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G., Frith,M.C. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
21. Landry,J.R., Mager,D.L. and Wilhelm,B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.*, **19**, 640–648.
22. Davuluri,R.V., Suzuki,Y., Sugano,S., Plass,C. and Huang,T.H. (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.*, **24**, 167–177.
23. Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.
24. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
25. Willingham,A.T. and Gingeras,T.R. (2006) TUF love for "junk" DNA. *Cell*, **125**, 1215–1220.
26. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science.*, **321**, 956–960.
27. Sakakibara,Y., Irie,T., Suzuki,Y., Yamashita,R., Wakaguri,H., Kanai,A., Chiba,J., Takagi,T., Mizushima-Sugano,J., Hashimoto,S. *et al.* (2007) Intrinsic promoter activities of primary DNA sequences in the human genome. *DNA Res.*, **14**, 71–77.
28. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
29. Blasco,M.A. (2007) The epigenetic regulation of mammalian telomeres. *Nat. Rev. Genet.*, **8**, 299–309.
30. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet.*, **15 Spec No 1**, R17–R29.
31. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
32. Matouk,I.J., DeGroot,N., Mezan,S., Ayesh,S., Abu-lail,R., Hochberg,A. and Galun,E. (2007) The H19 non-coding RNA is essential for human tumor growth. *PLoS ONE*, **2**, e845.
33. Cai,X. and Cullen,B.R. (2007) The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA*, **13**, 313–316.
34. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
35. Xie,J., Zhang,M., Zhou,T., Hua,X., Tang,L. and Wu,W. (2007) Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res.*, **35**, D183–D187.
36. Senoo,M., Matsumura,Y. and Habu,S. (2002) TAp63gamma (p51A) and dNp63alpha (p73L), two major isoforms of the p63 gene, exert opposite effects on the vascular endothelial growth factor (VEGF) gene expression. *Oncogene*, **21**, 2455–2465.
37. Matoba,S., Kang,J.G., Patino,W.D., Wragg,A., Boehm,M., Gavrilova,O., Hurley,P.J., Bunz,F. and Hwang,P.M. (2006) p53 regulates mitochondrial respiration. *Science*, **312**, 1650–1653.
38. The Gene Ontology (GO) project in 2006. (2006) *Nucleic Acids Res.*, **34**, D322–D326.
39. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
40. Iyer,N.V., Kotch,L.E., Agani,F., Leung,S.W., Laughner,E., Wenger,R.H., Gassmann,.M., Gearhart,J.D., Lawler,A.M.,

Yu,A.Y. *et al*. (1998) Cellular and developmental control of O2 homeostasis by hypoxia-inducible factor 1 alpha. *Genes Dev*., **12**, 149–162.

41. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M. *et al*. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*., **34**, D108–D110.
42. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
43. Barnes,W.M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl Acad. Sci. USA*, **91**, 2216–2220.
44. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
45. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
46. Faulkner,G.J., Forrest,A.R., Chalk,A.M., Schroder,K., Hayashizaki,Y., Carninci,P., Hume,D.A. and Grimmond,S.M. (2008) A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. *Genomics*, **91**, 281–288.
47. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
48. Meissner,A., Mikkelsen,T.S., Gu,H., Wernig,M., Hanna,J., Sivachenko,A., Zhang,X., Bernstein,B.E., Nusbaum,C., Jaffe,D.B. *et al*. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766–770.
49. Xi,H., Shulha,H.P., Lin,J.M., Vales,T.R., Fu,Y., Bodine,D.M., McKay,R.D., Chenoweth,J.G., Tesar,P.J., Furey,T.S. *et al*. (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet*., **3**, e136.
50. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al*. (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.