計画研究：2000～2004年度

ゲノム／cDNA情報を基盤とした疾患遺伝子解明法の開発

●菅野　純夫[1)]　◆大海　忍[2)]　◆加藤　宏幸[3)]

1）　東京大学大学院新領域創成科学研究科　2）　東京大学医科学研究所　3）　国立感染症研究所

〈研究の目的と進め方〉

2000年にはヒトゲノムのドラフト配列決定が行われ、染色体21番22番の全塩基配列も決定された。このようなヒトゲノム配列決定の進行を背景に、本計画研究では、ゲノム配列情報と遺伝子の転写開始点の情報をあわせ、プロモーター領域を体系的に明らかにし、プロモーター領域に変異を持つ疾患関連遺伝子を体系的に探査する上での基盤をあたえる事を目標にする。決定されたプロモーター領域にSNP情報を張り付け、疾患関連遺伝子探査グループへ公開する。

さらに、分離した多数の完全長cDNAクローンをタグ付きタンパク質の形で発現できるようにし、cDNAがコードするタンパク質と相互作用する一群のタンパク質を、迅速容易に同定可能なシステムの開発を行い、タンパク質相互作用解析を容易ならしめる。

〈研究開始時の研究計画〉

1）ヒトプロモーター領域の同定

われわれは、オリゴキャップ法を用いた完全長cDNAライブラリー作成法を開発し、ヘリックス研究所はじめ様々な機関と協力しながら、大量のヒト完全長cDNAクローンを収集し、その5‘端部分配列や全長配列を決定してきた。これらの活動を続けると共に、得られた配列データをゲノム配列上にマップすることで、タンパク質をコードする遺伝子を中心に、転写開始点を明らかにし、プロモーター領域を含むと考えられる転写開始点上流1000塩基長のゲノム配列を得、ホームページでのデータベースとして公開を行う。

2）プロモーターの解析とマウスとの比較

大量の5‘端データを利用して、遺伝子の詳細なプロモーター構造を、特に選択プロモーターを中心に解析する。また、同定したプロモーター領域が実際に活性を示すかどうか、ヒトゲノムよりプロモーター領域をクローン化し、培養細胞の系を用いてその活性を測定する。

さらに、他グループより発表されたマウスの完全長cDNAの5‘端ESTデータおよび、われわれが決定したマウスの完全長cDNAの5‘端データを用い、マウスの遺伝子のプロモーター部分を特定すると同時に、対応するヒト遺伝子のプロモーター領域との比較をおこなう。

3）タンパク質の細胞内局在解析と質量分析システムの構築

分離した多数の完全長cDNAクローンをタグ付きタンパク質の形で発現できるように、GATEWAYのエントリーベクターに、cDNAをクローン化する。さらに、GATEWAYシステムを用い、タグ付きのタンパク質の発現を検証するために、完全長cDNA由来のタンパク質の細胞内局在を検討する。

また、質量分析計を用いた高感度で迅速なタンパク質同定系を立ち上げ、タンパク質相互作用解析のためのシステムを構築する。

〈研究期間の成果〉

1）ヒトプロモーター領域の同定

2001年度にはタンパク質をコードしている遺伝子について約1000遺伝子、2002年度には約5000遺伝子、2003年度には約9000遺伝子について、プロモータ領域の同定をおこなった。最終的には、164種類の完全長cDNAライブラリーから、約178万の5'ESTデータを取得し、日立製作所、リバースプロテオミックス研究所、東京大学医科学研究所の中井謙太教授らと共に、約14,000種のタンパク質をコードしている遺伝子について、転写開始点を同定し、プロモーター領域を含むと考えられる転写開始点上流1000塩基長のゲノム配列を得ることが出来た。また、SNP情報をdbSNPを利用して、すべてのプロモーターに張り付けることが出来た。この結果をデータベースDBTSSとして、ホームページ（http://dbtss.hgc.jp/）で公開を行った（図1，2）。

2）プロモーターの解析とマウスとの比較

a）選択プロモーター領域の同定

転写は必ずしも一つの塩基から始まるわけではなく、あるプロモーター領域中の複数の塩基から始まることが一般的である。すなわち、転写開始点にはマイクロヘテロジェニュイティが存在する。逆に、転写開始点をゲノム配列上にマップすると、転写開始点はクラスターを形成することなる（図3）。これらのクラスターは､約100塩基の長さの範囲に収まっていることが多く、200塩基を超えることは希である。われわれは、転写開始点のクラスターが500塩基以上のながさの転写開始点を含まない領域で隔てられている場合、それを、異なるプロモーター領域と定義することとした。

