博士論文番号：9981043

転写開始点データベース DBTSS の構築と
ヒト・マウスプロモータ領域の網羅的解析

山下　理宇
奈良先端科学技術大学院大学
バイオサイエンス研究科　システム細胞学分野
（小笠原　直毅　教授）

平成１７年１月６日提出

バイオサイエンス研究科　博士論文要旨

所属 （主指導教官）	システム細胞学分野 （小笠原　直毅　教授）		
氏名	山下　理宇	提出	平成 17 年　1　月　6　日
題目	転写開始点データベース DBTSS の構築と ヒト・マウスプロモータ領域の網羅的解析		

要旨

現在、多くの生物種のゲノムが決定されてきている。例えば、ヒトゲノムは 2003 年に終了宣言が出され、遺伝子領域やリピート領域が推定された。また一方ではマイクロアレイなどによる大量の網羅的な発現データが蓄積されてきている。これらを合わせた発現制御ネットワークの解析は、生物学の主要な課題の 1 つである。

転写による遺伝子発現制御機構は、基本的には様々な転写因子とその DNA 結合部位であるシスエレメントとの相互作用によって行われている。この解析には、シスエレメントを含むプロモータ配列が必要であり、それには正確な転写開始点（transcription start site : TSS）の同定が不可欠である。しかし、遺伝子予測プログラムでは、コード領域の予測はできるが TSS の同定は困難である。また、TSS 予測プログラムもいくつかあるが、結果に多くの疑陽性を含むことが知られている。したがって、通常 TSS の推定には、cDNA の 5’端配列が用いられ、それらの代表配列を整理した Reference sequence(RefSeq)がよく使われる。しかし、RefSeq に登録されている mRNA でも、必ずしもその全長を反映しているわけではない。なぜならば、これらの配列は、単に今までに登録された最長の cDNA を代表配列としていることが多いからである。通常の Okayama-Berg 法等の cDNA 合成では、mRNA の 3’端から合成が行われるので、逆転写反応が必ずしも 5’端まで届いている保証がない。もちろん、1 つの遺伝子に対して、5’-RACE や primer extension 等で TSS を決めることもできるが、全遺伝子に網羅的に行うことは現実的ではない。また、1 つの遺伝子に複数のプロモータ（alternative promoter）がある場合、1 遺伝子につき 1 つの転写開始点の情報では、プロモータ領域の解析に誤りを起こす恐れがある。つまり、従来のデータベース中の配列を使用するだけでは、TSS の正確性と網羅性において十分であるとは言えない。

この問題を解決するために、著者らは、TSS のデータベース DataBase of Transcription Start Sites(DBTSS)を構築した。DBTSS には、mRNA の 5’端配列を選択的に抽出できる oligo-capping 法(ヒト)または Cap-trapper 法(マウス)によって決定された 5’端配列を、ゲノムにマッピングした情報が記載されている。これには、11,234 ヒト遺伝子と 7,524 マウス遺伝子の転写開始点が、それぞれ 190,964、195,446 クローンの配列として登録され

ている。さらに、両種の orthologous 遺伝子が NCBI の LocusLink データベースによって対応づけされているので、3226 遺伝子のプロモータ領域の配列比較ができる。そして、-1000~+200 領域に転写因子のデータベースである TRANSFAC の情報を付加し、プロモータ領域に予測される転写因子結合配列を参照することも可能である。

このデータベースを使って TSS を網羅的に見た結果、TSS には揺らぎがあり、そのパターンは二つに大別されることがわかった。一つは、揺らぎの標準偏差が 100 塩基以下であり、こちらは転写の揺らぎを反映していると考えられた。もう一つは、揺らぎの標準偏差が 100 塩基を越えるグループであり、こちらは alternative promoter の可能性が考えられた。また、従来 TSS 付近にあるとされている、4 種のシスエレメントを探索した。正確な TSS が決まると、TATA-box や initiator は、TSS から一定の距離に顕著に検出できたが、downstream promoter element(DPE)や TFIIB recognition element(BRE)は、明確には検出できなかった。さらに、TSS 付近の配列に着目すると、1）initiator を持つもの、2）Terminal oligo-pyrimidine tract(TOP)を持つもの　3）それ以外の 3 グループに分類できることがわかった。

続いて上記 2）で述べた、翻訳レベルで調節を受けるとされている TOP 配列を持つ遺伝子群（TOP 遺伝子）を、ヒトゲノムから網羅的に予測した。検出された TOP 遺伝子群は、リボソームタンパク質，翻訳伸長因子等、従来知られていたものに加え、いくつかの翻訳開始因子，構造タンパク質、サイクリン等、タンパク質合成および細胞分裂に関わる主要な遺伝子群を含んでいた。特に翻訳開始因子は、リボソームに直接結合するような eIF2, eIF3, eIF4A, eIF4B が TOP 遺伝子と推定される一方、直接には結合しないと思われる eIF4G, eIF4E,は TOP 遺伝子ではないと考えられた。この研究において TOP 遺伝子は、従来考えられていた翻訳関係の遺伝子にあるばかりではなく、一部の膜タンパク質、構造タンパク質等、より広範囲に存在する可能性が示唆された。

さらに、最近、プロモータ領域の CpG island と遺伝子発現の組織特異性について矛盾した報告があるので、これらの関係について検討した。まずプロモータ領域に CpG island を持つ遺伝子群（ヒト:6600、マウス 2948）と、持たない遺伝子群（ヒト:2948、マウス 1830）に大別した。そして、組織特異性の指標として expressed sequence tag(EST)をクラスタリングしたデータベース UniGene のソース数を用いた。すなわち UniGene には、ある遺伝子に属する cDNA 配列がどの組織・ライブラリー由来であるかというソースの情報があり、この数はおよその組織特異性の指標となり得る。その結果、CpG を持たない群では UniGene のソース数が少ないことがわかり、CpG を持つ群に比較してより組織特異的に発現していることが示唆された。

本研究で作成したデータベース、およびプロモータ領域の網羅的な解析に関しての情報は、今後の転写制御ネットワークの解析の上で重要な役割を果たすと考えられる。

博士論文番号：9981043

転写開始点データベース DBTSS の構築と
ヒト・マウスプロモータ領域の網羅的解析

山下　理宇
奈良先端科学技術大学院大学
バイオサイエンス研究科　システム細胞学分野
（小笠原　直毅　教授）

平成１７年１月６日提出

1. 序論

本論文は、真核生物のプロモータ領域解析に欠かせない転写開始点を、大規模な 5'EST 配列に基づき構築したデータベース DBTSS(Database of transcription start sites) と、正確な転写開始点情報に基づくプロモータ領域解析の研究報告である。まず、本章では、本研究の位置づけを明確にする。最初に、現存するゲノムデータベースをプロモータ領域解析に用いる場合の問題点について述べる。続いて、プロモータ領域解析を行う DBTSS 以外の既存のデータベースについて述べる。最後に、論文の構成を述べ、第 2 章以降の研究内容への導入とする。

1.1. ゲノムデータベースと遺伝子データベース

1.1.1. ゲノムデータベース

生物の進化や生体制御のメカニズムを解析する際、ゲノム配列を決定することは、非常に有効な手段と考えられている。1995 年にインフルエンザ菌の全ゲノムが決定(Fleischmann et al., 1995)されて以降、今日までに様々な生物種のゲノムが決定されてきた。最近では、酵母(Goffeau et al., 1996)を始めとする真核生物や、線虫(CESC, 1998)、ショウジョウバエ(Adams et al., 2000)といった多細胞生物も決められてきた、さらに、ヒト(Lander et al., 2001; Venter et al., 2001)やマウス(Waterston et al., 2002)のドラフト配列の公開は、シークエンス時代の大きな金字塔といえる。ヒトに関しては、2004 年に完了宣言が出され、遺伝子領域数が 20000~25000 であると見積もられた(IHGSC, 2004)。

決定された配列は、公共の機関を通して公開されており、誰でも利用可能である。多くの研究者に使われているものが、アメリカの National Center of Biotechnology(NCBI)の Map Viewer (www.ncbi.nlm.nih.gov/mapview)、University of Carifornia Santa Cruz(UCSC)の UCSC Genome browser (//genome.ucsc.edu/)、イギリスの European Bioinfomatics Instituite(EBI) の Ensembl Genome browser (www.ensembl.org)であり、同じゲノム配列を元にしてそれぞれ独自にブラウザーを構築している。

1.1.2. ヒト・マウスの遺伝子データベース

ゲノムが決まったとき最も重要なことの一つは、その生物の遺伝子領域を推

定することである。遺伝子領域を推定するには、ゲノム配列を開始コドンから終始コドンまで偶然とは考えられないほど長く続く領域(Open Reading Frame: ORF)の探索、コドンの使用頻度から隠れマルコフモデルを構築、スプライシングサイトを予測する等、後述する cDNA 配列の相同性による探索を用いないで、既存の遺伝子の数学モデルに基づく *ab initio* に決める手法がある。しかし、*ab initio* で遺伝子領域を予測すると、真核生物の場合精度は低い。従って、これだけで遺伝子領域を決定するのは現実的ではない。そこで、それぞれの生物種の転写産物に対応する cDNA 配列情報や、他生物種の遺伝子領域と相同性検索することを組み入れて、遺伝子領域の予測が行われている。この相同性検索を取り入れた手法でもっとも代表的な物が、EBI が提供する Ensembl gene(Hubbard et al., 2005; Hubbard et al., 2002)である。この配列セットには、22,287 の遺伝子領域が、スプライシングバリアントを含む 34,214 転写産物として登録されており、ヒトの遺伝子領域数を 25,000 と仮定すると 89%をカバーしていると見積もられている(IHGSC, 2004) 。

また、cDNA だけを元にしたデータベースもある。例えば GenBank(Benson et al., 2004)の中には、cDNA として決められた配列が納められている。しかし、この中の配列は、一つの転写産物に対して複数の配列が登録されているので、重複が非常に多い。このような重複を取り除く目的で、機械的に non-redundant（nr）なデータセットも作られている。しかし、ある遺伝子の登録 cDNA 配列が断片化している場合には、別々の登録配列として取り扱われている。このように一つの遺伝子に対して複数の cDNA がデータベース中に存在するということは、遺伝子の同定に支障をきたすばかりでなく、遺伝子数の過剰な見積りを引き起こす。この冗長性を取り除く目的で、NCBI では代表配列のセットである Reference sequence、通称 RefSeq(Maglott et al., 2000; Pruitt and Maglott, 2001)を提供している。この中には、cDNA 配列が、「NM_」という ID で始まって登録されている。この遺伝子セットは、ある遺伝子に対して Genbank 中の cDNA 配列で最も品質（長さ、精度）が良いものを、人の目で見て選び直して代表遺伝子としたものである。RefSeq には、ある遺伝子が単に cDNA としての予測だけなのか、それともコードするタンパク質が知られているのか、という信頼度も付加されている。現在は、Ensembl gene か RefSeq のどちらかが、遺伝子の標準セットとして研究者に使われている。

RefSeq や Ensembl gene は、代表遺伝子セットであるが、代表遺伝子領域では

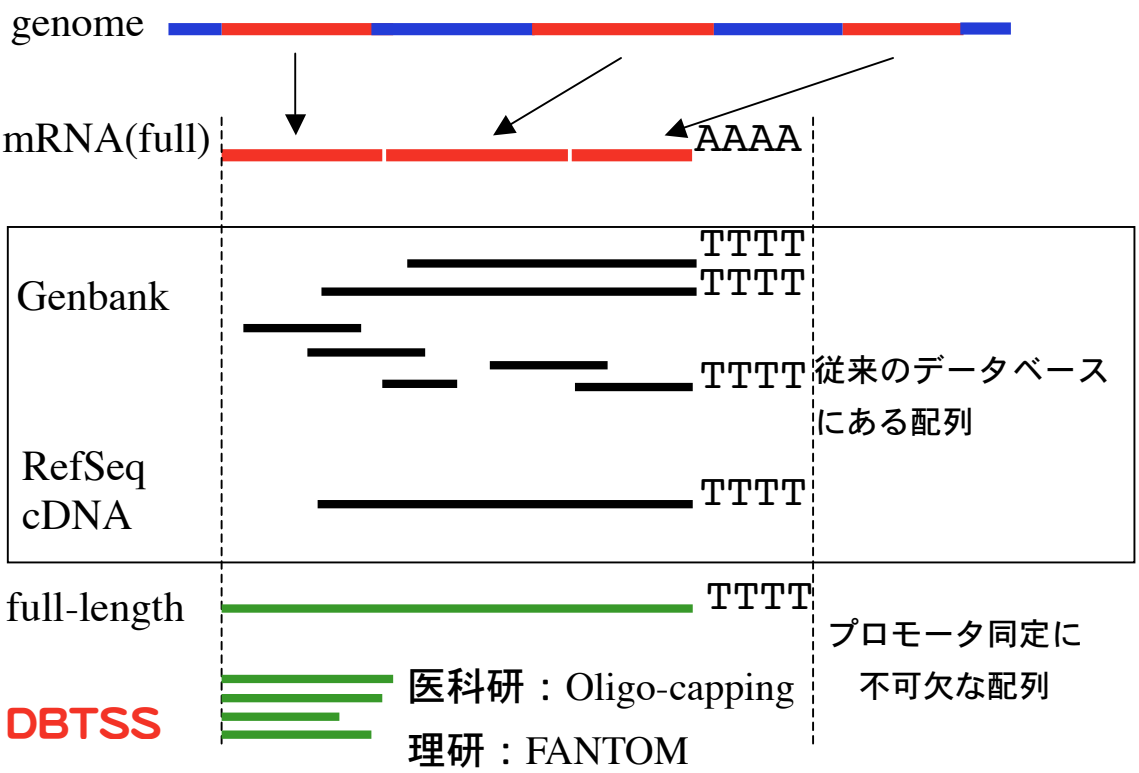


図 1-1　DBTSS の意義

現存するデータベースと、プロモータ領域の解析に必要なデータベースの関係。プロモータ領域同定には、5'端が保証された配列情報が欠かせない。

ない。なぜならば、スプライシングバリアントがある遺伝子の場合、ゲノム上では同じ領域に複数の遺伝子が存在するためである。遺伝子領域の代表的なデータセットは、NCBI が提供する LocusLink である。LocusLink には、その領域に存在する RefSeq 遺伝子、オーソログ遺伝子、及び、Gene ontlogy（遺伝子機能や発現部位に関する情報）の情報も併記されている。

