Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE: Supplementary text and methods.

Eivind Valen[1], Giovanni Pascarella[2], Alistair Chalk[3], Norihiro Maeda[4], Hiromi Sano[4], Miki Kojima[4], Chika Kawazu[4], Mitsuyoshi Murata[4], Hiromi Nishiyori[4], Dejan Lazarevic[2], Dario Motti[2,] Troels Torben Marstrand[1], Man-Hung Eric Tang[1], Xiaobei Zhao[1], Anders Krogh[1], Ole Winther[1], Takahiro Arakawa[4], Jun Kawai[4], Christine Wells[3], Carsten Daub[5], Matthias Harbers[7], Yoshihide Hayashizaki[4], Stefano Gustincich[2], Albin Sandelin[1,*], Piero Carninci[4,6,*]

Affiliations:
[1]The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaloes vej 5, DK2200, Denmark

[2]The Giovanni Armenise-Harvard Foundation Laboratory, Sector of Neurobiology, International School for Advanced Studies (SISSA), AREA Science Park, s.s. 14, Km 163.5, Basovizza, 34012 Trieste, Italy

[3]The National Centre for Adult Stem Cell Research, The Eskitis Institute for Cell and Molecular Therapies Room1.20 Eskitis Institute Building N75, Brisbane Research Park, Griffith University, Nathan Campus Kessels Rd, QLD 4111, Australia

[4] LSA Technology Development Group, [5]LSA Bioinformatics Team, [6] Functional Genomics Technology Team, Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045 Japan

[7]DNAFORM Inc., Leading Venture Plaza-2, 75-1, Ono-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0046, JAPAN

Contents:

1. SUPLEMENTARY FIGURE LEGENDS

Figure S1: Mapping of tags to known transcripts

The proportion (fraction) of tags that maps to know cDNAs from the FANTOM database (including all genbank sequences). Tags are mapped to five different categories: 5'prime, first exon, inner exon or last exon, based whether a tag overlaps that type of feature or not, on the same strand. The first category includes all tags no more than 50nt from the TSS of a known transcript. Note that the categories are not mutually exclusive and will therefore not add up to 1. If a transcript only has one exon it will be considered a first exon. While the libraries are not widely different, hippocampus tags hit 5' regions less often than other libraries, much like the embryonic library. The preference of brain tags to hit the last exon is described in (Carninci et al. 2006); the reason we do not see this in the hippocampus library is likely that the hippocampus library is random primed and not oligo-dt primed like the remaining libraries (See (Carninci et al. 2006)) for a further discussion on these types of promoters and dependency on priming technology)

Figure S2: Unique hippocampus promoters

The topmost graph shows the proportion of tag clusters that are unique to hippocampus given that you have X number of hippocampus tags in the tag cluster(TC). So, the first bar indicates that of those tag clusters that have exactly one tag from hippocampus 86% contain only that tag and nothing else (so, they are unique to hippocampus). This proportion rapidly decreases as the number of tags increases. After about 30 tags the number of promoters having exactly that number of tags dwindle so you observe larger perturbations. In particular the very large numbers only have a single TC with that number and therefore the fraction is one. The lower graph shows for comparison the actual number of unique hippocampus TCs. The y-axis is in log-scale and the first item is 83952.

Figure S3: Proportion of PEPs classified as single peaks

We defined two categories of shapes: broad (BR) and single peak (SP) as in (Carninci et al. 2006). We then classified the PEPs for each tissue into with the rule: if it is not a single peak it is by default broad. The proportion of broad is therefore 1 minus the value indicated in the barplot. We observe that some of the brain tissue PEPs (hippocampus and somotosensory cortex in particular) have less single peak PEPs than other tissue PEPs. The smaller dependence of P-value cutoffs in hippocampus, liver and lung is a sample size effect.

Figure S4: Proportion of PEPs overlapping CpG Islands
The proportion of PEPs for each tissue that map to CpG islands. Annotations for CpG islands were fetched from the UCSC browser. These were compared to the positions of the PEPs searching for overlaps. An instance was counted if they had at least one nucleotide in common. This was done for PEPs at five different significance levels (P = {"0.05", "0.01", "0.001", "0.0001", "0.00001"}). As above, brain and macrophage PEPs have a higher preference for CpG islands – this is particularly true for hippocampus PEPs

Figure S5: Proportion of PEPs with a TATA box
This figures shows the proportion of PEPs for each tissue that has a computationally predicted TATA-box. We extracted the promoters for all PEPs of each tissue at five different levels of significance (P = {"0.05", "0.01", "0.001", "0.0001", "0.00001"}). A promoter was defined as the whole PEP plus a downstream slack of 40nt and the TBP box model (MA0108) from the JASPAR database was used to scan each of these for TATA-boxes. A sequence was considered to contain a TATA-box if it had a substring scoring higher than 80% of the maximum score (assuming a uniform background model).
As noted previously (Gustincich et al. 2006), brain and macrophage PEPs have a smaller TATA-box preference.