この定義に従い、1）のデータから、ひとつの遺伝子について、複数のプロモーター領域（選択プロモーター）を持つものの割合を検討した。その結果、驚くべき事に、最終的に解析した14,000遺伝子のうち、約半数（52%）が、選択プロモーター領域を持つことを見出した（表）。25%の遺伝子には3つ以上のプロモーター領域が存在している。転写開始点が、イントロンの中にある例も予想以上に多く、転写開始点が異なると、選択スプライスのパターンも異なってくるので、遺伝子によっては、非常に複雑な遺伝子構造を持つ場合があることを示している（図4）。

b）プロモーター部位の活性測定

われわれがマップしたプロモーター領域が、現実に、活性を持つか検討する目的で、プロモーター領域を同定した遺伝子のうちヒト293細胞で転写されているもの473個を選び、そのプロモーター領域をクローン化し、その転写活性を293細胞を用いてルシフェラーゼアッセイで測定した。プロモーター領域としては、その領域で最も頻度の多い転写開始点から上流約1000塩基、下流約200塩基、合計で約1200塩基長の部分を選んだ。対照として、251の非プロモーター・非遺伝子領域のゲノム配列をランダムにえらび、そこから約1200塩基長の断片をクローン化して、その活性も測定した。

その結果、われわれがマップしたプロモーター領域と、対照としたランダム非プロモーター・非遺伝子領域配列の活性は明白に分かれ、われわれのプロモーター領域の推定が基本的に正しいことを示した。

されに詳しく見ると、a) われわれがマップしたプロモーター領域の活性は大まかに強いものと弱いものの2種類に分かれること、b) ランダム非プロモーター・非遺伝子領域配列の約1割が弱いプロモーター活性を示すこと、c)現在注目されているnon-codingRNAのプロモーター領域が、弱いプロモーター活性を示すことを見出した（図5)。

ｃ）マウスプロモーターの同定と比較

林崎らにより発表されたマウスの完全長cDNAのデータ、およびわれわれが決定したマウス完全長cDNAの５'端ESTデータを用い、2004年度には、マウスの既知遺伝子約6000個について、そのプロモーター領域を同定し、最終的には、約13,000種の遺伝子についてプロモーター領域を同定した。ヒトのプロモーター領域と比較したところ約40％が、一致していた。一致していたものについては、プロモーター領域全体の平均ホモロジーは40－50％である。

しかしながら、詳細に見ると、ホモロジーが60－70％の部分が転写開始点から上流へある程度続き、突然、ホモロジーが30－40％へ落ちるというブロック構造を取っていることがわかった（図６）。このブロックの長さが遺伝子ごとにさまざまであるため、平均すると40－50％のホモロジーとなる。面白いことに、約3割のケースで、ホモロジーが60－70％のブロックの境界部分にAlu配列などの反復配列を見いだした。したがって、組み換えが、ホモロジーブロックの生成に、何らかの役割を果たしている可能性がある。転写因子結合部位で高ホモロジー・ブロックに存在するものは、マウスとヒトで共通のものが、当然ながら多かった。

選択プロモーターも含めて、マウスとヒトを比較すると、ゲノム配列上は両者間でホモロジーが存在するものの、一方で、プロモーター領域が見いだせない例や、一方のプロモーター領域に対応する他方のゲノム領域が存在しない例などが、多数見いだされた。プロモーターが異なることによって、mRNAの構造が変わり、コードされるタンパク質が変化する場合がしばしばある。したがって、選択プロモーターのなかで、マウスとヒトで保存されないものが多数あるということは、タンパク質でヒト特異的というべきものが、存在する可能性を示し、種差について面白い示唆を与える。

３）タンパク質の細胞内局在解析と質量分析システムの構築

ａ）タンパク質細胞内局在の網羅的解析

分離した完全長cDNAの機能解析をめざして、多目的発現系であるGATEWAYシステムにcDNAをクローン化し、今後のタンパク質の網羅的機能解析に備えなることにした。

2000年度は、200種類、2001年度は1000種類、最終的には、約3000種類のcDNAをGATEWAYのエントリーベクター上に乗せると同時に、GFPをタグにしたGFP融合タンパク質の発現することの出来る発現クローンを作製し、これらのクローンを使用して、タンパク質の細胞内局在を検討した（図７）。

約3000種のcDNA由来のタンパク質を検討することで得られたミトコンドリアに局在するもの39クローンについて、さらに詳しい検討を加えた。このうち、既知のもの12クローン、既知のタンパク質であるが、他の細胞内局在を示すとされているもの５クローン、新規のもの22クローンであった。既知のタンパク質であるが、他の細胞内局在を示すとされているもの５クローンについては、最近の解析で、ミトコンドリアへの分布が言われているものが2クローンであり、されに、残りの３クローンはスプライスバリアントである。スプライスによって生じる一部の配列を欠失したタンパク質が、本来のものと異なる分布を示す例は、良く知られているが、これらもその例と考えられる。