1.2. 既存の転写開始点推定法とその問題点

ゲノムと遺伝子領域が決定されると、その遺伝子の制御領域の解析が、次の課題になる。近年、マイクロアレイなど大量に mRNA の発現量を測定する技術が登場してきた。これらの発現情報を使って転写因子結合部位の推定が、同じように発現している遺伝子は共通の転写機構により発現が制御されている、という仮定の下に行われている。このような解析には、正確な転写開始点に基づく転写開始点付近のゲノム配列情報が不可欠である。

現在、転写開始点付近の配列を得る手段は、文献情報または cDNA 情報のど

ちらに基づくかで二つに大別できる。前者に位置するもので代表的なデータベースは、EPD(Eukaryote Promoter Database)(Praz et al., 2002)である。これは、文献情報をデータベース化したものであり、信頼性は高く、多種の真核生物に渡ってプロモータが登録されている。しかし、登録されているデータは、2005 年の Release 81 で全真核生物で 4809 プロモータと未だ少なく、網羅的であるとは言い難い。

後者に位置するものとして、既知の遺伝子をゲノム上にマッピングし、その 5'端を基準にゲノム上の配列を抜き出すという手法がある。この手法は、データベースに記載されている cDNA が、mRNA の 5'末端を含んでいることが前提となる。しかし、上述した遺伝子データベースの RefSeq や Ensembl gene に登録されている配列は、5'末端が保証されていない。そこで、後述するヒト・マウスなどで報告されている完全長 cDNA 配列を、GenBank から抜き出してゲノム上にマッピングし、その 5'末端前後をプロモータ配列として利用するという方法がある。これらは、5'末端の保証があるという点では有用なデータセットである。しかし、一つの遺伝子領域に対応する完全長 cDNA は数が少ないので、複数の転写開始点がある alternative promoter の検出や、転写開始点の多様性を解析することは困難である。

その他の手法として、プログラムを用いて転写開始点を予測するという方法がある。しかし、数多くのプロモータ予測プログラムが知られているが、どれも性能に問題があるため実際に使えるとは言い難い。例えば Bajic らの８種のプロモータ予測プログラムを性能比較した報告によると、NNPP は最も高い 93％の sensitivity を記録したが、specificity は 2.8%と非常に低い。逆に、McPromoter は 78%の spcificity を示したが、sensitivity は 27%程度である(Bajic et al., 2004)。

1.3. 5'末端を保証した cDNA 構築法と完全長 cDNA プロジェクト

1.3.1 mRNA の 5'末端を含む cDNA 合成法

真核生物の完全長の mRNA の 5'末端には、三個のリン酸基を介して 7 メチル化グアノシンが結合しているキャップ構造がある。これに対して、5'末端が欠けている mRNA は、リン酸基や水酸基が露出している。したがって、5'端にキャップ構造を持つ mRNA を選択的に抽出すれば、5'末端が保証されている mRNA を得ることができる。代表的な手法として、以下の oligo-capping 法と Cap-Trapper 法という二つがよく使われる。

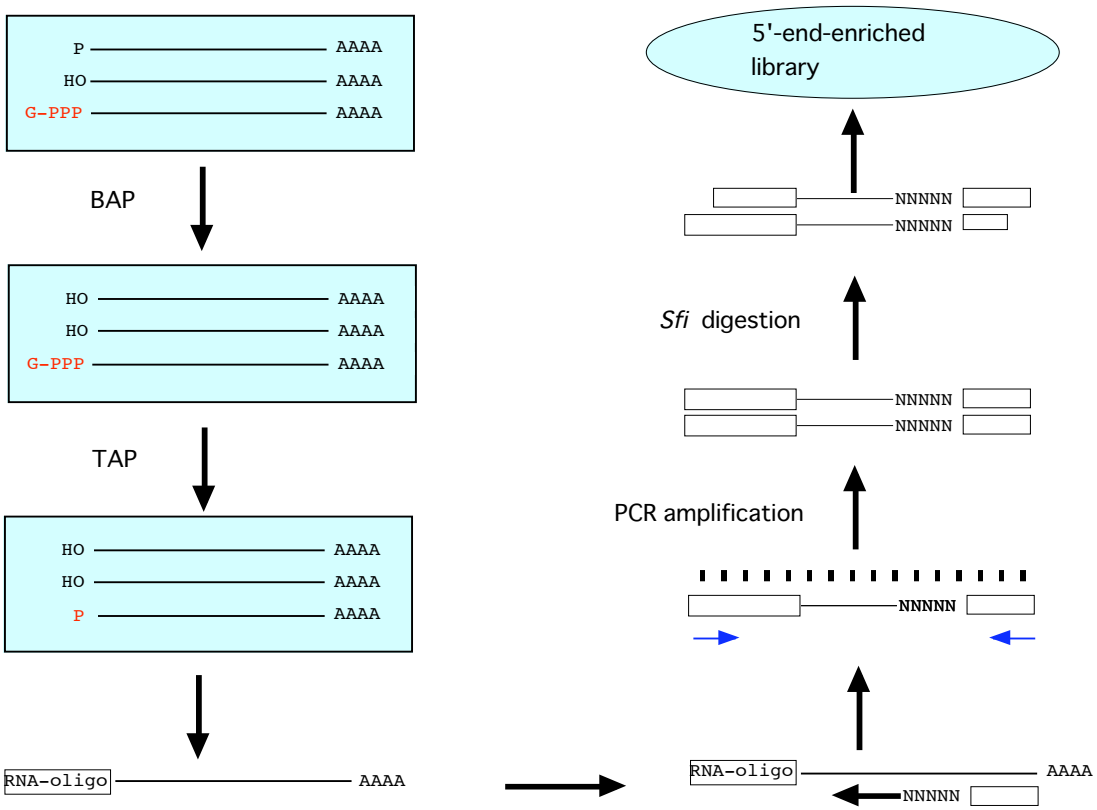


図 1-2　oligo-capping 法概略

mRNA 集団に対して BAP 処理を行うと、不完全長の 5’端である P が OH に置き換わる。この後、TAP 処理を行うと、完全長の 5’端のみが P 基が露出する。これに RNAoligo を反応させると、5’端に P 基があるものすなわち完全長の RNA のみが結合する。この後、ランダムプライマーや oligodT プライマーを用いて逆転写反応を行い、RNAoligo 内の配列と、逆転写の時に使ったプライマー内に含まれる配列で PCR を行うと、完全長の cDNA のみが、増幅されることになる。

Oligo-capping法(Maruyama and Sugano, 1994; Suzuki et al., 1997)の原理は、以下の通りである。mRNAに対してBAP(Bacteria Alkaline Phsphatase)処理を行うと、不完全長の5’端であるリン酸基が水酸基に置き換わる。その後、TAP(Tobacco Acid Pyrophosphatase)処理を行うと、mRNA集団のうち完全長の5’端リン酸基のみが露出していることになる。これにRNA-oligoを反応させると、5’端にリン酸基があるmRNAのみと結合する。この後、タグの付いたランダムプライマーやoligo-dTプライマーを用いて逆転写反応を行い、RNA-oligo内の配列と、逆転写の時に使ったプライマーのタグに含まれる配列をプライマーにしてPCRを行う

と、5’端にキャップ構造のあったmRNAに対応するcDNAのみが、増幅されることになる（図1-2）。

もう一つのCAP-Trapper法(Carninci et al., 1996; Carninci et al., 1997)の原理について述べる。まず、完全長mRNAの7メチル化グアノシンを$NaIO_4$によって開裂させる。次に、この開裂したグアノシンにビオチンを付加させる。5’端にメチル化グアノシンのないmRNA、すなわち不完全長mRNAにはビオチンが付加されない。このmRNA集団に対して、oligo-dTプライマーを用いて逆転写反応を行い、DNA-RNAのハイブリッドを作成する。さらにこのDNA-RNAハイブリッドにRNaseIを反応させると、逆転写反応が5’端まで達していないハイブリッドは、RNAの5’側が分解され、結果としてビオチン部分が外れることになる。したがって、アビジンビーズを用いてビオチンを持つハイブリッドを回収すると、CAP構造のあったmRNAに対応する一本鎖cDNAのみが回収できる。アビジン回収後、RNAを加水分解、二本鎖DNA合成を行うことで、5’末端の保証された完全長cDNAを得ることができる。

1.3.2 ヒト・マウスの完全長 cDNA プロジェクトと 5’EST 配列

ヒト・マウスでは、1.3.1 で述べた作成技術によって、完全長の転写産物を大規模に配列決定、及び、その機能推定を行う cDNA プロジェクトが行われている。この配列解析は、ヒトでは oligo-capping 法を用いて、東京大学医科学研究所とかずさ DNA 研究所が中心となって行われてきた(Imanishi et al., 2004; Ota et al., 2004)。またマウスでは、理化学研究所が中心となって Cap-Trapper 法により配列解析が進められてきた(2004; Kawai et al., 2001; Okazaki et al., 2002)。各プロジェクトでは、ヒトでは 21,243 種、マウスでは 60,770 種の転写産物を報告している。

上記のプロジェクトでは、未知遺伝子の発見・機能解析が主眼に置かれている。従って、oligo-cappin 法や Cap-Trapper 法によってクローニングされた配列が、5’側から配列決定され既知遺伝子であると同定されると、そのクローンはそれ以上配列決定されない。既知遺伝子に当たったそのような 5’端配列は、データベースに登録される以上に利用されていない。しかし、図 1-1 のように、このような 5’末端が保証された大量の配列こそが、転写開始点の同定には必要な配列なのである。

1.4. 本論文の構成

本研究では、多量の 5'端配列をゲノム上にマッピングすることによって、大規模な転写開始点データベース DBTSS(DataBase of transcription start sites)の構築とそれを利用したプロモータ領域の網羅的解析について述べる。

以降、2 章では、DBTSS の構築法を中心に述べる。DBTSS は、5'端配列を既知遺伝子と対応付けしてゲノムにマッピングしたデータベースである。ここには、11,234 ヒト遺伝子と 7,524 マウス遺伝子の転写開始点が、それぞれ 190,964、195446 クローンの配列として登録されている。この情報は、web ページを通して公開している。さらに DBTSS に登録されている遺伝子を観察した結果わかった 1)転写開始点がそろっている、2)転写開始点に揺らぎがある、3)alternative promoter を持つ遺伝子群について述べる。

3 章では、転写開始点近傍配列の特徴について述べる。まず従来、転写開始点付近にあるとされている 4 種のシスエレメントの探索を行った。正確な転写開始点が決定されると、TATA-box や、initiator は転写開始点から一定の距離に観察されたが、downstream promoter element、TFIIB recognition element は顕著に見いだすことはできなかった。さらに、転写開始点付近に着目すると、initiator を持つもの、terminal oligo pyrimidine(TOP)配列を持つもの、それ以外の 3 グループに分けることができた。

4 章では、3 章で明らかになった翻訳レベルで調節を受けるとされている TOP 配列を持つ遺伝子群（TOP 遺伝子）を、DBTSS とゲノム情報を合わせて予測した。検出された TOP 遺伝子群は、リボソームタンパク質、翻訳伸張因子等従来知られていたものに加え、いくつかの翻訳開始因子、構造タンパク質、サイクリン等タンパク質合成や細胞分裂関わる主要な遺伝子を含んでいた。また、この研究において TOP 遺伝子は、一部の膜タンパク質、構造タンパク質等、より広範囲に存在する可能性があることがわかった。

5 章では、プロモータ領域の CpG island と遺伝子発現の組織特異性について矛盾した報告があるので、これらの関係について検討した。まず、プロモータ領域に CpG island を持つ遺伝子群（ヒト:6600、マウス:2948）と持たない遺伝子群に分けた。そして、組織特異性の指標として、expressed sequence tag(EST) クラスタリングしたデータベース UniGene のソース数を用いた。UniGene には、一つの遺伝子に属するクラスタリングされた cDNA 配列が、どの組織・細胞由

来であるかというライブラリのソース情報があり、この数はおおよその遺伝子の組織特異性の指標となり得る。その結果、ヒト・マウス共に、CpG island を持たない遺伝子群では、持つ遺伝子群に比べて組織特異的に発現していることがわかった。

2. 転写開始点データベース DBTSS の作成

2.1 序論

１章で述べたように、DBTSS が登場する 2001 年まで、ヒトやマウスのプロモータの情報を得ようとした場合の情報源は、大きく二つに大別される。１つは、Eukaryotic promoter database(EPD)(Praz et al., 2002)に登録されている情報である。ここには、文献を基に構築されたプロモータ領域の情報が載っている。しかし、2005 年現在、ヒトで 1871、マウスで 196 遺伝子と少なく、網羅的なデータベースであるとは言い難い。もう一つは、cDNA 配列をゲノムにマッピングしてある情報を、ゲノムブラウザとして公開されている UCSC，NCBI、Ensenbl などから取り出してきて利用する方法である。しかし、それぞれの cDNA 配列の 5’端は、mRNA の 5’末端が保証されているわけではないので、真の転写開始点ではない可能性がある。