Figure S6: Correlation examples of tag distributions in promoters used in brain
We selected all tag clusters that were not preferentially expressed in any tissue, and had at least 30 tags each from hippocampus, cerebellum and somotosensory cortex. Each such tag cluster will have one vector of expression values for each tissue, of the same length as the cluster, corresponding to the tag usage of each nucleotide within the cluster, (expressed as tags per million). We randomly selected 20 of these 491 promoters and visualized them as barplots. By eye, the tag usage per nucleotide is correlated between the tissues. Pearson correlation statistics for the whole set is shown below.

Figure S7: Correlation between tag distributions in promoters used in brain
We used the 491 promoters as defined above and calculated the Pearson correlation coefficient between the TPM usage on nucleotide level in each promoter for hippocampus and somotosensory cortex, using cerebellum as a reference. The reason for this is twofold: we want to see whether most promoters have a shared tag distribution regardless of tissue, and we want to see if there are more divergences from this distribution in the 454-based library (hippocampus) than a Sanger-based (somotosensory cortex). Note that low correlations could be due to methodological bias, but could also be due to different mechanisms in determining the TSS selection: we have previously showed that there are clear cases of this phenomenon also in Sanger-based data (Kawaji et al. 2006).
We use a hexbin plot to indicate where data points overlap (several overlaps will

give stronger colored hexagons).
The picture shows that most promoters have a reasonable correlation between tag usage in all the three tissues – a promoter with high correlation between hippocampus to cerebellum generally also have a high correlation between somotosensry complex and cerebellum. Divergences from this trend are spread about equally for both tissues. This indicates that promoters that are used in all these brain tissues generally use similar mechanisms for TSS selection, and that the 454 technology behave similarly as the Sanger technology in terms of capturing this.

Figure S8: CAGE identifies promoter activity from small sub populations of hippocampal cells (expanded versions of Figure 6 also showing CAGE tag locations on the corresponding genes)
Examples of correspondence between CAGE tags and signal detected by *in situ* hybridization, ordered from relatively high expression (from the top panel), expressed as the number of CAGE tags from hippocampus hitting the transcription unit to low expression (bottom panel). On the left, the original *in situ* signal is shown, in the middle panel the *in situ* hybridization signal is quantified with pseudo-colors where red corresponds to high expression. The right panel shows a genome browser-like representation of where hippocampus CAGE tags hit the genes in questions. In these panels, the gene is shown with cyan-colored boxes corresponding to exons (only one transcript per gene is shown due to space constraints). CAGE tags are shown as two bar plots under the gene (one for each strand). Thy Y axis of the bar plot corresponds to the number of CAGE tags hitting a certain nucleotide position. Note that in almost all cases the CAGE tags hit the annotated 5' end of the gene (and sometimes other, alternative promoters.
In situ hybridization images were obtained from the Allen Brain Institute (Lein et al. 2007). Notice the signal or tags corresponding to less than 5 tags/1.3 millions mapped tags) correspond to RNAs which are expressed only in a subset of cells.

2. SUPPLEMENTARY TABLES

Table S1: CAGE libraries used. RNA library IDs refer to the CAGE database presented in (Carninci et al. 2005)

Tissue	RNA libraries used	Sequenced tags	Mapped tags	Ratio between sequenced tags and mapped tags	Tag clusters with >30 TPMs and preferential use in this tissue (PEPs)
Hippocampus	Hcamp	1943592	1388237	71.4%	1182
Somatosensory cortex	CAD, CAF, CAH, CAJ, CAB	253904	208943	82.2%	396
Visual cortex	CAA, CAC, CAE, CAG,CAI	284502	232690	81.8%	317
Cerebellum	BC	343606	253234	73.7%	1089
Liver	CBI ,CBJ ,CBK ,CBL , CBM, CBN, CBO,CBP, CBQ,CBR,CBS, CBT	1642666	1218713	74.2%	1106
Lung	CAW ,CAX, CAY, CAZ, CBA, CBB , CBC , CBD, CBE , CBF, CBG, CBH	1835646	1298243	70.1%	840
Macrophage	CBU, CDV	214195	169919	79.3%	1100
Embryo, stage TS-26	IN	995460	632197	63.5%	506