この結果と、Psort、Mitoprot、TargetPなどの既存のタンパク質の細胞内局在予想プログラムで検討したところ、SOSUIにより膜貫通部位の存在が洋装されるタンパク質については、ミトコンドリアの局在を良く予想できているものの、膜貫通部位があると予想されるタンパク質についてはミトコンドリアへの局在を充分に予想できていなかった。このことから、ミトコンドリアの膜タンパク質について新しいミトコンドリア局在モチーフの存在が考えられる。

ｂ）細胞内の小タンパク質の網羅的解析とUTR-ORF発現の確認

都立大学磯部俊明教授、産総研夏目チームリーダーとの共同研究により、ナノＬＣと高感度質量分析計システムを利用し、タンパク質の高感度で迅速な検出同定系を作ることができた。この系を用いて、Ｋ562細胞内の100アミノ酸以下の長さの小タンパク質の検出を行った。

この結果、既知の小タンパク質は52種類同定できた。主なものは、リボゾームタンパク質や膜輸送、転写、スプライス、細胞周期、エネルギー産生に関係する小タンパク質である。既知の小タンパク質の総数は約600種であるので、約一割のものをＫ562細胞で検出できたことになる。

新規のものは合計７種見いだされたが、そのうち4種はわれわれのクローンしたcDNAから、その存在が予想されていたものである。他の３つは思いがけずに既知のmRNAの５'UTRに存在するORFでコードされるタンパク質をであった（図８）。選択スプライスにより特別なmRNAができて、このタンパク質が作られている可能性は捨てきれないものの、このような小タンパク質が、ヒトを含む多細胞真核生物で具体的に検出されたのは、本研究がはじめてである。

５'UTRに存在するORFでコードされるタンパク質がみいだされたｍRNAの最長ORFがタンパク質を作っているのかどうかを、NM_015532について抗体で確認したところ、予想通りの51.7kDaのバンドを見いだした（図９）。これにより見出された小タンパク質が、UTR-ORFである可能性が高まった。

〈国内外での成果の位置づけ〉

本研究で使用している５'端のcDNA配列は、われわれの開発したオリゴキャップ法により、mRNAの転写開始点を高い効率で反映していると考えられる。このような情報を持つものは、国際的にもわれわれだけであり、本研究は国際的にユニークな位置を占めている。

ナノＬＣを用いた高感度質量分析システムも我が国独自のものであり、国際競争力を持つと考えられる。

〈達成できなかったこと、予想外の困難、その理由〉

特になし。

〈今後の課題〉

疾患関連の遺伝子を中心に、プロモータ活性や、相互作用するタンパク質を明らかにしていく必要がある。

〈研究期間の全成果公表リスト〉

1）論文

1．受付番号：0602071207
Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A., Sugano, S. Statistical analysis of 5' untranslated region of human mRNA using "Oligo-capping" cDNA libraries. Genomis 64: 286-297, 2000.

2．受付番号：0602071214
Ohmori, Y., Tanigami, A., Sugano., S Comparative PCR: A Simple and Sensitive Method to Quantify Low-Abundant mRNA Species. Genomics 67: 140-145, 2000.

3．受付番号：0602071218
Yudate, H. T., Suwa, M., Irie, R., Matsui, H., Nishikawa, T., Nakamura, Y., Yamaguchi, D., Peng, Z. Z., Yamamoto, T., Nagai, K., Hayashi, K., Otsuki, T., Sugiyama, T., Ota, T., Suzuki, Y., Sugano, S., Isogai, T., Masuho, Y. HUNT: launch of a full-length cDNA database from the helix research institute. Nucl. Acid Res. 29: 185-188, 2001.

4．受付番号：202281642
Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 11: 677-684, 2001.

5．受付番号：202281647
Suzuki Y, Taira H, Tsunoda T, Mizushima-Sugano J, Sese J, Hata H, Ota T, Isogai T, Tanaka T, Morishita S, Okubo K, Sakaki Y, Nakamura Y, Suyama A, Sugano S. Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep. 2: 388-393, 2001.

6．受付番号：202281651
Suzuki Y, Sugano S. Construction of full-length-enriched cDNA libraries. The oligo-capping method. Methods Mol Biol. 175: 143-153, 2001.

7．受付番号：303221955
Watanabe J, Sasaki M, Suzuki Y, Sugano S. Analysis of transcriptomes of human malaria parasite Plasmodium falciparum using full-length enriched library: identification of novel genes and diverse transcription start sites of messenger RNAs. Gene. 291:105-113, 2002.

8．受付番号：303221951
Omori Y, Imai J, Suzuki Y, Watanabe S, Tanigami A, Sugano S. OASIS is a transcriptional activator of CREB/ATF family with a transmembrane domain. Biochem Biophys Res Commun. 293: 470-477, 2002.