近年、5’末端が保証された完全長 cDNA を大量に得る目的で、大規模な cDNA 配列決定がヒト・マウスに対して行われてきた。ヒトに関しては、東京大学医科学研究所とかずさ DNA 研究所が中心になって oligo-capping 法によって作られた 21,243 種の完全長 cDNA を配列決定している(Imanishi et al., 2004; Ota et al., 2004)。また、マウスにおいては、理化学研究所が中心になり、Cap-Trapper 法によって作られた完全長 cDNA 60,770 種を報告している(Kawai et al., 2001; Okazaki et al., 2002)。両者の配列解析では、まず、完全長の cDNA 集団からランダムに選ばれたクローンが、5’端から配列決定される。もし、決定された 5’端配列が既知遺伝子に存在しなかった場合、そのクローンはさらに 3’側に伸ばして全長が決定される。この過程で大量に配列決定される cDNA の 5’端配列は、mRNA の 5’末端、すなわち転写開始点が保証されている。したがって、この 5’端配列をゲノムにマッピングすることで、ゲノム上の転写開始点を大規模に発見することができる。

我々のグループでは、2001 年に 217,402 のヒト由来の 5’端配列をゲノムにマッピングして、7889 遺伝子の転写開始点を 111,382 のクローンの配列として登録したデータベース DBTSS（Databse of Transcription start sites）を構築し報告した(Suzuki et al., 2002)。その後 ver.2 では、東大医科学研究所菅野研究室で配列決定されたマウスの 33,482 の 5’末端配列も利用して、マウスの DBTSS も構築した。また、ヒトの 5’端配列を 400,225 に増やすことによって、9336 遺伝子に関して 183,845 配列を登録し、より多くの遺伝子に関する転写開始点情報を付

表 2-1　DBTSS の歴史

左側の列から、バージョン、公開時期(open)、マッピングしたゲノム(genome)、元になった 5'est 数（#clones)、DBTSS に登録されているクローン数（mapped)、NCBI　RefSeq 数(#NM)、NCBI LocusLink 数#LocusLink)を示す。

		human					mouse				
	open	genome	#clones	#mapped	#NM	#LocusLink	genome	#clones	#mapped	#NM	#LocusLink
ver.1	2001Sep.	hg7	217402	111382	7889	-	-	-	-	-	-
ver.2	2002Mar.	hg11	400225	183845	9336	-	mm1	33482	14606	2789	-
ver.3	2003May.	hg13	400225	190964	11234	9470	mm2	580209	195446	7524	6875
ver.4	2004Nov	hg16	400225	277794	15536	12780	mm3	580209	290714	11116	10933
ver.5	not yet	hg17	400225	278700	17411	13292	mm5	580209	364487	14745	14162

加させた。Ver3 では、理化学研究所で配列決定されたマウスの 5'端配列を加え、580,209 クローン配列のうち、195,446 配列を 7524 遺伝子として登録した。また、ver.3 からは、ヒト-マウス間でオーソログ関係のものについて、プロモータ領域の比較ゲノム解析を視覚的にできるようにした(Suzuki et al., 2004)。Ver4 以降は、5'端配列数は増えていないが、ゲノム配列が更新されるたびにマッピングを行い、現在では、ヒト 17411 遺伝子、マウス 14745 遺伝子の各転写開始点が 278,700、364,487 のクローン配列として登録されている（表 2-1)。

DBTSS の作成手順は、ver.3 の時に固定され、それ以降のバージョンではルーチンワークとして行っている。以下この章では ver.3 の DBTSS の作成手順、特徴、及び、その問題点について述べる。

2.2　材料と方法

2.2.1. DBTSS 構築に用いたデータセット

ヒトの 5'端配列は、東大医科研菅野研究室で配列決定された、143 種のライブラリ由来の 400,225 配列を用いた。マウスに関しては、菅野研究室で oligo-cappinng 法によって合成され配列決定された 4 ライブラリ 33,482 配列に加え、理化学研究所の林崎研究室で Cap-Trapper 法によって決定された 126 ライブラリ由来の 5'端配列 546,727 配列を用いた。

2.2.2. ヒトの 5'端配列の前処理

今回、元になった 5'端クローン配列は、ベクター部分が切り取られている予定であった。しかし、ヒトの 5'端配列には、明らかにベクターサイトが取り除

図 2-1　mapping 時に問題になる配列例

クローン DMC01903 は、ベクターサイト（赤字非下線部）が取り除けていなかった。また、ゲノム上にマッピングされず、アーティファクトと考えられる長い A が連続した部分（赤字下線部）があった。なお、ゲノムに正しくマッピングされるのは、黒字部分のみであり、A が連続した部分は、約 10 万塩基離れている A の連続した simple repeat 領域にマッピングされた。このような配列は、手作業で処理するのが困難なため、除去することにした。

かれていない配列が見つかった。また、原因は不明であるが、5'端に極端に長い A の連続配列があるクローンが存在した。例えば図 2-1 の例では、ベクターサイトがある上に、ベクター部分と cDNA の 5'端の間にゲノム上に存在しない極端に長い A の連続配列がある。しかしながら、ゲノム中には A の連続配列のような単純な反復配列が多数存在するため、誤ってマッピングされることがある。こうしたクローンを避けるために、5'端 100base 以内に 10 塩基以上 A が連続して存在しているクローンは解析の対象から外した。また、全 5'端配列においてベクターサイトの再検索を 5'端 100base 以内に対して一塩基の欠失、置換、挿入は許すようにして検索した、見つかった場合には、そのベクターサイトを取り除いた部分を 5'端配列とした。

2.2.3. Refseq に対する相同性検索

既知遺伝子との対応付けを行うため、最初に BLAST を用いて(Altschul et al., 1997)RefSeq の NM で始まるもの（重複無しの mRNA 遺伝子セット）に対し、相同性検索を行った。この検索では、BLAST の期待値（e-value）が 10^{-100} 以下であった場合、その RefSeq の登録配列と同一遺伝子であるとした。RefSeq と対応が取れなかった 5'端配列は、まず非常に高速だがやや精度が劣る BLAT(Kent, 2002)を用いて直接ゲノムへマッピングしておおよその位置を定めた。次に、そ

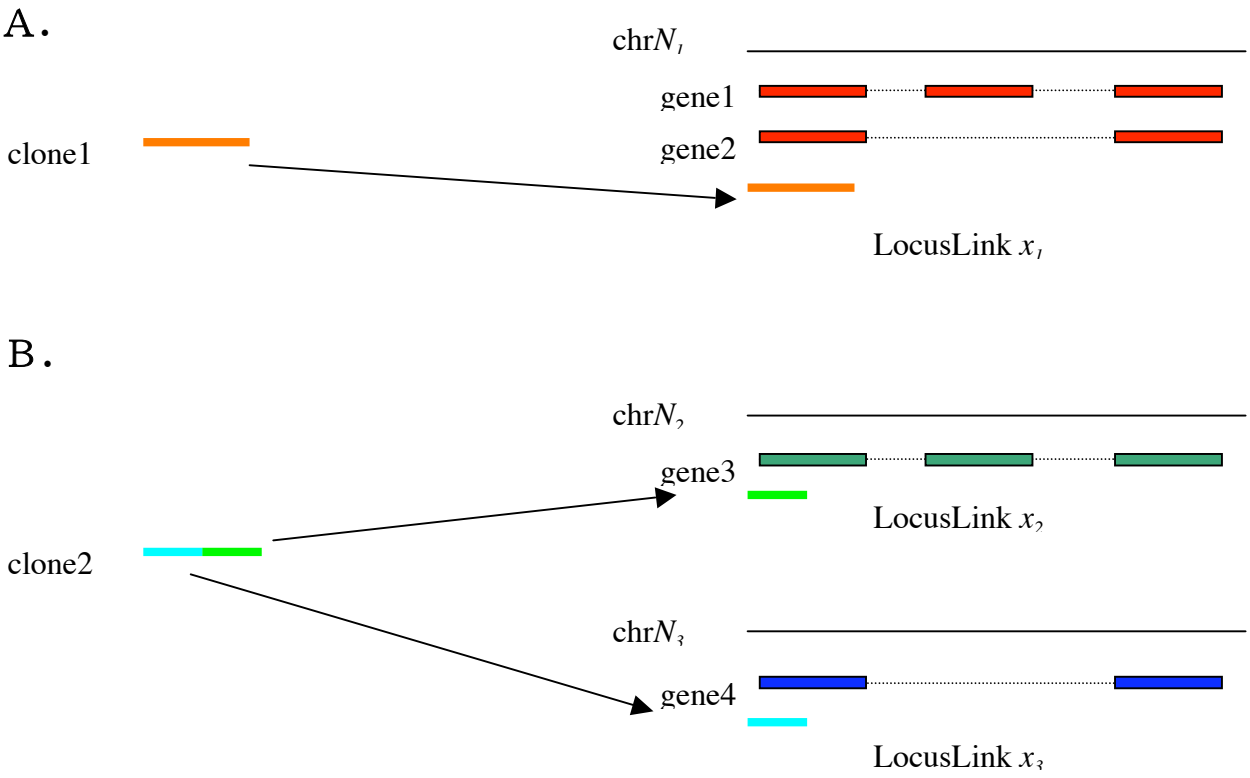


図 2-2　一つのクローンが複数の遺伝子に対応する状況

一つのクローンが複数の遺伝子に対応する状況として考えられるのは、次の二つの状況が考えられる。A:複数の遺伝子がゲノム上の同じ領域にスプライシングバリアントとして登録されている場合。この場合、クローンは、複数の遺伝子に相当し、どちらの遺伝子を代表しているかについては決定できないが、ゲノム上は同じ座標にあり転写開始点は決められる。B:一つのクローンがキメラになっている場合。この場合のように、一つのクローンがゲノム上の別の領域にマッピングされる遺伝子に対応するとき、そのクローンはアーティファクトの一つのキメラの可能性が高い。従って、このようなクローンは除去の対象となる。この場合、遺伝子に相当するLocusLinkID が異なるので A の様な場合と区別できる。

の 5'端から-1M~+1M のゲノム配列に対して、低速であるが、精度が高いマッピングソフトである sim4(Florea et al., 1998)で再度マッピングした。このマッピング結果が、既知の RefSeq のエクソン領域と一部でも重なっている場合、5'端配列はその RefSeq 遺伝子由来であるとした。

いくつかの遺伝子において、1 つの clone が複数の RefSeq 遺伝子に当たってしまう状況が観察された。この現象は、1)遺伝子の同一領域の複数のスプライシングバリアントに当たっている場合（図 2-2 A)、2)クローンがキメラである場合（図 2-2 B）が考えられる。従って、それぞれの Refseq が、ゲノム上の遺伝子領域である LocusLink 一つのみに対応する場合には、単一の遺伝子領域由

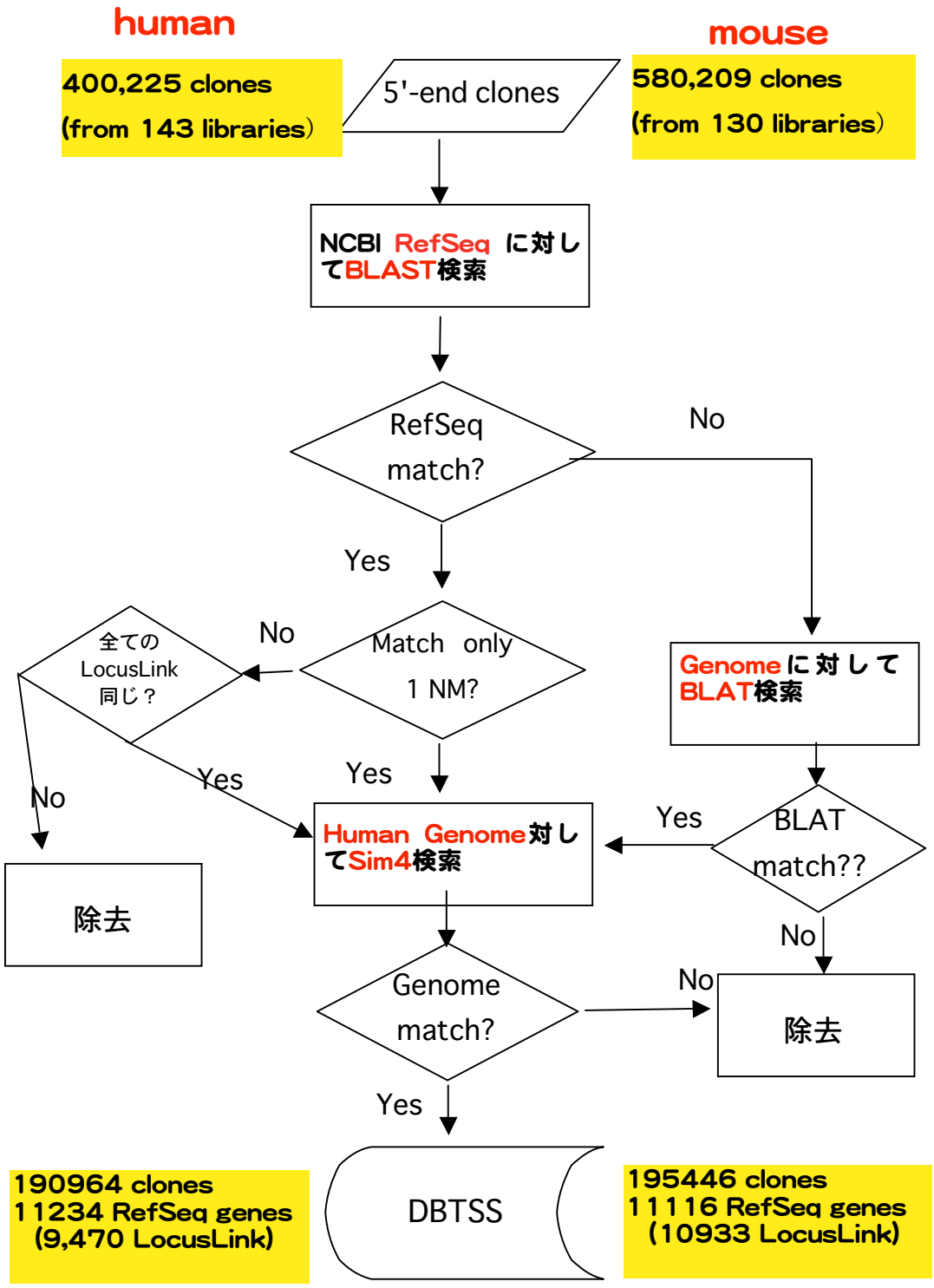


図 2-3　マッピングの手順

マッピング手法は、ver3 以降ではこのときに確立された手順をそのまま用いていて、クローン数のみが増加している。

来のクローンと判断した。LocusLink ID が複数になっているクローンは、キメラになっている可能性が高く、解析の対象から外した。

2.2.4. ゲノムの遺伝子領域の切り出し

University of Carifornia Santa Cruz(UCSC)が提供する UCSC Genome browser の Web ページには、RefSeq 遺伝子がゲノムのどの領域にマッピングされているかという情報がある。これは、refGene.txt という名前のテキストファイルになっており、RefSeq cDNA とその cDNA のエクソン部分のゲノム上の位置を対応付けできるようになっている。この遺伝子マッピング情報の中で最長の遺伝子は 1.9M のゲノム配列を持つ low density lipoprotein-related protein 1B であった。したがって、全ての遺伝子は 2M base の領域にマッピングが可能であると推測できる。今回、全ての RefSeq に登録された cDNA の配列に対してその 5'端配列から-1M~+1M base 抜き出した。そして、その cDNA にヒットした 5'端配列を、この 2M 塩基の範囲内に対して sim4(Florea et al., 1998)を用いてマッピングを行った。このように範囲を絞って検索した最大の理由は、マッピングにかかる時間の節約である。sim4 でマッピングすると、この 2M 塩基の領域の場合、Sun 400Mhz 1CPU の環境下で 1 クローン当たり 1 秒前後の時間を要した。もしヒト・マウス合わせて約 100 万のクローンに対して、この作業を約 3G 塩基のゲノム全体に行うとすると、100CPU 使ったとしても約半年かかる計算になり現実的ではない。