Table S2
RACE primer list
IDs in bold refer to tag cluster identifiers

To validate **C1R18949D1**,:
GCTCGGGCTCACCTTGTTCAGTACC (GSP1)
AGCTCCGGGAGGCTCTTGATCTAAG (GSPnested);

To validate **C11R235CA53**;
GATGTCGCCGGTCCTTTACATCCAT (GSP1)
CAGTAGCTTCGCTCCCTGACACCAT (GSPnested);

To validate **C1FAC028B9**,
TCGTAAGCCTTGCGGTACCAGCTCT (GSP1)
GGAAAGGTTGGAACGGCTGGTGGTT (GSPnested);

To validate **C9F67A5FEF**,
AGTGGCAGGCTACCTCTTCGTAAAG (GSP1)
TTTGCCTACCTACTCGGAGCCTCAT (GSPnested);

To validate **C1R935EAE3**,
GACACCACTTGAGTAGGCTCTCCCAC (GSP1);

To validate **C18R37A2246**
CCATCTGTGCCATGCGCCCAAAG (GSP1)
TAGGCTTGCGGTGTCCTGGTGTAGC (GSPnested).

Table S3

Motif models from JASPAR(Bryne et al. 2008) with an over-representation of predicted sites in hippocampus PEPs compared to all strong promoters (See table S4 for the same table but PEPs from all tissues)

JASPAR model ID	Transcription factor name
MA0080	SPI1
MA0098	c-ETS
MA0006	Arnt-Ahr
MA0079	SP1
MA0056	ZNF42-1-4
MA0004	Arnt
MA0104	Mycn
MA0048	NHLH1
MA0003	TFAP2A
MA0093	USF1
MA0024	E2F1
MA0028	ELK1
MA0067	Pax2
MA0102	cEBP
MA0057	ZNF42-5-13
MA0117	MafB
MA0072	RORA1
MA0088	Staf
MA0100	Myb
MA0062	GABPA
MA0076	ELK4
MA0073	RREB1
MA0119	Hox11-CTF1

Table S4

Motif models with an over-representation of predicted sites in respective PEPs compared to all strong promoters (See table S3 for the same table but only PEPs from hippocampus). "1" refers to a detected over-representation, while 0 indicates no over representation (see Methods)Note that since we compare the PEPs to al other promoters, including promoters that have no tissue preference, it is possible for a motif to be over-represented in many different PEPs. "Brain PEPs" refers to PEPs calculated from tags from pooled brain tissues libraries.

	Names	cere	embryo	hcamp	liver	lung	macro	som	vis	brain
MA0003.pfm	TFAP2A	0	0	1	0	0	1	0	0	1
MA0009.pfm	T-box	0	0	0	1	0	0	0	0	0
MA0056.pfm	ZNF42-1-4	1	0	0	0	1	1	0	1	1
MA0080.pfm	SPI1	1	0	1	1	1	1	1	1	1
MA0018.pfm	CREB1	0	0	0	0	0	0	0	1	0
MA0117.pfm	MafB	0	0	1	0	0	0	1	0	1
MA0067.pfm	Pax2	1	0	1	0	0	0	0	0	1
MA0100.pfm	Myb	0	0	0	0	0	0	0	0	1
MA0103.pfm	deltaEF1	1	0	0	1	1	1	1	0	1
MA0004.pfm	Arnt	0	0	1	1	0	1	1	1	1
MA0043.pfm	HLF	0	1	0	0	0	0	0	0	0
MA0052.pfm	MEF2A	0	0	0	0	0	0	0	0	1
MA0055.pfm	Myf	0	1	0	1	0	0	0	0	0
MA0076.pfm	ELK4	0	0	1	0	0	0	0	0	1
MA0081.pfm	SPIB	0	0	0	0	0	1	0	0	0
MA0098.pfm	c-ETS	1	0	1	1	1	1	1	1	1
MA0102.pfm	cEBP	0	0	1	0	0	0	0	0	1
MA0069.pfm	Pax6	0	1	0	0	0	0	0	0	0
MA0041.pfm	Foxd3	0	0	0	1	0	0	0	0	0
MA0047.pfm	Foxa2	0	0	0	1	0	0	0	0	0
MA0105.pfm	NFKB1	0	0	0	0	0	0	1	0	0
MA0058.pfm	MAX	1	0	0	0	0	0	0	0	0
MA0079.pfm	SP1	1	1	1	0	1	1	1	1	1
MA0104.pfm	Mycn	0	0	1	1	0	1	1	1	1
MA0093.pfm	USF1	0	0	1	0	0	1	0	1	1
MA0050.pfm	IRF1	0	0	0	0	1	0	0	0	0
MA0088.pfm	Staf	0	0	1	0	0	0	0	0	0
MA0057.pfm	ZNF42-5-13	0	0	1	0	0	1	0	0	1
MA0048.pfm	NHLH1	0	0	0	0	1	1	0	1	1
MA0090.pfm	TEAD	0	1	0	0	0	0	0	0	0
MA0024.pfm	E2F1	0	0	1	0	0	1	0	0	1
MA0006.pfm	Arnt-Ahr	1	0	1	1	1	1	1	1	1
MA0014.pfm	Pax5	0	0	0	1	0	0	1	1	0