9．受付番号：303221946
Osada N, Kusuda J, Hirata M, Tanuma R, Hida M, Sugano S, Hirai M, Hashimoto K. Search for genes positively selected during primate evolution by 5'-end-sequence screening of cynomolgus monkey cDNAs. Genomics. 79: 657-662, 2002..

10．受付番号：303221941
Suzuki Y, Yamashita R, Nakai K, Sugano S. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. 30: 328-331, 2002.

11．受付番号：303221937
Ueda R. H, Chen W, Adachi A, Wakamatsu H, Hayashi S, Takasugi T, Nagano M, Nakahama K, Suzuki Y, Sugano S, Iino M, Shigeyoshi Y, Hashimoto S. A Transcription Factor Response Element for Gene Expression During Circadian Night. Nature 418: 534-539, 2002.

12．受付番号：303222000
Komatsu T, Suzuki Y, Imai J, Sugano S, Hida M, Tanigami A, Muroi S, Yamada Y, Hanaoka K. Molecular cloning, mRNA expression and chromosomal localization of mouse angiotensin-converting enzyme-related carboxypeptidase (mACE2). DNA Seq. 13:217-220. 2002.

13．受付番号：0403261324
Suzuki Y, Sugano S. Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. Methods Mol Biol. 221: 73-91, 2003.

14．受付番号：0403261329
Sakate R, Osada N, Hida M, Sugano S, Hayasaka I, Shimohira N, Yanagi S, Suto Y, Hashimoto K, Hirai M. Analysis of 5'-end sequences of chimpanzee cDNAs. Genome Res. 13: 1022-1026, 2003.

15．受付番号：0403261335
Matsuda A, Suzuki Y, Honda G, Muramatsu S, Matsuzaki O, Nagano Y, Doi T, Shimotohno K, Harada T, Nishida E, Hayashi H, Sugano S. Large-scale identification and characterization of human genes that activate NF-kappaB and MAPK signaling pathways. Oncogene. 22: 3307-3318, 2003.

16．受付番号：0403261341
Shibui-Nihei A, Ohmori Y, Yoshida K, Imai J, Oosuga I, Iidaka M, Suzuki Y, Mizushima-Sugano J, Yoshitomo-Nakagawa K, Sugano S. The 5' terminal oligopyrimidine tract of human elongation factor 1A-1 gene functions as a transcriptional initiator and produces a variable number of Us at the transcriptional level. Gene. 311: 137-45, 2003.

17．受付番号：0403261344
Kikuchi S. et al. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. Science. 301: 376-379, 2003.

18．受付番号：0403261350
Ohira M, Morohashi A, Inuzuka H, Shishikura T, Kawamoto T, Kageyama H, Nakamura Y, Isogai E, Takayasu H, Sakiyama S, Suzuki Y, Sugano S, Goto T, Sato S, Nakagawara Expression profiling and characterization of 4200 genes cloned from primary neuroblastomas: identification of 305 genes differentially expressed between favorable and unfavorable subsets. Oncogene. 22:5525-5536. 2003

19．受付番号：0403261407
Yamashita R, Suzuki Y, Nakai K, Sugano S. Small open reading frames in 5' untranslated regions of mRnas. C

R Biol. 326:987-991, 2003.
20. 受付番号：0403261411
Ota T, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 36:40-45, 2004.
21. 受付番号：0403261422
Watanabe, J., Sasaki, M., Suzuki, Y. and Sugano, S. Full-malaria 2004: an enlarged database for comparative studies of full-length cDNAs of malaria parasites, Plasmodium species. Nucleic Acids Res. 32: D334-338, 2004.
22. 受付番号：0403261415
Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res. 32: D78-81, 2004.
23. 受付番号：0602071226
Hashimoto SI, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K. 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol. 22: 1146-1149, 2004.
24. 受付番号：0602071231
Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isobe T, Sugano S. Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res. 10B: 2048-2052, 2004.
25. 受付番号：0602071235
Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res. 14:1711-1718, 2004.
26. 受付番号：0602071242
Baajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. Nt Biotechnol. 22: 1467-1473, 2004

Baajic VB, Tan SL, Suzuki Y, Sugano S. Promoter prediction analysis on the whole human genome. Nt Biotechnol. 22: 1467-1473, 2004