2.2.5. 低クオリティクローンの除去

oligo-capping クローンは、5'端を持つクローンを選択的に収集する手法である。クローンの 1 塩基目からゲノムにマッピングされていない場合には、2.2.1 で述べたように、ベクターサイトが完全に取り除けていない場合や、原因不明の A が付加されている可能性が考えられる。従って、ヒト遺伝子に関しては、クローンの 1 塩基目からマッピングされていないデータは全て除去した。

Cap-Trapper 法由来の 5'端配列は、原因は不明であるが、最初の 1 ないし数塩基がゲノム上にマッピングされない場合がある。数塩基程度の短い配列であれば、sim4 はゲノム上の誤った位置にマッピングすることはない。従って、Cap-Trapper 法由来の 5'端配列は、ゲノム上にマッピングされた最初の 5'端の配列を転写開始点とした。

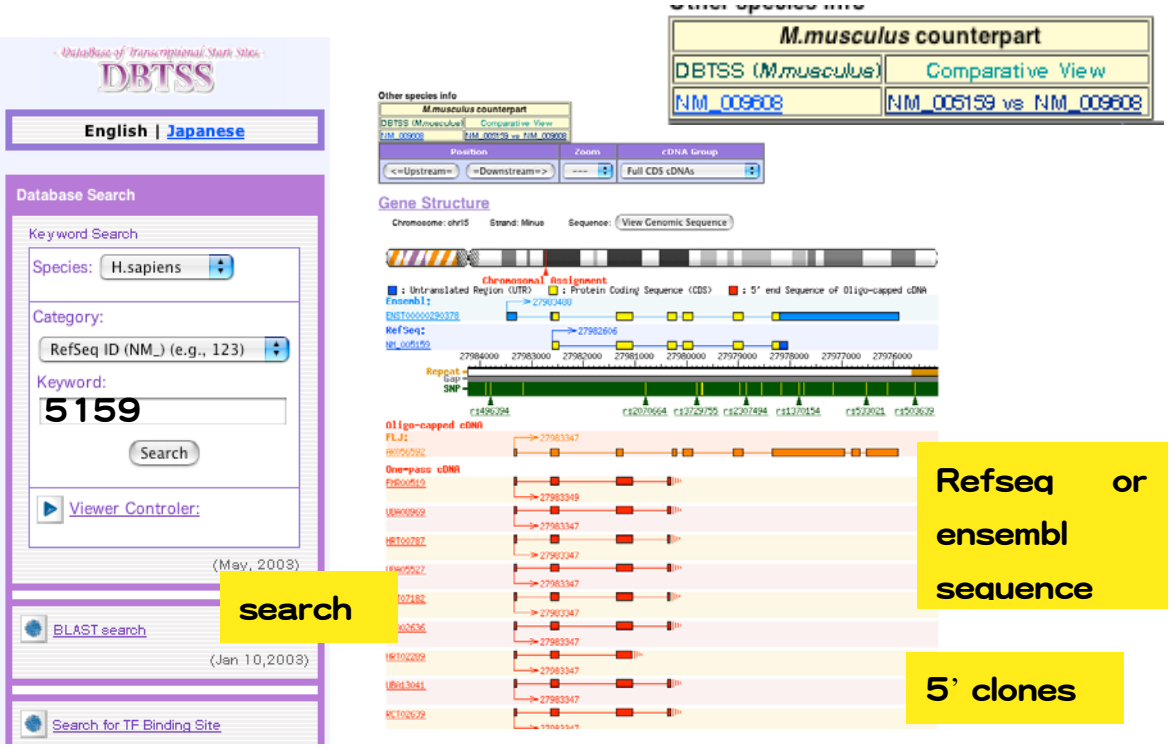


図 2-4　DBTSS 初期画面と結果の例

DBTSS で NM_005159 を検索したときの例を示した。検索 winedow に 5159 と入力して、refseq を選択すると、右のような結果が得られる。Oligo-capping clone のマッピング情報により、転写開始点がどこにあるかを見ることができる。この例では、NCBI の Refseq は、最も 5'側のエクソン（第一エクソン）の情報がないことがわかる。マウスとオーソログ関係が取れる場合、上の「comprative view」をクリックすることで、図 2-5 のプロモータ領域の comparative genome の画面に移ることができる

2.2.6. データベースの構築

データベースのスクリプトは全て perl 言語で書かれている。画像は、perl の GD モジュールを使って書き出し、HTML 言語で表現している。データの管理には Mysql を用いている。実際の web ページの作成は、株式会社ダイナコムに委託した。データは DBTSS として web 経由（http://dbtss.hgc.jp）で公開されている。

2.3　結果

2.3.1 DBTSS データ統計

RefSeq に対してマッピングを行ったヒト 400,225 クローン、マウス 580.209 クローンのうち、RefSeq 遺伝子にヒットしたものは、ヒト 307,734 クローン、

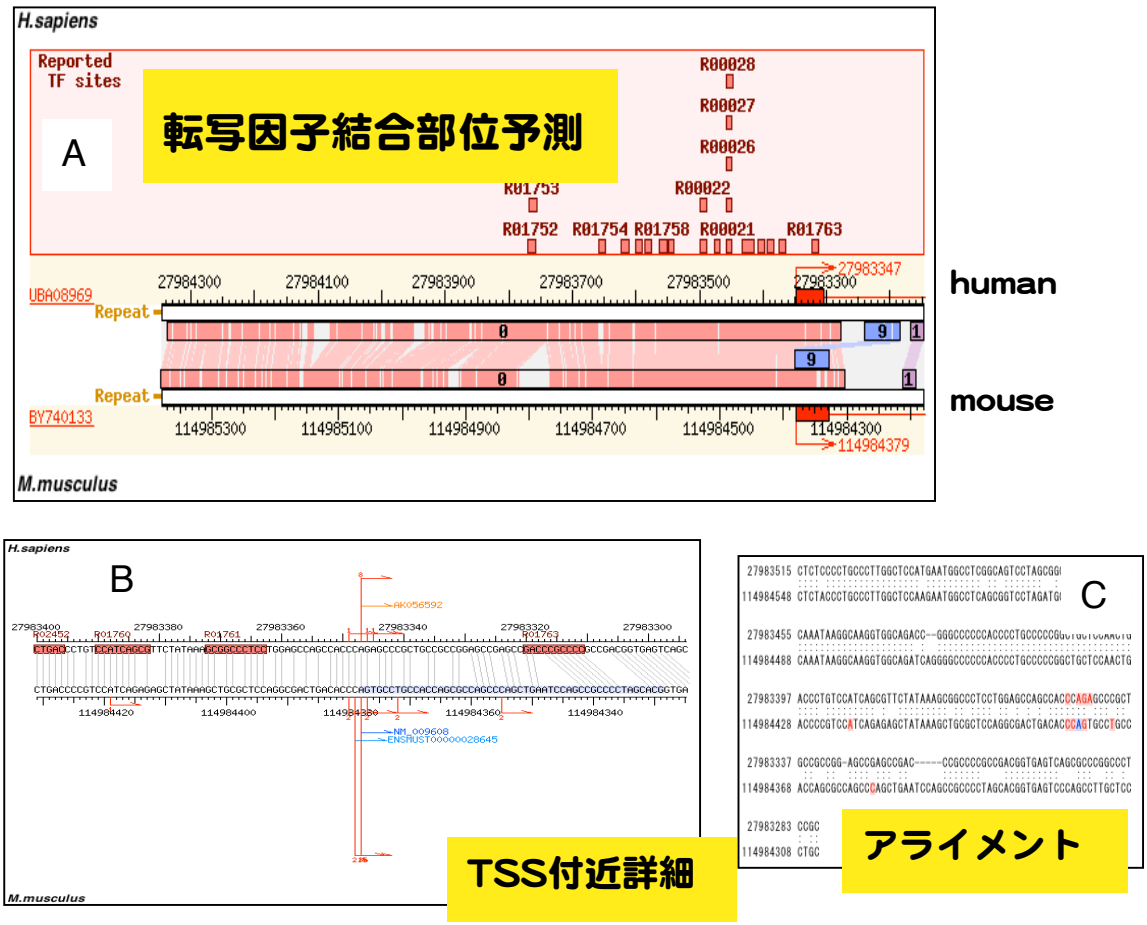


図 2-5　comparative genome 例

Comparative genome の例を示した。A では、lalign により動的にローカルアライメントを作成し、その結果を図示している。同じ数字は、lalign でアライメントされた同じブロックに相当する。また、match による TRANSFAC による転写因子結合部位の予測結果も示している。TSS 付近の詳細を見るために、B の画面が用意されている。また、各部分の lalign の結果を C で詳細に見ることができる。

マウス 232,070 クローンあった。Sim4 により、ゲノムにマッピングを行った結果、ヒト 233,747 クローン、マウス 195,446 クローンがマッピングできた。ヒトの場合低クオリティのクローンを取り除くと、ヒト 190,964 クローンとなった。最終的には、ヒト 11,234 種マウス 7,524 種の RefSeq 遺伝子について 5'端配列をマップすることができた。この流れ図は、図 2-3 に示した。

RefSeq より 5‘端に伸ばすことができた配列は、ヒトでは 6,042 遺伝子、マウスでは 6,848 遺伝子あった。延びた長さは、ヒトの場合では mRNA レベルで見ると平均 71.6 塩基と短いが、DNA レベルで見るとイントロンを含むことがあるので、平均 4396 塩基と長くなっていた。

2.3.2　DBTSS 初期画面

DBTSS の初期画面を図 2-4 に示す。DBTSS データの検索手法は大きく 2 つに大別される。調べたい遺伝子がわかっている場合、その遺伝子の名前か RefSeq ID,を入れると検索できるようにした。その他、LocusLink ID, UniGene ID 等でも検索できるようにした。

もう一つの検索手段として未知配列を BLAST によって検索させる方法もある。BLAST の結果は、通常の検索結果と同様に表示されるが、これに DBTSS へのリンクを張っており、直接転写開始点情報を参照できるにした。

2.3.3　comparative genome 結果

2,048 種の遺伝子については、NCBI のデータベース homologene により、ヒトとマウスではオーソログ関係にあるとされている。これらに関しては転写開始領域の比較ゲノム解析を行えるようにした。図 2-5 はその例である。転写開始点から上流 1000 塩基、下流 200 塩基をゲノム配列から抜き出し、clustalW(Thompson et al., 1994)を使った大域アライメント、もしくは lalign(Huang and Miller, 1991)を用いた局所アライメントの結果を動的に表示することができる。

2.3.4 モチーフ検索

モチーフ検索は、TRANSFAC7.2 の公開データベースに対して行えるようになっている。モチーフ検索プログラムである MATCH、または正規表現を利用して、ある特定の転写因子結合部位をプロモータ領域に持つ遺伝子群を抽出することができる。

2.4　考察

DBTSS の特徴は、転写開始点を大規模に登録したデータベースである。既存のデータベースである EPD(Praz et al., 2002)は、2005 年 1 月の段階で登録されている遺伝子数が 1800 程度と少ない。これに対して DBTSS では、ヒト 11,234 遺伝子、マウス 7524 遺伝子に対して転写開始点の情報を付加することができた。これらのうちヒト 6,042 遺伝子、マウス 6,848 遺伝子を 5'側に伸長することができた。これらの遺伝子群に対しては、従来知られていたものよりもより正確な転写開始点情報を付加できた。

図 2-6　転写開始点の例

転写開始点の例を図示した。A では、全ての転写開始点がそろっている。B では、全ての転写開始点が、短い範囲でずれている。C では、JTH で始まる甲状腺由来のクローンとそれ以外のクローンとで、転写開始点が異なっている。

今回のマッピングに関しては、最初に 5’端配列を RefSeq cDNA 配列と対応を取り、直接ゲノムにマップしなかった。この理由は、大きく二つある。１つは、クローンがどの遺伝子に当たるのかマッピング前にクラスタリングしたかったからである。もう１つは、sim4 でのゲノムのマッピングは、記憶領域的にも時間的にも多量のリソースを消費するので、それらを短縮させるためである。あ