3. SUPPLEMENTARY DATA FILES

All of these files are available at
http://people.binf.ku.dk/albin/supplementary_data/hcamp/

CAGE wig and bed tracks

The CAGE tracks can be directly used in the UCSC browser at http://genome.ucsc.edu/ ,

1) Either by first downloading the .bed and .wig files of interets and then, at the main UCSC site clicking Genomes->add custom track, and upload the files in this page.

2) Or, by using the pre-established links at http://people.binf.ku.dk/albin/supplementary_data/hcamp/ . Clicking one of the links will add that track as custom track to the ucsc browser as above, but without having to download the file.

WIG tracks (single nucleotide resolution barplots).

CAGE tags from:

* Cerebellum: plus strand
* Cerebellum: minus strand
* Embryo: plus strand
* Embryo: minus strand
* Hippocampus: plus strand
* Hippocampus: minus strand
* Liver: plus strand
* Liver: minus strand
* Lung: plus strand
* Lung: minus strand
* Macrophages: plus strand
* Macrophages: minus strand
* Som. cortex: plus strand
* Som. cortex: minus strand
* Vis. cortex: plus strand
* Vis. cortex: minus strand

BED tracks (blocks, corresponding to clusters of tags)

Summary tracks:

* All CAGE tag clusters (regardless of number of tags)
* All CAGE tag clusters (>30 Tags per million (TPM))

Preferentially expressed promoters (PEPs): Subsets of the tag cluster above that have more that have >30 TPMs and where >50% of tags come from a particular tissue (normalized for sample size)

* Cerebellum PEPs
* Embryo PEPs
* Hippocampus PEPs
* Liver PEPs
* Lung PEPs
* Macrophages PEPs
* Som. cortex PEPs
* Vis. cortex PEPs

PEP FASTA sequence files

- Sequences corresponding to the tag cluster width plus X upstream and Y downstream, used for the over-representation analysis, broken down by tissue

Over-represented motifs in each set

Lists of motifs that are statistically over-represented in each PEP fasta set defined above, using the methods as described in Methods.

Images of genes where hippocampus alternative promoter usage is predicted to give differential protein domain content.

Genome browser style images as in Figure 5 showing the 50 genes where hippocampus PEPs are predicted to results in protein products which exclude Interpro-annotated promoter domains – in other words giving promteins which might have different function from the full-length isoform.

4. SUPPLEMENTARY METHODS

Supplementary information: Full protocols on line.

CAGE library preparation

This protocol was used in this study to prepare CAGE libraries for the 454 Life Sciences sequencer.

Synthesis of First-strand cDNA

The cDNA synthesis was carried out by using 50μg of total RNA in 20 μl of water and 2 μl of random primer (N20; 6μg/μl) with M-MLV Reverse Transcriptase RNase H Minus, Point Mutant (Promega). We heated the RNA and primer to 65^{o}C for 5 min and then placed them on ice. Then we added to the reaction mixture, 75 μl of 2x GC I LA Taq buffer (TaKaRa), 4 μl of 10mM dNTPs, 20 μl of 4.9M Sorbitol, 10μl of saturated Trehalose (>75% w/v), 4 μl of water and 15 μl of Reverse Transcriptase (200U/μl), followed by reverse transcription in a thermocycler at: 30 s at 25^{o}C, 30 min at 42 ^{o}C, 10 min at 50 ^{o}C, 10 min at 56; hold at 4 ^{o}C until further processing. We stopped the reaction with 2μl of EDTA and added 3μl of proteinase K. We purified cDNA/RNA hybrids by CTAB precipitation as described and dissolved the pellet in 46 μl of water.