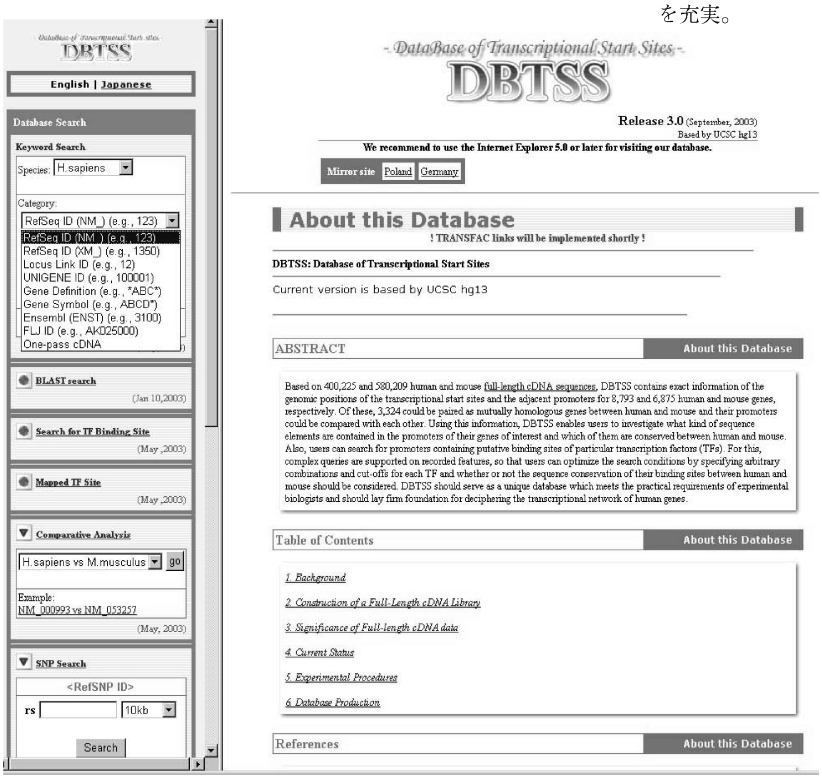
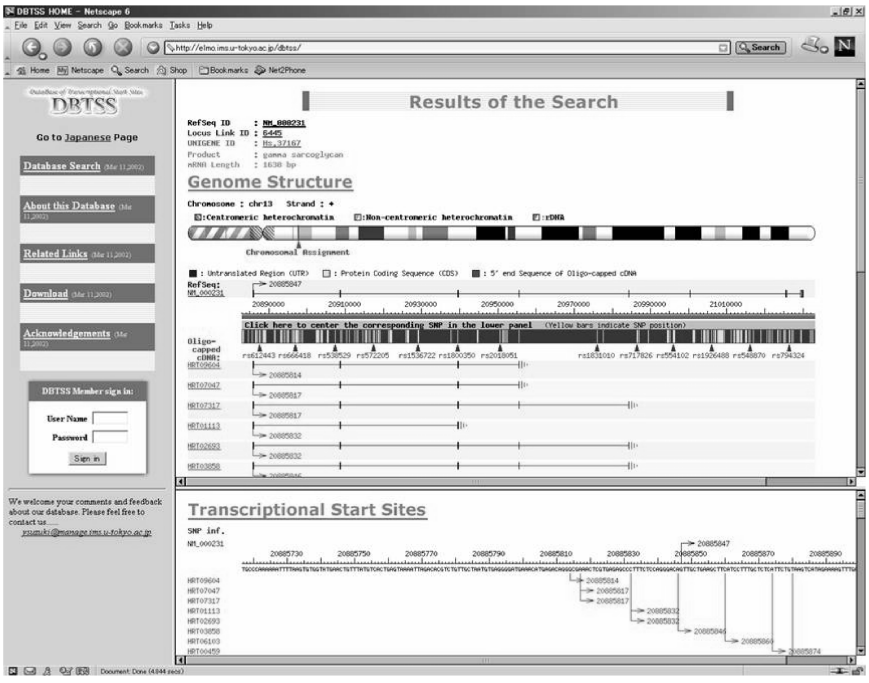


図1 データベース DBTSS のフロントページ



2）データベース／ソフトウェア

DB名：DBTSS URL：http://dbtss.hgc.jp/

100万を越すヒト5'ESTデータを使用した、ヒト遺伝子の転写開始点の情報を中心に、遺伝子の転写開始点をデータベースとしたもの。現在は、ヒトに加えマウスのデータが多い。今後他の生物を充実。