A kinase anchor protein 1

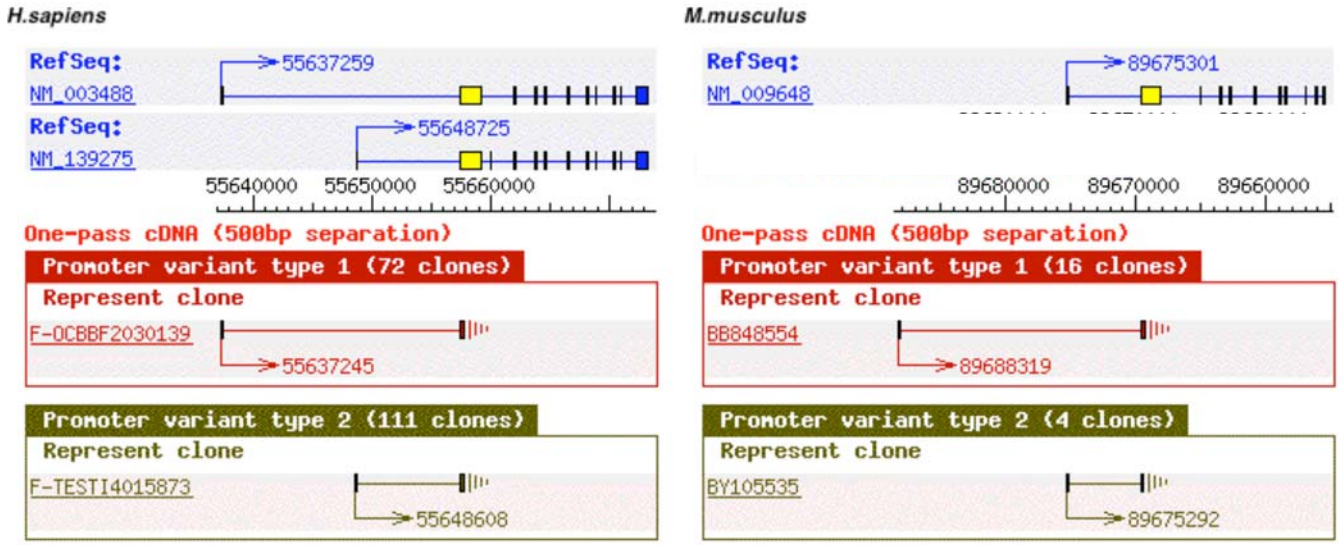


図 2-7　Alternative promoter が考えられる例

Human では、一つの遺伝子に対して二つの NM（NM_003488,NM_139275）が報告されている。この両者に対応する 5'端クローンが DBTSS 内にあった。しかし、mouse では NM_009648 の片方しか、報告されていないが、DBTSS では、NM_003488 に対応すると考えられるクローンが登録されていた。

らかじめ既知遺伝子に当てれば、その遺伝子がマップされているゲノム上の配列のみに探索を行うことが可能になる。この作業によって、リソースの節約だけでなく、偽遺伝子領域にマッピングされたり、5'端配列の一部が誤った領域にマッピングされるといったエラーも回避できたと考えている。なお、RefSeq cDNA と対応が取れなかった 5'端配列については、BLAT でマップし、大まかなゲノム領域を決めた後に、sim4 で正確に再マップした。そのマッピング結果が、一部でも RefSeq cDNA のエクソンとオーバーラップしている場合は、その 5'端配列をその cDNA 由来とした。

DBTSS の特徴の 1 つは、ヒト・マウスの転写開始点付近の比較ゲノム解析が容易であるということである。例えば、図 2-5 のように転写開始点付近で保存されている配列が容易に見いだせる。TRANSFAC にあるマトリックスを使った転写因子結合部位予測はよく行われるが、疑陽性が多いことが問題となる。これを避けるために、よく保存されている部分を選択的に選び出す手法を加えることで、より正確に転写因子結合部位を推定することができる。

図 2-6 で示した通り、転写開始点のパターンは遺伝子によって、1)転写開始点に揺らぎがない（図 2-6A）、2)一つのプロモータ内で転写開始点が揺らいでいる（図 2-6B）、3)複数のプロモータを持つ（図 2-6C）という大きく３つのパターンに分類できることがわかった。同じように、転写開始点が揺らいでいる上記 2), 3)で起こっている生物学的要因は大きく異なると考えられる。前者の場合、ORF は同じなので翻訳産物は変わらない。したがって、こちらは一つのプロモータ内での揺らぎであると考えられる。これは、RNA polymerase の転写開始の曖昧性にあるかもしれない。しかし後者の場合は、図 2-6C の様に甲状腺由来（JTH で始まる ID）5'端配列と、それ以外の配列という異なる転写開始点を持つだけでなく、alternative splicing form を示し ORF も異なる。後者の場合は、より積極的な制御システム、つまり alternative promoter(Ayoubi and Van De Ven, 1996; Landry et al., 2003)による転写制御システムが甲状腺で働いていると推測できる。図 2-7 に、alternative promoter の典型例と考えられる例を示した。A kinase anchor protein 1 は、ヒトでは２つの遺伝子(NM_003488,NM129275)が同じ領域にマッピングされている。ところがマウスでは、短い方の一つ(NM_009648)しか報告されていない。しかし DBTSS では、マウスにも BB848554 等の NM_003488 に対応すると考えられる遺伝子が観察されている。このように、DBTSS は、alternative promoter を発見するのに非常に有効である。転写開始点の揺らぎに関しての詳細な解析は、次章に述べる。

3. 転写開始点付近の基本的な特徴について解析

3.1 序論

転写は、ゲノムの転写開始点付近にあるプロモータ上に、RNA polymerase や基本転写因子が集合して開始複合体を形成することによって開始される。プロモータ部分とは別に、エンハンサーやサプレッサーと呼ばれる配列が DNA 上にあり、開始複合体と作用する様々な転写因子が結合することにより mRNA の転写制御が行われている。

Molcular Biology of The Cell 第４版によると、プロモータ部分には、転写開始点から-40 塩基付近に TFIIB recognition element(BRE)、-35 塩基付近に TATA-box、転写開始点付近に Initiator(Inr)、+30 塩基付近に Downstream promoter element(DPE)という４つの転写因子結合部位がある。全ての遺伝子がこれらの配列を持つわけではないが、持つ遺伝子同士では、結合部位がコンセンサス配列として保存されている。近年、ヒト 1031 遺伝子の転写開始点を利用した解析の結果、TATA は 32%、initiator は 85%の遺伝子にあることが報告されている。また、驚くべきことに、転写開始点は従来考えられていたほど単一ではなく、複数の転写開点から成立していることが示唆されている(Suzuki et al., 2001b)。しかし、この報告以外に、どの遺伝子がどのコンセンサス配列を持つのか、またどのコンセンサス配列が保存されているのかについての大規模な解析は行われていない。これは、DBTSS 構築以前は、転写開始点の大規模なデータの入手が困難だったからである。

以降第３章では、ヒトの DBTSS の転写開始点情報を用いて、転写開始領域の基本的な解析を試みた。まず、一つの転写開始点に二つ以上の 5'端配列がある転写開始点は信頼性が高いとして、そのようなものを DBTSS から抜き出した。その結果 4,863 遺伝子、17,426 転写開始点情報を抽出した。このデータセットを用いて、まず転写開始点の揺らぎが、どのくらいの遺伝子でどの程度見られるのか調べた。続いて、プロモータに存在していると考えられている上述の４つの転写因子結合部位(BRE、TATA-box、Inr、DPE)の探索を行った。さらに、ヒトプロモータにおいては、CG に富む領域すなわち CpG islands があるプロモータとないプロモータが存在することが知られている(Gardiner-Garden and M., 1987)ので、この有無も調べた。

本研究で行われた解析は、正確な転写開始点に基づくヒトの転写開始領域の大規模な解析ということでは初めての報告である。

3.2 材料と方法

3.2.1 データセット

データセットは、ヒトの DBTSS ver.2（9336 遺伝子：183,8455　5’端配列）を用いた。転写開始点情報がはっきりわかっているものを使用するため、ゲノムの複数箇所にマッピングされているクローンは除去した。さらに、転写開始点の精度を上げるために、１つの転写ユニットに 2 つ以上クローンが存在しているもののみを抜き出した。これは、17,426 転写開始点(4,863 遺伝子)に相当した。

3.2.2 転写開始点揺らぎの定義

一つの遺伝子の転写開始点の揺らぎ(valiation)は、以下のように、標準偏差を用いて定義した。

$$variation = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - ave\right)^2 \qquad \text{(式 3-1)}$$

ただし、*Xi*: 5’端配列のゲノム上の位置、n: 一つの遺伝子で観察された 5’端配列数、*ave*: 一つの遺伝子で観察された 5’端配列のゲノム上の位置の平均値である。これを、17,426 転写開始点(4,863 遺伝子)について行った。

3.2.3 クラスター解析

転写開始点に揺らぎがあることは、２章で述べた。このような揺らぎのある遺伝子に対して転写開始点の解析を行うのは、どの転写開始点に着目すればよいのか判断が難しい。したがって、最初に、転写開始点に揺らぎが見られない遺伝子群のプロモータのみを解析対象とすることにした。まず、4,863 遺伝子セットから、一遺伝子あたり 10 以上の 5’端配列があるものを抽出した。さらに、5’端配列の 50%以上が一つの転写開始点から始まっている場合、その遺伝子を揺らぎのない遺伝子と定義した。これは 334 転写開始点(334 遺伝子)であった。

抽出した 334 種の TSS のうち-3~+5 領域を抜き出して 8 塩基の配列セットを用意した。この 334 配列の距離をミスマッチの割合で定義した。この距離を Clustan Graphics ver.5(Clustan Ltd.) を用いて Unweighted pair-group method using arithmetic averages(UPGMA 法)によって 表現した。

3.2.4 転写因子結合部位予測

転写因子結合部位予測には、334 遺伝子の-500～＋200 領域を抜き出した配列を用いた。この配列に対して、転写因子の位置特異的スコア行列（position specific score matirix：PSSM）を利用して、式 3-2 を用いてスコアを計算した。

$$score = \sum_{i=1}^{l} \log \frac{(n_{if} + 1)}{(N_i + 4)} / \frac{1}{4} \qquad \text{(式 3-2)}$$

ただし、l:スコア行列の長さ、n_{if}：i 番目の塩基 n の数、N_i：i 番目の塩基数合計とする。スコア行列の長さの window size で、-500～＋200 領域の全てに対して一塩基ずつずらしながらスコアを求めた。TATA-box と Initiator の検出には TRANSFAC(Wingender et al., 2000)の V$_TATA_01 と V$_CAP_01 をそれぞれ用いた。BRE に関しては、Lagrange (Lagrange et al., 1998) らによって報告されているものをスコア行列とした。各スコア行列の長さをヒトゲノムから 100 万配列ランダムに切り出してスコアを計算し、上位 5%以上となるスコアを敷居値として、それ以上の値であったとき転写因子結合部位が陽性であると判断した。

Downstream promoter element(DPE)に関しては、スコア行列が作られていない。したがって、Burke らの報告(Burke and Kadonaga, 1996)に従って G(A/T)CG という正規表現による検索を行った

3.2.5 CpG island の検出

転写開始点から-100~+100 塩基を取り出して、データーセットとした。CpG island の定義は Gardinar-Garden ら(Gardiner-Garden and M., 1987)の定義に従った。すなわち、GC 含量と CpG score を

GC 含量 = (G+C/200)

CpG score = (CG/G*C)*200

で定義し、GC 含量 >= 0.5 かつ CpG score >= 0.6 ならば CpG ありと判定した。

3.3 結果

3.3.1 TSS の分布

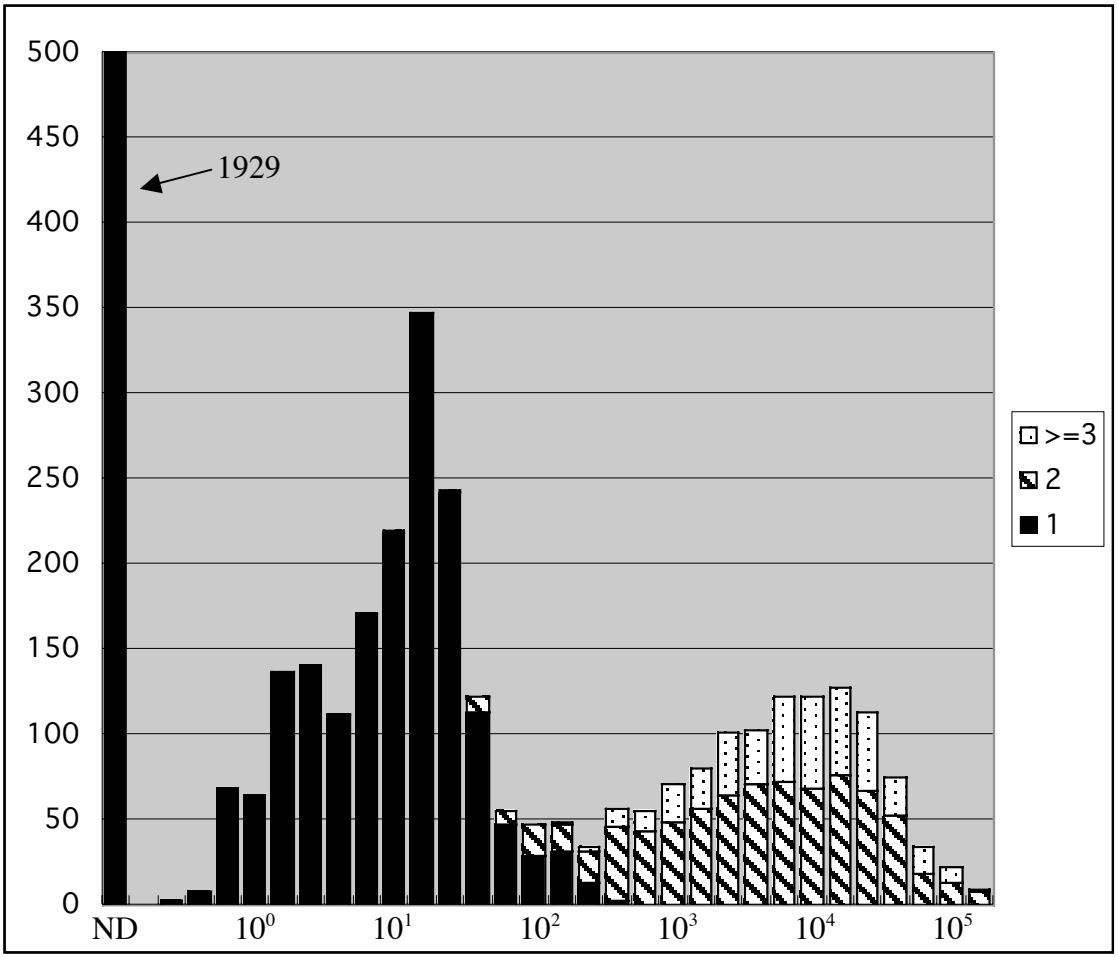


図 3-1　一遺伝子の転写開始点距離分布

4,863 遺伝子（17,426 転写開始点）の標準偏差とその頻度を示した。横軸は転写開始点の標準偏差の対数値、縦軸はその標準偏差をもつ遺伝子数を示した。棒グラフの模様は、遺伝子内で観察された第一エクソンが、いくつであったかを示している。