Oxidation/Biotinylation

3.3 μl of 1M Sodium Acetate (pH4.5) and 2 μl of 250mM $NaIO_4$ were added to cDNA/RNA hybrids and the tube was kept for 45 min on ice in the dark. After the reaction was stopped with 1μl of 80% glycerol, the cDNA/RNA hybrids were precipitated with isopropanol. The pellet was dissolved in 50 μl of water and 5 μl of 1M Sodium Citrate (pH6.1) were added with 5 μl of 10% SDS and 150 μl of 10mM biotin (long arm) hydrazide. The reaction mixture was kept for 10-12h at room temperature. Then the reaction was stopped with 75 μl of 1M Sodium Acetate (pH6.1) and then precipitated with ethanol. The pellet was dissolved in 180 μl of 0.1xTE and then treated with RNase ONE^{TM} Ribonuclease (1U/each μg starting RNA, Promega). cDNA/RNA hybrids were then purified by proteinase K digestion, phenol/chloroform extraction and isopropanol precipitation. The pellet was dissolved in 50 μl of 0.1xTE.

Full-length cDNA capture-release

500 μl of Dynabeads MP-280 Streptavidin (Dynal) were blocked by incubation with 100 μg of tRNA for 30min on ice with occasionally shaking. Beads were then washed 3 times with 500 μl of wash buffer (4.5M NaCl / 50mM EDTA) and resuspended in 500 μl of wash buffer. Beads were added to cDNA/RNA hybrids and incubated with mild agitation at 50 ^{o}C for 30min to bind the biotinylated cap to the beads. This was followed by collection of the cap/beads complex with a magnetic stand. Beads were then washed with 500 μl of a series of solutions including wash buffer (two times), 0.3M NaCl / 1mM EDTA (one time), 0.4% SDS / 0.5M NaOAc / 20mM Tris-HCl pH8.5 / 1mM EDTA (three times) and with 0.5M NaOAc / 10mM Tris-HCl pH8.5 / 1mM EDTA (two times). The selected full-length cDNAs were released 3 times with 100 μl of release buffer (50mM NaOH / 5mM EDTA). After setting the cDNAs on ice, 100 μl of 1M Tris-HCl (pH7.0) and 10 Units of RNase ONE^{TM} Ribonuclease (Promega) were added and

incubated at 37 °C for 10min. cDNAs underwent a series of purification steps including proteinase K digestion, phenol/chloroform extraction and isopropanol precititation as well as a S400 MicroSpin Column and isopropanol precipitation. The pellet was finally resuspended in 5 µl of water.

Single-strand linker ligation

A specific linker, containing recognition sites for XmaJI and MmeI ("upper oligonucleotide GN5" sequence: biotin-agagagagacctcgagtaactataacggtcctaaggtagcgacctaggtccgacGNNNNN, and "upper oligonucleotide N6" sequence: biotin-agagagagacctcgagtaactataacggtcctaaggtagcgacctaggtccgacNNNNNN) were mixed in a ratio of 4:1, and then this mixture in turn was mixed at 1:1 to the "lower oligonucleotide" sequence: Pi-gtcggacctaggtcgctaccttaggaccgttatagttactcgaggtctctctct-NH2). We added 0.2 µg of linker to the single-stranded cDNA. Using TaKaRa Ligation Kit ver.2.1, we ligated the linker to the single-stranded cDNA by incubating at 16 °C overnight. After Proteinase K treatment and phenol/chloroform extraction, linker and linker dimmers were eliminated by filtering through a S400 spun column followed by ethanol precipitation. The pellet was dissolved in 10 µl of 0.1x TE.

2nd strand synthesis

10 µl of cDNA sample were mixed with 6 µl of 100 µg/µl second-strand primer (5'-biotin-agagagagacctcgagtaactataa-3'), 7.2 µlof 5x buffer A, 4.8 µl of 5x buffer B (Elongase Enzyme Mix, Invitrogen), 6 µl of 2.5mM dNTPs and water up to 58 µl. The reaction mixture was heated to 65 °C before 2 µl of ELONGASE polymerase (Invitrogen) was added. The reaction was performed in a thermal cycler with the following settings: 5min at 65 °C, 30min at 68 °C, and 10min at 72 °C. cDNA was then filtered through a S400 spun column to eliminate extra primers, ethanol precipitated and dissolved in 10 µl of 0.1x TE.

Tagging

The double-stranded cDNA was cleaved with MmeI (3U/µg cDNA, New England Biolabs) in 100 µl, and incubated at 37 °C for 1h. After purification (proteinase K digestion, phenol/chloroform extraction, ethanol precipitation), 1.6 µl of the 2nd linker (upper oligo sequence: Pi- cctaggtcaggactcttctatagtgtcacctaaagacacacacac-NH2, lower oligo sequence: gtgtgtgtgtcttaggtgacactatagaagagtcctgacctaggNN) were added to the sample dissolved in water in a final volume of 20 µl and heated to 65 °C for 2min, then set on ice. The 2nd linker was ligated to cDNA with T4 DNA ligase (NEB) at 16 °C overnight. After stopping the reaction by heating at 65 °C for 5 min, 80 µl of 0.1xTE buffer were added.