図2 DBTSS の遺伝子情報

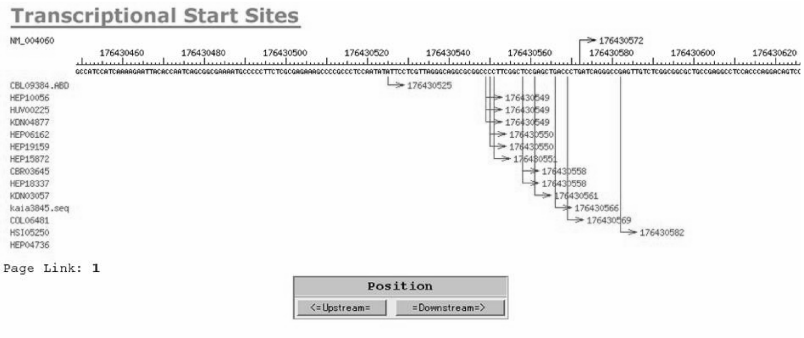


図３ゲノム配列にマップされた個々の転写開始点（赤い矢印）

Table 1. Distribution of the putative alternative promoters

No. PAPs	No. locus	No. included TSS positions	No. cDNA clones (avg.)
1 (PAP-less)	6954 (48%)	70,175	43
2	3724 (26%)	67,846	83
3	1821 (12%)	44,455	115
4	1003 (7%)	32,582	160
5	490 (3%)	19,962	166
6	294 (2%)	13,937	159
7	147 (1%)	7948	184
8	85 (0.6%)	4912	194
9	42 (0.3%)	2167	163
10	25 (0.2%)	1650	164
>10	43 (0.3%)	4140	341
total	14,628	269,774	80

表　遺伝子あたりのプロモーター数　ＰＡＰ:推定選択プロモーター

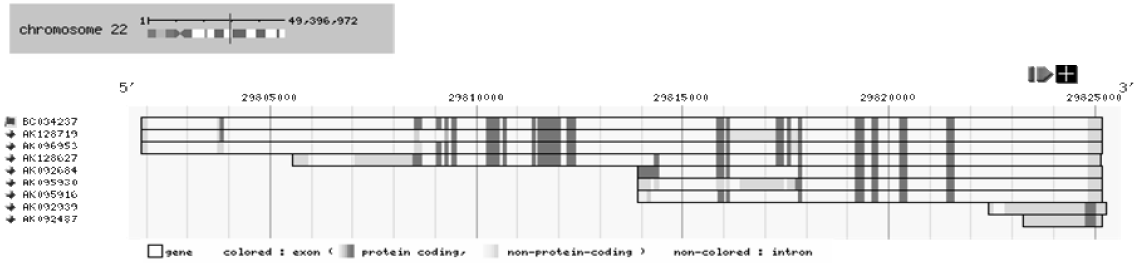


図４　５つの選択プロモーターと複雑なスプライスパターンを示す遺伝子例。

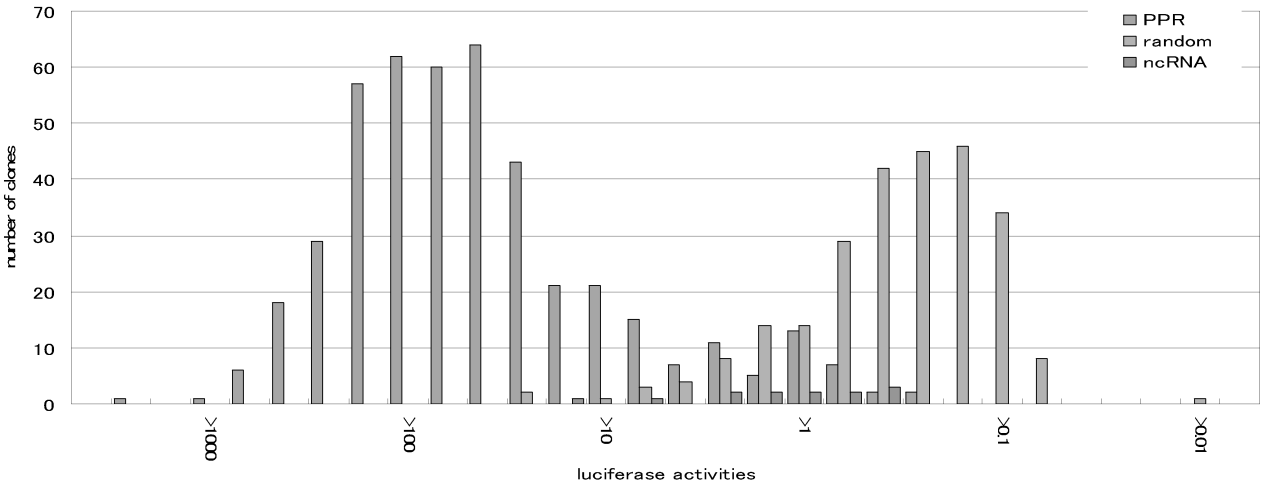
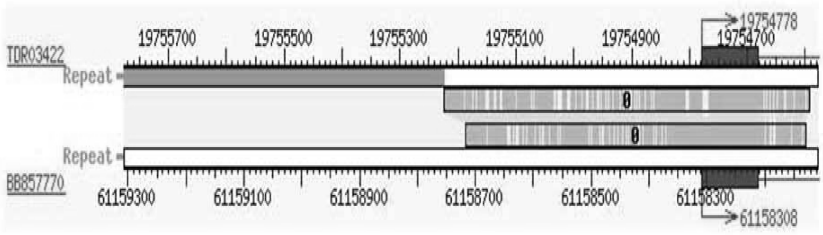