4,863 遺伝子(174,26 転写開始点)において、転写開始点の分布を標準偏差で表した（図 3-1）。1,929（39.7%）の遺伝子は、全て同じ転写開始点から始まっていた。しかし、残りの遺伝子には、転写開始点の揺らぎが見られ、標準偏差約 10^2 塩基を境界に二つの分布が認められた。最初のピークは 1,761 遺伝子を含み標準偏差 10^2 塩基以下である。それに対して次のピークは標準偏差 10^2 塩基以上であり、1,173（24.1%）の遺伝子群を含む。図 3-1 では、オーバーラップしていない第一エキソン数も同時に示した。1,761 遺伝子の分布の転写開始点は、ほとんど同じ第一エキソンを共有している。しかし、第二の分布では、ほとんどが 2 つ以上の第一エキソンから構成されている。したがって、第二の分布に

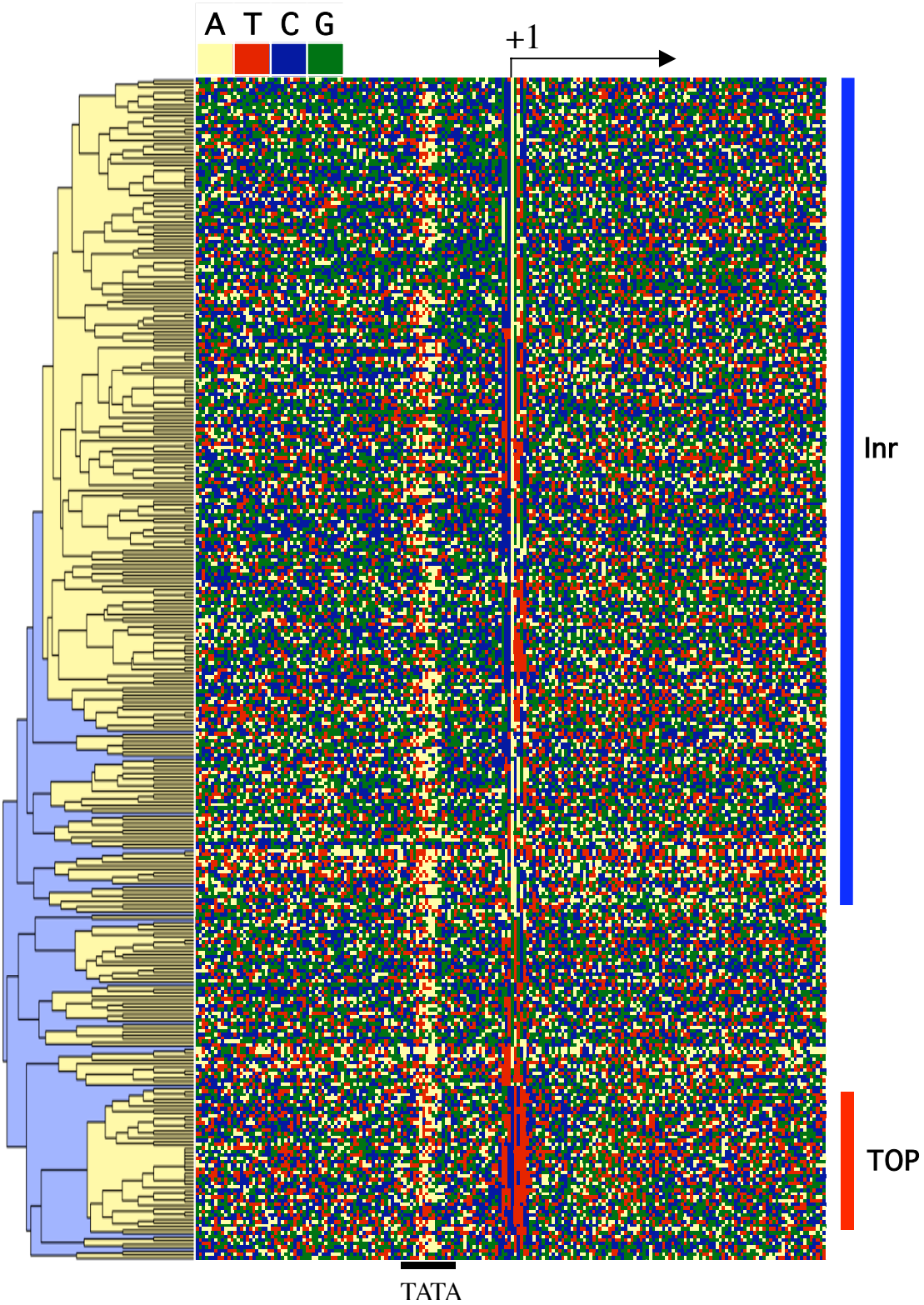


図 3-2　–100~+100の塩基情報

転写開始点のそろっている 334遺伝子の TSSを ATCGの塩基ごとに色を付けて図示した。-4~+7までを UPGMA法によるクラスター解析の結果順に並べた。Initiator(Inr)と Teminal oligo pyrimidine（TOP）を持つ遺伝子部分を示した。

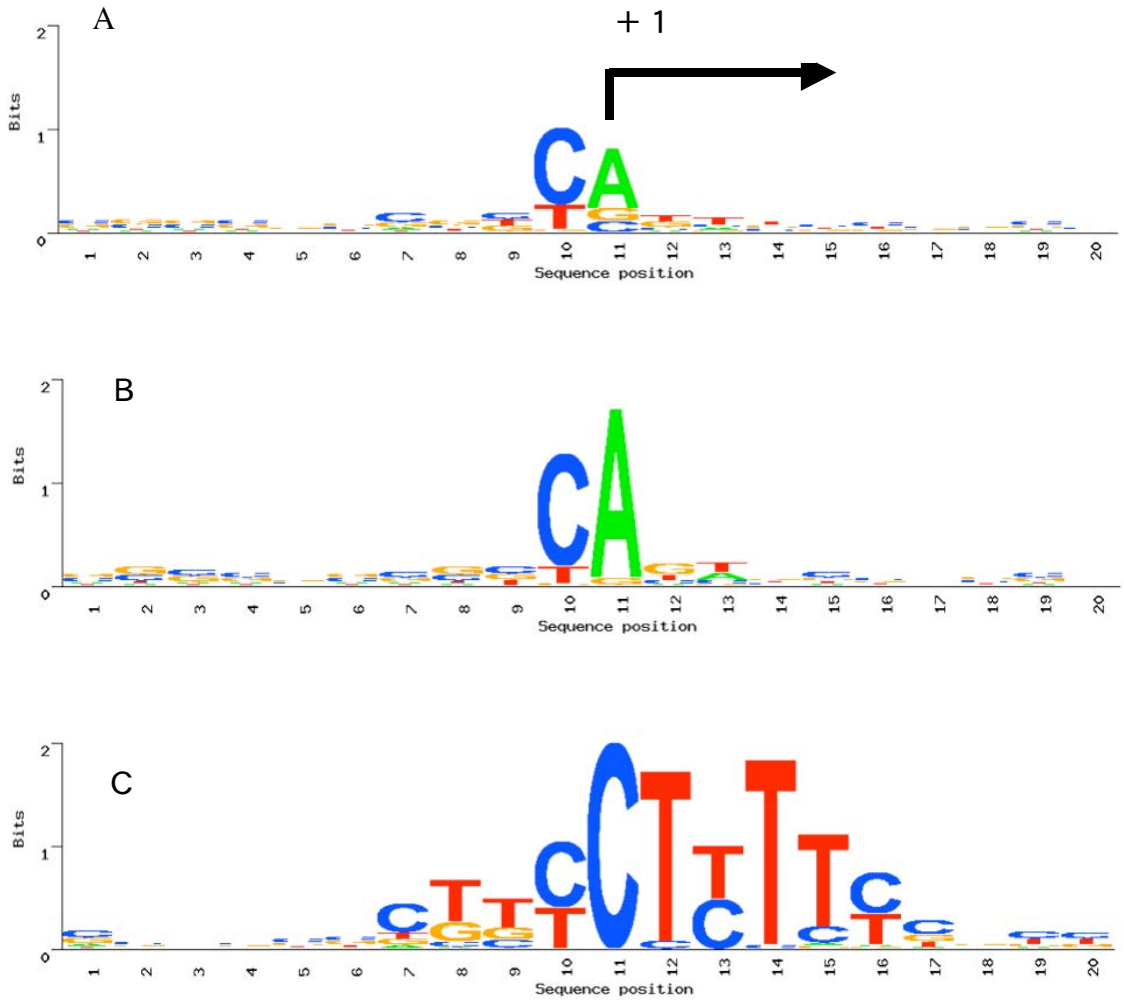


図 3-3　-10~+10 の塩基組成

-10~+10 までの配列を sequence logo で表示した。転写開始点がそろっている 334 遺伝子全て（A)、図 3-2 のクラスターB236 遺伝子(B) 、クラスターC42 遺伝子(C)に対応する。

含まれる遺伝子群は alternative promoter を構成する可能性が考えられる。

3.3.2 転写開始点のクラスター解析

1 つの遺伝子に 5’端配列が 10 個以上ありかつそのうち 50％以上が 1 つの転写開始点によって占められている遺伝子群 334 個を選び出した。これらの遺伝子の-100~+100 をとってきて表記したのが図 3-2 である。Sequence logo(Schneider and Stephens, 1990)で 334 遺伝子の-10~+10 領域を図示したものが図 3-3 である。ここで、-3~+5 の領域が比較的よく保存されていることが見て取れたので、この領域を取り出し類似度に基づいてクラスター解析を行った。その結果 2 つの

顕著なクラスターが認められた。前者は 236（70％）の遺伝子からなり、sequence logo で表示（図 3-3B）させると「CANT」という普遍的転写因子 TFIID の結合部位である Initiator(Kraus et al., 1996)に似た配列を持つことがわかる。後者は 42 個（13％）の遺伝子からなり、翻訳時の制御領域として知られる 5’-terminal oligopyrimidine tract(TOP)(Loreni and Amaldi, 1997)を持つ。この遺伝子群の特徴として、全ての遺伝子が-3～+5 の領域に 60％以上のピリミジンをもち、TT(C/T)|CT(C/T)TT（ただし｜は TSS を意味する）というパターンを持つ（図 3-3C)。その内訳は、ribosomal protein33 遺伝子、翻訳伸長因子 1、翻訳開始因子 3、その他 TOP として報告されている 2 遺伝子(LAMR1 と TPT1)が含まれている。特に ribosomal protein に関しては、334 の遺伝子セットに含まれる 34 遺伝子のうち 33 遺伝子が含まれることがわかった。

3.3.3 基本転写因子結合部位の探索

図 3-2 の転写開始点付近の塩基組成の図から、TATA-box が-30 付近に固まって存在していることが見て取れる。どのくらいの遺伝子が TATA-box を持つのか調べるために TRANSFAC(Matys et al., 2003; Wingender et al., 2000)の中に含まれている位置特異的スコア行列を用い、式 3-2 により計算を行った。敷居値を設定するために、ゲノムからランダムに 100 万配列を抜き出してスコアを計算した。この結果、-0.087 より上にランダムサンプリングの上位 5%が入ったので、これを敷居値に設定した。図 3-4 では、転写開始点に揺らぎのない 334 遺伝子の-1000~+200 領域のスコアの分布を示している。図 3-4A の TATA-box の検出では-36~+28 の領域に鋭いピークが認められた。さらに、残りの転写開始点群にも TATA があるかどうかを調べるために 17,476 の転写開始点に対して-36~+28 のみを調べた。3,724 転写開始点に対して TATA が認められ、遺伝子数では約 1/3 の 1649 個に相当した（図 3-5)。

その他、基本的転写因子結合部位として知られている、BRE、Inr、DPE の検出も試みた。揺らぎのない 334 遺伝子のうち 55％が Inr を持っていた。しかし、BRE と DPE の存在は顕著に認められなかった。BRE は TATA-box のすぐ上流にあると Lagrange らは報告している。しかし、BRE があるとされている-44~-34 の領域では、757 の陽性がみつかったが、この部分を取り除いた他の領域でも 12,201 個検出された（図 3-4B)。DPE に関しては、＋30 に存在する(Burke and Kadonaga, 1996)とされているので、前後 2 塩基ずつ幅を取り、+28~+32 領域に