200 µl of Dynabeads M-280 Streptavidine beads were blocked by incubation with 200 µg of tRNA for 30 min on ice with occasionally shaking. Then the beads were washed 3 times with 200 µl of 1xB+W buffer (1M NaCl / 0.5mM EDTA / 5mM Tris-HCl pH7.5) and resuspended in 100 µl of 2xB+W buffer. The washed beads were added to CAGE tags and incubated with mild agitation at room temperature for 15 min to bind the biotin to the beads. This was followed by the collection of the CAGE tags/beads complex with a

magnetic stand. Beads were washed twice with 200 μl of 1xB+W containing a BSA buffer, twice with 200 μl of 1xB+W buffer and finally twice with 200 μl of 0.1xTE buffer. The 5'-end cDNA tags were released from the beads by treatment with free biotin in excess. The biotin was solved at 1.5% (wt/vol) in 4 M guanidine thiocyanate/25mM sodium citrate pH7.0/0.5% sodium N-laurylsarcosinate. CAGE tags were then released from the beads by incubation at 45 ^{o}C for 30 min under occasional agitation. This elution was repeated three times, and fractions were pooled. After the addition of 3.5 μg of glycogen the sample was ethanol precipitated and resuspended in a 50 μl of 0.1xTE buffer. Then CAGE tags were treated with RNase ONE^{TM} Ribonuclease (Promega) and further purified on a G50 spin column followed by ethanol precipitation. The remaining DNA was dissolved in 24 μl of water.

Amplification of CAGE tags

DNA fragments were amplified in a PCR step by using the following two linker-specific primers: Primer1: 5'-biotin-CTATAGAAGAGTCCTGACCTAGG-3'; Primer2: 5'-biotin- CGGTCCTAAGGTAGCGACCTAG-3'. Ten parallel PCRs were performed in a total volume of 50 μl each by using 1.6 μl of cDNA-tags /5μl of 10xPCR buffer /3μl of DMSO /12μl of 2.5mM dNTPs /0.5 μl of Primer1 (350μg/μl) /0.5 μl of Primer2 (350μg/μl) / 26.6 μl of water / 0.8 μl of Accuzyme DNA polymerase (2.5 U/ μl, BIOLINE). After incubating at 94 ^{o}C for 1min, 20cycles were performed for 30sec at 94 ^{o}C, 20sec at 55 ^{o}C, 20 sec at 70 ^{o}C, followed by 5min at 72 ^{o}C. PCR products were then pooled, purified by ProteinaseK digestion and phenol/chloroform extraction, isopropanol precipitated, and finally resuspended in 24μl of 0.1xTE buffer.

PCR products were further purified on a 12% polyacrylamide gel. The appropriate 75bp band was cut out of the gel, crushed, and incubated with 150 μl of elution buffer (2.5mM Tris-HCl pH7.5 / 1.25M ammonium acetate / 0.17mM EDTA pH7.5) overnight at room temperature. The purified tags were filtrated through MicroSpin Columns. 150 μl of elution buffer were then added to the gel and the tube rotated at room temperature for 30min. This step was repeated additional three times. The tags were then precipitated with ethanol and dissolved in 30 μl of 0.1xTE. The concentration was measured with Picogreen.

Purified DNA from the previous PCR was PCR-amplified once more in a total of 100 μl by using 0.2-6ng of cDNA-tags/10 μl of 10x PCR buffer/6 μl of DMSO/12 μl of 2.5mM dNTPs/0.75 μl of Primer1 (1 μg/μl)/ 0.75 μl of Primer2 (1 μg/μl)/ 0.8 μl of Accuzyme DNA Polymerase (2.5 U/ μl) and water up to 100 μl. 30-100 tubes were heated to 94 ^{o}C for 1min, then 8 cycles were performed for 30 sec for 94 ^{o}C , 20 sec for 55 ^{o}C, 20 sec for 70 ^{o}C , followed by a final elongation at 72 ^{o}C for 5 min. The PCR products were pooled, purified, ethanol precipitated and dissolved in 50 μl of 0.1xTE. To eliminate excess primers, PCR products were further purified with MinElute columns (QIAGEN), ethanol precipitated and finally dissolved in 100 μl of 0.1x TE. The concentration was measured with Picogreen.