図５　２９３細胞におけるプロモーター活性。PPRはわれわマップしたプロモーター領域、randomはランダム非プロモーター・非遺伝子領域配列、ncRNAはnon-codingRNAのプロモーター領域を示す。



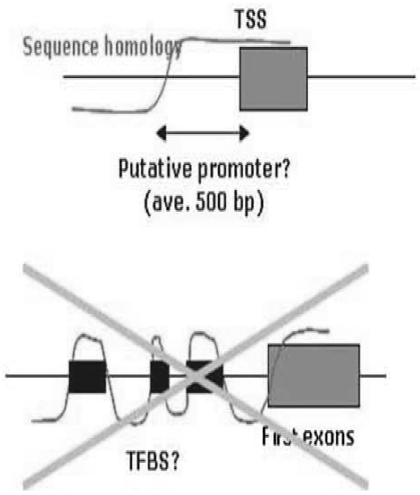
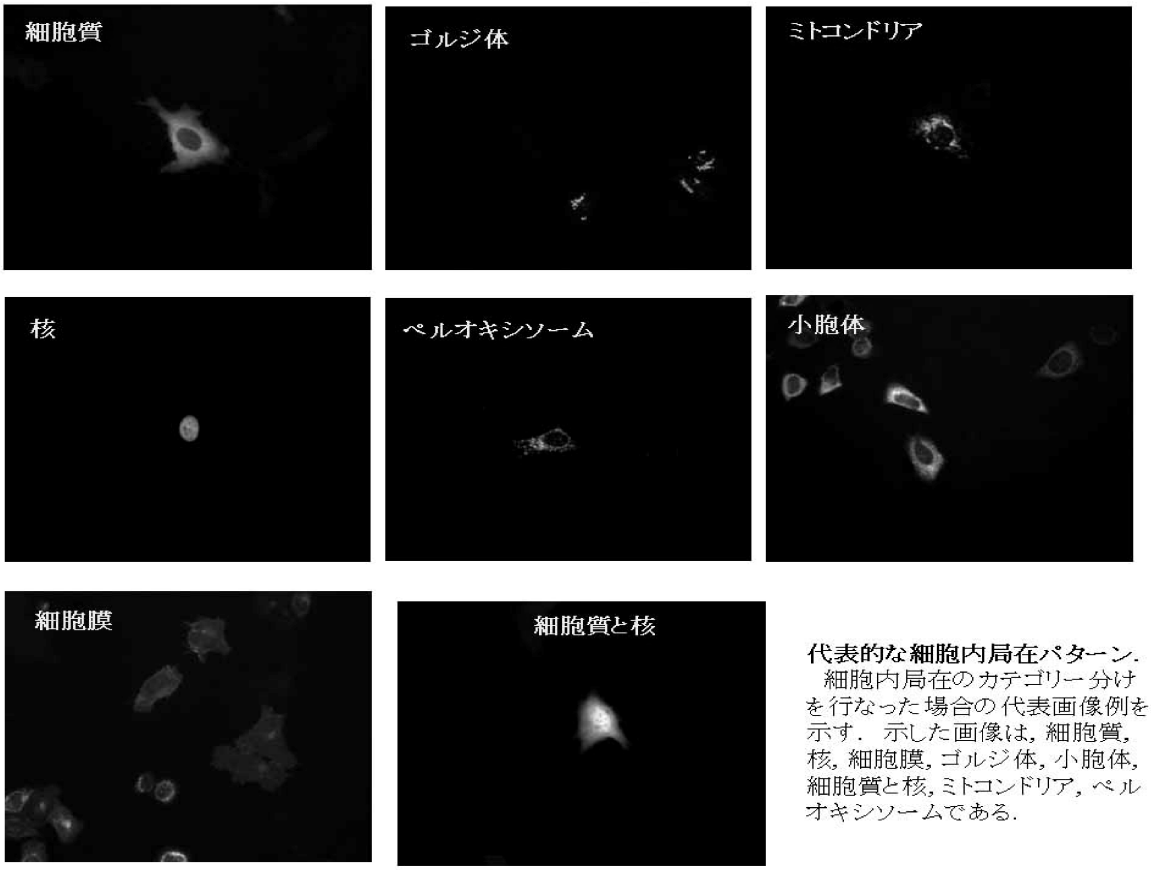


図６　ヒトとマウスのプロモーター領域におけるホモロジーのブロック構造



代表的な細胞内局在パターン.
細胞内局在のカテゴリー分けを行なった場合の代表画像例を示す．示した画像は，細胞質，核，細胞膜，ゴルジ体，小胞体，細胞質と核，ミトコンドリア，ペルオキシソームである．