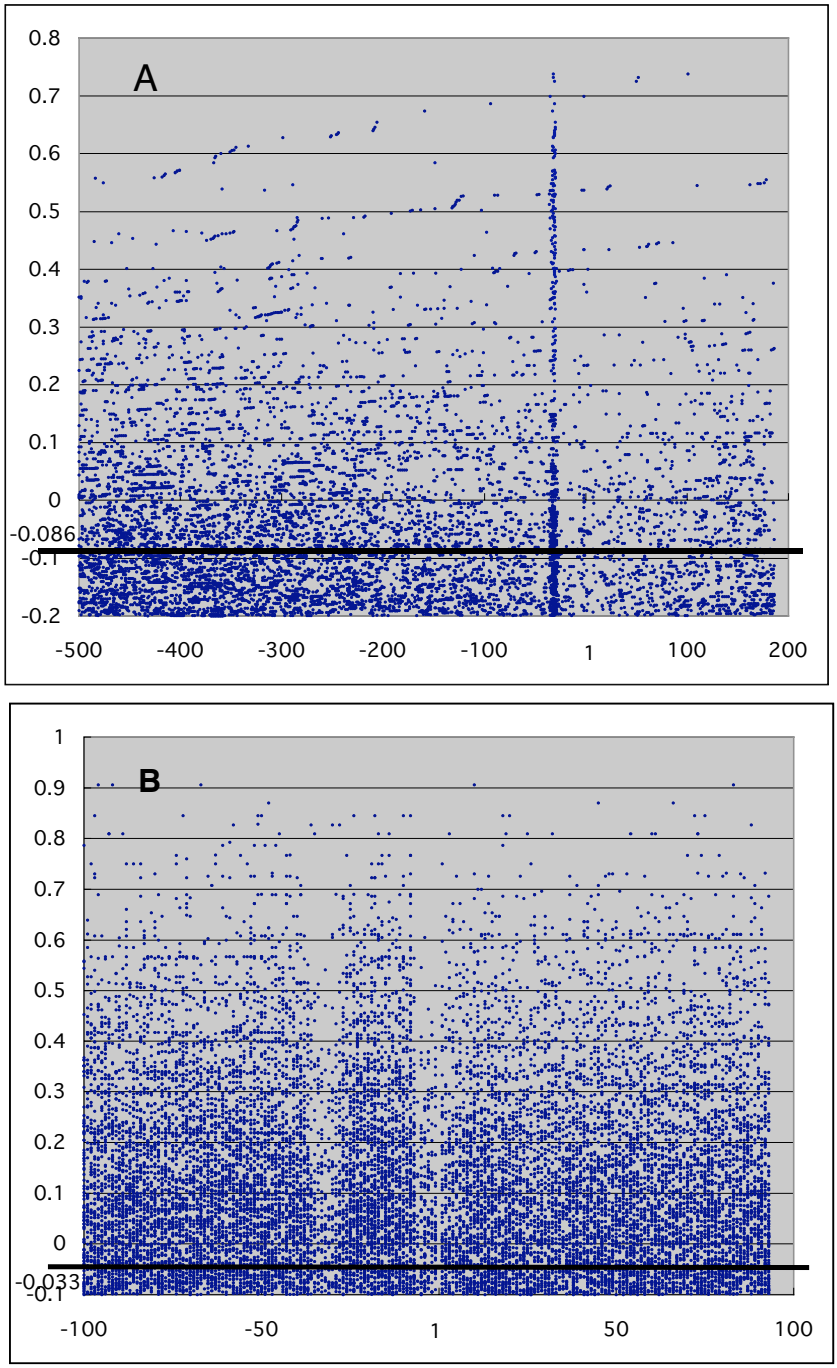


図 3-4　PSWM 検索

334TSS の PSWM 検索の結果のうち、score>0 のみ表示した。それぞれに設定した threshold を図の中に示した。A:TATA motif(500~200)、B:BRE motif(-100~100)。

対して検索した。その結果、334 遺伝子中で検出されたのは 16 個しかなかった。

最後にプロモータ領域(-100~+100)に CpG islands があるかをを見た。転写開

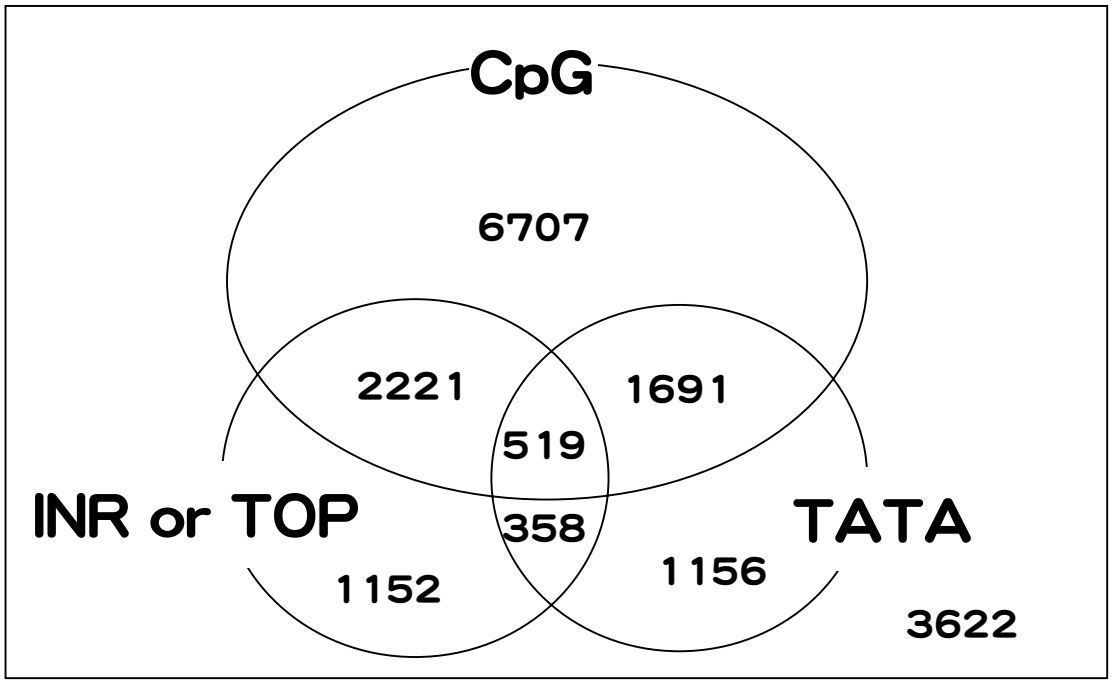


図 3-5　17,426 転写開始点のベン図解析

データセット全てである 17,426 転写開始点（4,863 遺伝子）がプロモータ上に持つ配列を示した。CpG は CpG isllnads, TATA は TATA-box, INR は Intiator,TOP は Terminal ologo pyrimidine 配列を意味する。上記の配列を持たなかった転写開始点は 3,622 あった。

始点に揺らぎのない 334 遺伝子のうち CpG islands 陽性と考えられたのは 245 遺伝子（73.4%）であった。また、データセット全体の 17,426 の遺伝子セットでは、11,138（63.9%）の転写開始点に CpG islands が認められ、これは 3696 遺伝子(76.0%)が少なくても１つの CpG islands をプロモータ領域に持つことになる（図 3-5）。

3.4　考察

転写開始点は、単一ではなく複数個ある可能性があることは、Suzuki らによって示されていた(Suzuki et al., 2001a)。DBTSS でも図 2-6 で示したように、遺伝子ごとに転写開始点の分布の違いがあることがわかる。しかしながら、この現象が、oligo-capping clone 作成時のアーティファクトによって起こっている可能性を否定できない。本研究では、183,845 の 5’端配列という巨大なデータを用いて、一つの転写開始点に２つ以上の 5’端配列のサポートがある転写開始点のみを抜き出した。これは、104,226clone から構成される 17,426 転写開始点であり、それは 4,863 遺伝子に相当する。もちろん PCR の副産物というものもあ

るだろうが、それ以外の要因で偽の転写開始点に複数のクローンがマッピングされることは少ないと考えられるので、このデータセットの信頼性は、非常に高いことが考えられる。このデータセットに置いても一遺伝子内での転写開始点の揺らぎが観察されたことは、この現象が、*in vivo* で普通に起こっていることが示唆される。図 3-5 で示したように、従来言われていた TOP や TATA、Inr といった配列が抽出できたことも、このデータセットの信頼性の高さを裏付ける。なお、TOP 遺伝子の詳細な解析は次章で述べる。

TSS の標準偏差の分布から、遺伝子の転写開始点は 3 つに大別できることが考えられる。すなわち、まったく揺らぎのない 1,929 遺伝子、揺らぎの標準偏差が 100 塩基以内の 1,761 遺伝子群、揺らぎが、100 塩基を越える 1,173 遺伝子である(図 3-1)。三つ目のグループは、基準が異なるため Suzuki らによって報告されていない。注意しなくてはならないのは、x 軸にログスケールをもちいていることがある。つまり、2 つ目のクラスターに比較して、3 つ目のクラスターは揺らぎの分布が非常に大きい。さらに一番目のクラスターに属する遺伝子には、転写開始点を表現するクローンが少数である可能性も考えられる。しかし、この図 3-1 から、少なくとも 2/3 の遺伝子は、転写開始点に揺らぎを持つことが示唆される。この揺らぎが生物的な意義をもつのか不明であるが、いくつかの可能性が考えられる。まず、最初の揺らぎの少ない遺伝子群であるが、これらの mRNA は、同じエキソンから始まっているので一つのプロモータ内で同じような発現制御機構によって起こっていると考えられる。これに対し、第三の揺らぎが大きい遺伝子群は、第一エキソンを共有していない場合が多いので、翻訳産物も異なる可能性が強く考えられる。このような遺伝子群では、alternative promoter(Ayoubi and Van De Ven, 1996)によって異なる転写制御のメカニズムによが行われている可能性がある。図 3-1 から、4,863 遺伝子のうち約 1/4 の 1,173 遺伝は alternative promoter を持つことが示唆される。

334 の遺伝子セットには、TBP 結合部位である TATA-box が-34~-28 の領域に観察された。これは、TATA-box binding protein(TBP)が転写開始点を決めているという報告(Schmidt et al., 1989)と矛盾しない。Suzuki らの報告によると、TATA-box が、揺らぎの無い遺伝子群で多く観察された(Suzuki et al., 2001b)。今回の解析で転写開始点の最初のデータセットの 334 遺伝子群では,80%に TATA-box が認められたが、4,863 遺伝子群では 34％(1,649)にしか認められなかった。この差は、最初の 334 遺伝子のデータセットには、50％以上の遺伝子が同じ位

置から始まっているという条件を付加したため、TATA を持つ遺伝子をより選択的に抽出したためであると考えられる。その他の、基本的転写因子結合部位とされているものでは、Inr は 334 遺伝子で 55%に認められ、17,426 転写開始点（4,,863 遺伝子）全体でも 53%(2,578)の遺伝子群に認められた。しかし、BRE や DPE に関しては、はっきり抽出ができなかった。これは、BRE や DPE が従来考えられていたほど、たくさんの遺伝子に含まれていないか、転写開始点からの距離が一定で無いためであると考えられる。また CpG island の検出も試み、4,863 遺伝子のうち 76%が少なくても 1 つの転写開始点に CpG island を持つことがわかった。これらをまとめると、図 3-5 のようになりほとんどの promoter と呼ばれる領域には、TATA, Inr or TOP, CpG island のいずれかが存在していることがわかった。

従来まで、転写開始点情報のわかっている遺伝子は非常に少なく、プロモータ領域の解析に大きな支障をきたしてきた。本研究で用いたデータセットから、core promoter 領域に関して有用な情報を得ることができた。今後のプロモータ領域の研究にはこのような大規模なデータセットが非常に有用であろう。

4　ヒト TOP 遺伝子の網羅的探索

4.1　序論

TOP 遺伝子は、5'端に terminal oligo-pyrimidine tract を持つ遺伝子群の総称である(Meyuhas et al., 1996; Meyuhas, 2000b)。TOP 遺伝子としては、リボソームタンパク質群(Meyuhas, 1996)、翻訳伸長因子(EEFA1、EEFA2)、poly(A)-binding protein(PABP)(Meyuhas, 2000a; Meyuhas, 2000b)、nucleophosmin(NPM1)(Zatsepina et al., 1999)、laminin receptor 1 (LAMR1)(Ford et al., 1999; Tohgo et al., 1994)等、リボソームの構築あるいは翻訳に関わるものがよく知られている。また、pre-mRNA のスプライシングと RNA 輸送に関わる heterogeneous nuclear ribonucleoprotein A1 (HNRNPA1)(Dreyfuss et al., 1993; Izaurralde et al., 1997)、機能が明らかではないが tumor protein translationally controlled 1 (TPT1)(Gachet et al., 1999)も TOP 遺伝子としてあげられている。

TOP 遺伝子の最大の特徴は、翻訳レベルでタンパク質合成の制御を受け、TOP 構造がシスエレメントとして働いていることである。培養細胞において栄養制限をかけると、TOP 遺伝子は polysome から mRNP 顆粒（subpolysomal fraction）に移行して翻訳が抑制される。もし、この TOP 配列を C→A に変えた時には、この現象は起きなくなる。TOP 遺伝子以外では、このような切り換えは観察されず、栄養制限をかけても翻訳は継続される(Meyuhas, 2000b)。

TOP 遺伝子研究に関しては２つの大きな課題が残されている。一つ目は、TOP 遺伝子が受ける翻訳制御パスウェイである。現在、栄養制限等のシグナルが受容体の phosphatidylinositol 3-kinase(PI3-K)から(3-phophoinositide-dependent kinase 1(PDK1)を経由して(protein kinase B)PKB に伝わり、翻訳制御に関わる可能性が示唆されている。しかし、実際に TOP の構造を認識する物質が何なのか、そしてその物質がどのように TOP 遺伝子を制御しているのかについては、全くわかっていない(Stolovich et al., 2002)。もう一つの課題は、どの遺伝子が TOP 遺伝子であるかと言うことである。近年リボソームタンパク質について網羅的に 5'端配列を決定した結果、全てに poly-pyrimidine tract があることが示唆された(Yoshihama et al., 2002)。しかし、それ以外に遺伝子全体を網羅的に調べた報告はない。