The purified PCR products were digested with XmaJI (50 Units/μg in a series of tubes, using 2 /μg of DNA/tube. A proteinase K treatment was then carried out.

The 37 bp DNA tags were separated from the free DNA ends during restriction by incubation with streptavidin-coated magnetic beads, which retain the biotin-labeled DNA ends. The cleaved tags were mixed with 500 μl of beads and incubated at room

temperature for 15min with mild agitation. The supernatant was then collected after removal of the magnetic beads. Beads were rinsed with 50 µl of 1x B+W buffer. Pooled 37-nt tags from both supernatants were extracted with phenol/chloroform followed by ethanol precipitation and dissolved in 45 µl of TE.
Tags were further purified on a 12% polyacrylamide gel. The appropriate 37bp band was cut out of the gel, crushed, and eluted with the previously used elution buffer overnight at room temperature, followed by ethanol precipitation. The tags were dissolved in 6 µl of 0.1xTE. The concentration was measured with Picogreen.
500ng of CAGE tags were ligated to form concatemers. 6 µl of tags were incubated overnight at 16°C with 1/20 amount of tags of 454 adaptors A/B as described in the original publication . 1.0 µl of 10x T4 DNA ligase buffer, 1.0 µl of T4 DNA ligase and water up to 10 µl. A ProteinaseK treatment followed. The sample was then purified on a GFX column to eliminate short concatemers.

The eluted sample was sequenced with 454 GS20 Life Science sequencer.

CAGE mapping procedure
Step1: Vmatch
CAGE tags were mapped to chromosomes 1-22, X, Y and the mitochondrial genome (Genome build: Hg18) using the BLAST/Vmatch alignment program, and the longest full-matched (meaning no mismatches in the middle) positions were selected. These tags were referred to as ‘single-mapped’ tags. Tags that map to multiple locations on genome (with same length) were called 'multi-mapped' and tags that did not map (mapped less than 18bp long) were called 'unmapped'. These 'multi-mapped' and 'unmapped' tags were passed to the rescue stage to increase the number of 'single-mapped' tags. Rescued tags were incorporated into the single-mapped tag collection, and other tags were discarded.
Steps 2 and 3: Rescue of unmapped and multi-mapped CAGE tags.

Failure to map a particular CAGE tag could arise because of: 1) polymorphisms in the genome from which the tags derive compared to the reference genome; or 2) heterogeneity in the CAGE tag sequence itself owing to post-transcriptional modifications or experimental artifacts such as annealing errors or technical sequencing errors. This process focused on rescuing tags that cannot be mapped as a result of known experimental artifacts. Multi-mapped tags were also submitted to the rescue process when an artifact correction may result in specific mapping.
CAGE tag rescue were performed in two sequential steps: 1) mono-base replacement and 2) homopolymer adjustment. The general rescue process pipeline is shown in the Figure 3. These steps corresponded to overall CAGE mapping steps 2 and 3 respectively, and will be referred to as such. Tags with no unique Vmatch mapping were passed first to step 2, and then to step 3 if they were not ‘rescued’ in step 2. Tags that were not uniquely mapped in any of the three mapping steps remained classified as unmapped and multi-mapped, as described above. All steps use Hg.18 and inspect chromosome 1-22, X,Y & M including their repetitive region denoted in small capital for masking, but “random” and “haplotype” sequences were excluded.

Step 2: Mono-base replacement rescue.

Artifacts in the CAGE sequences may arise during anealing.

There are two typical types of artifacts in CAGE sequences. One is the well known tendency of capping G(guanine). Theoretically, the capping G may occur a single time at the head of the sequence, but experimentally, there seem to be multiple artifact span of G's at the head of the sequence. The number of artifact G's is unpredictable and it leaves the problem of how many heading G's have origins in the genome.

The possibility of a second type of artifact was revealed as the result of the preliminary study shown in Figure 4. This figure shows the proportion of non-single-mapped human CAGE tags from GNP deep CAGE libraries, that could be rescued by allowing a mono-base replacement. Tags with single nucleotide mutations in the first 6 or final 2 nucleotides were more frequently rescued, indicating that single nucleotide substitutions occurred more frequently at these positions. This region just matched to the annealing site in the CAGE extraction protocol, and the curve supported mono base replacement probability. This assumption naturally suggested a rescue process that permits a mono base replacement in these heading 6 and tailing 2 regions within a CAGE tag sequence.

Mono-base replacement rescue definition.