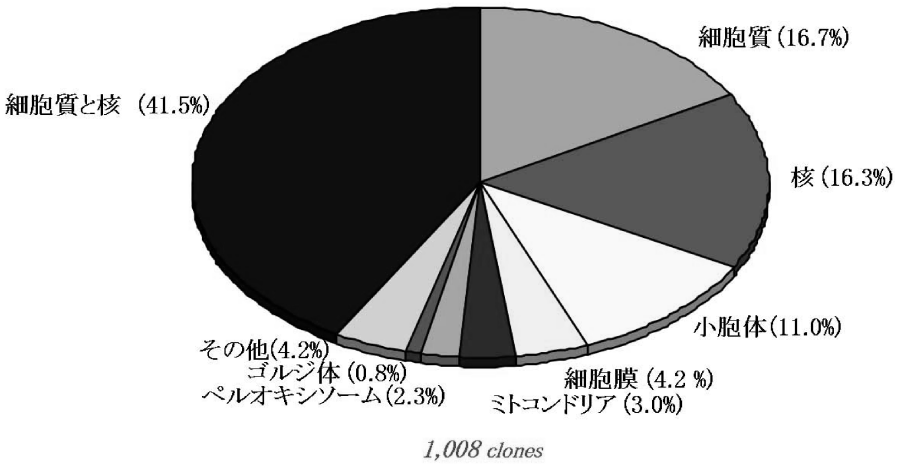
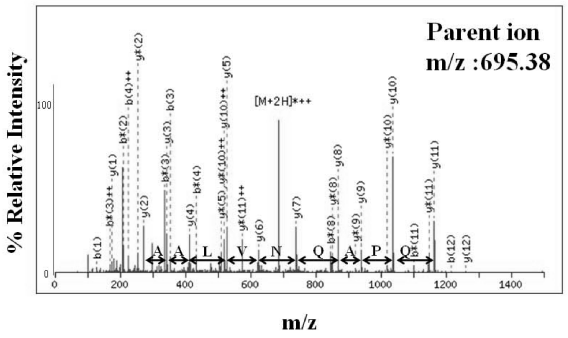


図７代表的な細胞内局在パターンと完全長 cDNA 由来のタンパク質の細胞内局在



NM_015532 novel short CDS (86 a.a.)

MATPARAPESPPSADPALVAGPAEEAECPPPRQPQPAQNVLAAPR
LRAPSSRGLGAAEFGGAAGNVEAPGETFAQRKIHLQIARQR

図８Ａ　質量分析計によるMS/MSパターンと同定されたペプチッドの対応するORFのアミノ酸配列

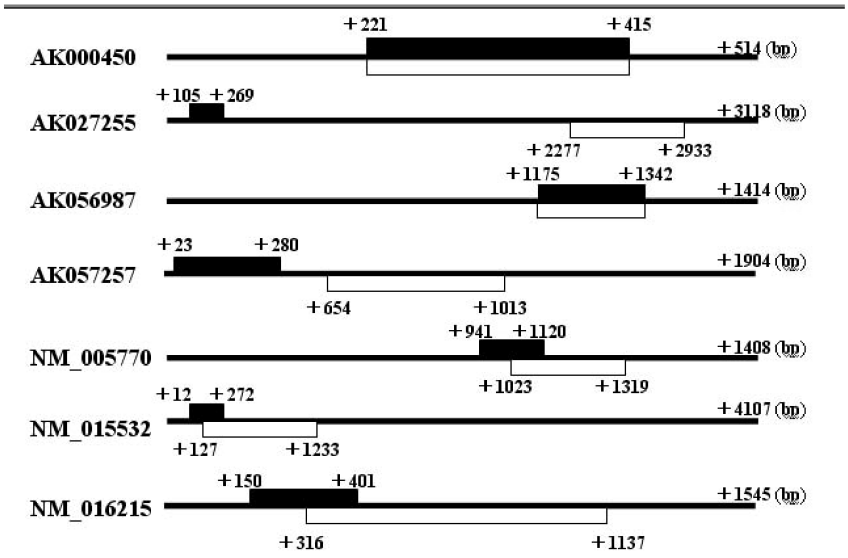


図８Ｂ　新規に同定された小タンパク質をコードするORF（黒の塗り）と同一cDNA中の最長ORF（白抜き）との関係。

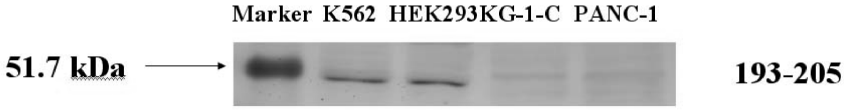


図９　　NM_015532 の最長 ORF の発現の確認。最長 ORF の 193-205 番に対応するペプチドに対する抗体が最長 ORF の分子量に対応するタンパク質をＫ５６２細胞と HEK２９３細胞で検出