A CAGE tag was defined to match onto the genome by mono-base replacement relaxation, if and only if the tag had only one base difference in its heading 6 or tailing 2 regions on the alignment to the genome. Then, if a tag had a single (chromosome, strand, and locus) match in the whole genome, it was classified into the specific tag category that had a single hit. If a tag had multiple matches in the whole genome, regardless of their mismatch position, it is classified into the non-specific multi-map tag category. For a neither single- nor multi- hit tag, if the head nucleotide of the tag was G(guanine), the same definition was applied recursively without its heading G. Otherwise, the tag was classified into the category of "unmapped tag with no hits".

Step 3: Homopolymer rescue.

Artifacts in CAGE sequences due to pyrosequencing.

Pyrosequencing is an advanced technology for high throughput DNA sequencing. However, in contrast to conventional Sanger sequencing, it detects homopolymeric (continuous single) nucleotides at a single cycle by fluorescence intensity, and the accuracy of calling each base in a homopolymer decreases as the homopolymer increases in size as the dynamic linear fluorescent range is limited. CAGE tags that were sequenced by pyrosequencing were under the influence of homopolymer errors. The third CAGE mapping step was intended to correct for mismatching homopolymer counts between genome and tag sequences.

Homopolymer rescue definition.

Homopolymer rescue utilized the penalty score value P between tag and genome sequences that had an exact common order of nucleotides irrespective of each number of occurrences. The penalty score P was defined as:

$P \equiv (\text{head\_penalty}) + \Sigma i=2,n-1(\text{internal\_penaltyi}) + (\text{tail\_penalty})$

$\text{head\_penalty} \equiv \max(0, \log(\text{tag\_polymer\_degree1} / \text{genome\_polymer\_degree1}))$

$\text{internal\_penaltyi} \equiv \text{abs}(\log(\text{tag\_polymer\_degreei} / \text{genome\_polymer\_degreei}))$

$\text{tail\_penalty} \equiv \max(0, \log(\text{tag\_polymer\_degreen} / \text{genome\_polymer\_degreen}))$

The "suffix i" denoted the i-th nucleotide, where the "polymer degreei" was for the polymerization degree number at i-th nucleotide in the order. Absolute value of

logarithmic ratio of tag and genome degrees denoted the elementary dissimilarity of the two sequences. The score P essentially summed up these elements. In case that the genome displayed a longer homopolymer at either end than the tag being compared, head and tail definitions did not increase the score if the tag homopolymer degree was smaller.
If a tag corresponded to a single specific region on the genome that had a minimum value of P under maximum threshold (tentative value used is log2), then this tag was classified into the specific tag category that had a single hit. The trailing processes were the same as mono-base replacement rescue. If a tag had multiple matches in the whole genome, with an equal minimum value of P under the maximum threshold, it was classified into the non-specific multi-map tag category. For a neither single- nor multi- hit tag, if the head nucleotide of the tag was G(guanine), the same definition was applied recursively without its heading G. Otherwise, the tag was classified into the category of unmapped tag with no hits. The detail on the high throughput implementation of this method was beyond the scope of this paper (in preparation: High throughput DNA sequence mapping system that focuses and evaluates homopolymer adjustment).

CAGE tag rescue summary

Single-mapped results from full-match (first step) and rescues (second and third steps) were combined with priority of (1) full-match, (2) SNP rescue, and (3) homopolymer rescue. Order was decided by the quality of mappings (by looking at percentage of tags mapped within transcription units). Multimap and unmapped results from first step were pooled with tags that did not have any single-mapped results in all three steps.
In the rest of the analysis we use only the single-mapped tags; note that the same mapping procedure was applied to all CAGE libraries in the study.

REFERENCES

Bryne, J.C., E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36:** D102-106.

Carninci, P. T. Kasukawa S. Katayama J. Gough M.C. Frith N. Maeda R. Oyama T. Ravasi B. Lenhard C. Wells et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309:** 1559-1563.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C.A. Semple, M.S. Taylor, P.G. Engstrom, M.C. Frith et al. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*.

Gustincich, S., A. Sandelin, C. Plessy, S. Katayama, R. Simone, D. Lazarevic, Y. Hayashizaki, and P. Carninci. 2006. The complexity of the mammalian transcriptome. *J Physiol* **575:** 321-332.

Kawaji, H., M.C. Frith, S. Katayama, A. Sandelin, C. Kai, J. Kawai, P. Carninci, and Y. Hayashizaki. 2006. Dynamic usage of transcription start sites within core promoters. *Genome Biol* **7:** R118.

Lein, E.S. M.J. Hawrylycz N. Ao M. Ayres A. Bensinger A. Bernard A.F. Boe M.S. Boguski K.S. Brockway E.J. Byrnes et al. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445:** 168-176.