TOP 遺伝子の検出に関しては、3 章でも述べた。しかし、3 章の解析では、典型的な TOP 遺伝子とされている遺伝子群がいくつか検出されていない。例えば eukaryotic translation elongation factor 1 alpha 1(EEF1A1)が検出されなかった。

DBTSS では、EEF1A1 に相当する RefSeq:NM_001402 がゲノム上で染色体 6 番と 9 番の 2 ヶ所にマッピングされている。このような場合は、どちらの領域から転写されたのか曖昧になるために、3 章の解析では利用しなかった。しかし、両方の領域を目で見た場合、明らかに EEF1A1 は TOP 遺伝子であった。また、HNRNP1 も検出できなかった。これは、DBTSS ver2 においては HNRNP1 に相当するクローンは 1 つしかなく、2 つ以上のクローンをもつ転写開始点のみを取り扱った 2 章では、取りこぼしてしまった。こちらも TOP 遺伝子の条件は満たしていた。

以上のように、2 章の方法では、偽陰性が明らかにあり、TOP 遺伝子の探索としてはやや不適切であった。本章では、TOP 遺伝子の抽出に関していくつかの改善方法を加え、偽陰性をできるだけ減らすように抽出することを目指した。また、マウスでも同様の探索を行い、オーソログ遺伝子同士で比較した。

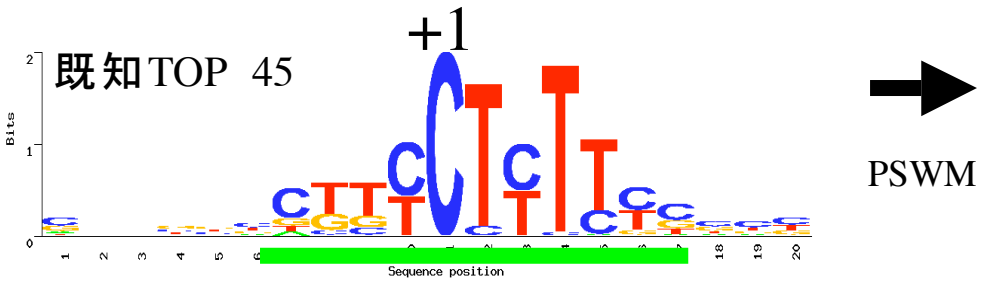
4.2 材料と方法

4.2.1 転写開始点のデータセット

転写開始点のデータは DBTSS ver.3 を用いた。3 章で使った解析とは異なり、ヒトの全ての転写開始点情報 113,875(11,234 遺伝子に由来)を解析対象とした。Ver.3 では、DBTSS の転写開始点情報を、RefSeq 遺伝子ではなく NCBI の LocusLink を単位にしている。これは、ゲノム上の一つの領域(locus)に複数の RefSeq cDNA がスプライシングバリアントとして登録されているので、複数の遺伝子に一つの 5'端配列に基づく転写開始点が与えられている場合がある。このような場合、転写開始点をそのゲノム上の領域での転写開始点として見なすべきであり、LocusLink が有用なためである。ヒトの 11,234 遺伝子は、9,470 loci に相当する。また、ゲノム上に複数箇所にマッピングされている遺伝子に関しては、全てを探索し、一つでも条件に当てはまる転写開始点があれば陽性とした。マウスに関しては、88,078 TSS(7,524 遺伝: 6875 loci)のデータを用いた。ヒトとマウスのオーソログ情報は NCBI の homologene により 2.3.3 で作成したものを用いた。

4.2.2 TOP 遺伝子の位置特異的重み行列作成、及びスコア計算

3.2.3 と同様に、ヒトの 9,470 loci (113,875 転写開始点)から、一つの遺伝子に 5'端配列が 10 以上存在し、かつそのうち 50%以上の転写開始点が同じであ



PO	A	C	G	T
-4	4	29	7	5
-3	1	4	14	26
-2	1	6	10	28
-1	0	26	0	19
1	0	45	0	0
2	0	3	0	42
3	0	23	0	22
4	0	1	0	44
5	1	11	0	33
6	1	21	4	19
7	2	23	12	8

図 4-1　既知 TOP 遺伝子から構築した PSWM

DBTSS ver3 より選び出した、転写開始点がそろっている遺伝子から既知の 45TOP 遺伝子（EEF2,LAMR1,TPT1,ribosomal protein 42 種）を選び出し、PSWM を作成した。

る遺伝子を 404 選び出した。この中から既知の TOP 遺伝子 45 種（表 4-1）を選び出して、良く保存されている-4~+7 に対して位置特異的重み行列を作成した（図 4-1）。さらに、113,875 転写開始点から-4~+7 の配列をゲノムから抜き出して、この行列と式 3-2 によってスコアを求めた。

4.3　結果

4.3.1 404 遺伝子からの抽出

DBTSS ver.3 の転写開始点のデータセットには、9,470 loci の転写開始点 113,875 が含まれている。この中から一つの locus 当たり 10 個以上のクローンを持ち、そのうち半分以上の転写開始点がそろっているもの 404 個（404 loci）を選び出した。さらにこの中から、確実に TOP 遺伝子だと考えられている 45 遺伝子（42 リボソームタンパク質、EEFA2、TPT1、LAMR1）を抜き出し、それより得られる TOP 配列のコンセンサス配列を sequence logo で表示した（図 4-1）。このとき、-4~+7 の領域がよく保存されていたので、この領域から、位置特異的重み行列を作成した。その行列と式 3-2 を用いて 404 転写開始点を検索すると、重み行列のスコア ＞ 0.062 で全ての既知 45 TOP 遺伝子が抽出できた(表 4-1)。従ってスコアのの閾値を 0.06 と設定した。さらに、45 遺伝子全てで観察されたものを TOP 遺伝子として抽出する条件に加えた。つまり、

1. 重み行列のスコア ＞0.06
2. +1 は必ず C

表 4-1　既知 TOP 遺伝子

DBTSS ver3 より選び出した、既知の 45TOP 遺伝子。これから作成した位置特異的重み付きスコア行列による再建策の結果は、PSWM score の列に示した。敷居値とした値 0.062 は、赤字で示してある。

NM id	chr	strand	TSS	#TSS clones	#all clones	m4p7	definition	PSWM score
NM_001961	chr19	−	4054739	97	157	CGTTCTCTTCC	eukaryotic translation elongation factor 2 (EEF2), mRNA	0.844
NM_002295	chr3	+	38681430	32	60	CTGTCTTTTCC	laminin receptor 1 (ribosomal protein SA, 67kDa) (LAMR1), mRNA	0.805
NM_006013	chrX	+	150007389	67	116	GCGCCTCTTTC	ribosomal protein L10 (RPL10), mRNA	0.554
NM_007104	chr6	+	35432595	50	81	TAGTCTCTTTT	ribosomal protein L10a (RPL10A), mRNA	0.328
NM_000975	chr1	+	23089282	50	60	CTTTCTCTTCC	ribosomal protein L11 (RPL11), mRNA	0.897
NM_000976	chr9	−	121860223	101	136	CGGCCTCTCGG	ribosomal protein L12 (RPL12), mRNA	0.498
NM_002948	chr3	+	23641155	96	184	CTTTCCTTTCC	ribosomal protein L15 (RPL15), mRNA	0.677
NM_000985	chr18	−	46818521	84	127	CTTCCTCTTTC	ribosomal protein L17 (RPL17), mRNA	0.916
NM_000979	chr19	−	49490557	36	62	CTCTCTTTCCG	ribosomal protein L18 (RPL18), mRNA	0.614
NM_000982	chr7	+	19686691	53	88	TTGCCCTTTCG	ribosomal protein L21 (RPL21), mRNA	0.415
NM_000983	chr1	−	6068571	27	51	CTCCCTTTCTA	ribosomal protein L22 (RPL22), mRNA	0.499
NM_000978	chr17	−	39007795	33	64	CTTCCTTTTTT	ribosomal protein L23 (RPL23), mRNA	0.823
NM_000987	chr17	−	9380108	293	370	AGTTCTCTTCC	ribosomal protein L26 (RPL26), mRNA	0.681
NM_000990	chr11	+	9150892	72	97	CTTCCTTTTTC	ribosomal protein L27a (RPL27A), mRNA	0.912
NM_000967	chr22	−	36330158	288	356	CGGCCTCTACC	ribosomal protein L3 (RPL3), mRNA	0.525
NM_000993	chr2	+	100071483	72	112	CTTCCTTTCCA	ribosomal protein L31 (RPL31), mRNA	0.637
NM_000995	chr4	+	109862813	19	29	CTTCCTCTTCC	ribosomal protein L34 (RPL34), transcript variant 1, mRNA	0.924
NM_015414	chr19	+	5758909	12	23	AGCCCTTCCGC	ribosomal protein L36 (RPL36), transcript variant 2, mRNA	0.063
NM_021029	chrX	+	97617140	21	40	CTTTCTTTCCG	ribosomal protein L36A (RPL36A), mRNA	0.743
NM_000997	chr5	−	41927498	48	60	CGGTCTTTCTG	ribosomal protein L37 (RPL37), mRNA	0.593
NM_000968	chr15	−	59896199	443	675	CTTCCTTTTCC	ribosomal protein L4 (RPL4), mRNA	0.921
NM_021104	chr12	−	56872911	55	89	CTTTCTCTCGG	ribosomal protein L41 (RPL41), mRNA	0.612
NM_000971	chr8	−	74146410	30	54	CTTCCTCTTTT	ribosomal protein L7 (RPL7), mRNA	0.827
NM_001015	chr19	+	50367828	33	63	CTTTCTTTTTT	ribosomal protein S11 (RPS11), mRNA	0.796
NM_001016	chr6	+	132982826	55	76	AGGCCTCTTTC	ribosomal protein S12 (RPS12), mRNA	0.611
NM_001017	chr11	−	18042975	19	36	CTCTCCTTTCG	ribosomal protein S13 (RPS13), mRNA	0.492
NM_001018	chr19	+	1507698	16	30	CGATCTCTTCT	ribosomal protein S15 (RPS15), mRNA	0.512
NM_001019	chr16	−	18231920	81	101	CGTCCTCTTTC	ribosomal protein S15a (RPS15A), mRNA	0.862
NM_001020	chr19	−	40317744	79	117	TTTCCTTTTCC	ribosomal protein S16 (RPS16), mRNA	0.774
NM_001021	chr15	−	76142596	29	50	TTTCCTCTTTT	ribosomal protein S17 (RPS17), mRNA	0.680
NM_001023	chr8	−	56926068	103	170	CTTTCTTTTTG	ribosomal protein S20 (RPS20), mRNA	0.829
NM_001025	chr5	−	81801096	63	111	TTCTCTCTTTC	ribosomal protein S23 (RPS23), mRNA	0.613
NM_001026	chr10	+	79003250	61	91	GGTTCTCTTTT	ribosomal protein S24 (RPS24), transcript variant 2, mRNA	0.626
NM_001028	chr11	−	120400848	98	124	CTTCCTTTTTG	ribosomal protein S25 (RPS25), mRNA	0.856
NM_001030	chr1	+	149694319	32	56	GCTCCTTTCCG	ribosomal protein S27 (metallopanstimulin 1) (RPS27), mRNA	0.497
NM_002954	chr2	+	55646009	38	52	CTTCCTTTTCG	ribosomal protein S27a (RPS27A), mRNA	0.865
NM_001031	chr19	+	8476429	19	26	ACTCCTCTCCG	ribosomal protein S28 (RPS28), mRNA	0.458
NM_001032	chr14	−	43849063	47	79	CTTCCTTTTAC	ribosomal protein S29 (RPS29), mRNA	0.703
NM_001007	chrX	−	68731855	77	153	GGTCCTCTTTC	ribosomal protein S4, X-linked (RPS4X), mRNA	0.742
NM_001010	chr9	−	19569196	75	125	GGCCCTCTTTT	ribosomal protein S6 (RPS6), mRNA	0.524
NM_001011	chr2	+	3265224	68	98	GGGTCTCTTCC	ribosomal protein S7 (RPS7), mRNA	0.635
NM_001012	chr1	+	44241669	143	190	GTTTCTCTTTC	ribosomal protein S8 (RPS8), mRNA	0.768
NM_001004	chr11	−	735497	16	29	CTTCCTTTTCC	ribosomal protein, large P2 (RPLP2), mRNA	0.921
NM_001002	chr12	−	120165589	78	154	CCTTCTCTCGC	ribosomal protein, large, P0 (RPLP0), transcript variant 1, mRNA	0.515
NM_003295	chr13	−	39901802	247	386	CGGCCTTTTCC	tumor protein, translationally-controlled 1 (TPT1), mRNA	0.779

3.　-1~+4 は必ずピリミジン

を条件とした。これらの条件を用いて再度 404 遺伝子に対して検索した結果、最初の 45 遺伝子以外にさらに 7 つの遺伝子が TOP 候補として検出された。

表 4-2　検出された ribosomal protein 以外の既知 TOP 遺伝子

本研究で検出した ribosomal protein 以外の TOP 遺伝子について示した。各列は、NM ID: RefSeq ID、chr: 染色体、strand: 遺伝子の方向、TSS: TOP 遺伝子であると判断された転写開始点、clones: DBTSS に登録されている 5'端配列、-4~+7: 転写開始点から-4~+7 の配列、definition: 遺伝子名を示す。

NM ID	chr	strand	TSS	clones	score	-4~+7	definition
NM_001402	chr6	-	74197456	3215	0.83	CGTTCTTTTTC	eukaryotic translation elongation factor 1 alpha 1 (EEF1A1)
NM_001959	chr2	+	205749235	10	0.74	GGTCCTTTTTC	eukaryotic translation elongation factor 1 beta 2 (EEF1B2)
NM_001961	chr19	-	4054739	97	0.84	CGTTCTCTTCC	eukaryotic translation elongation factor 2 (EEF2)
NM_002136	chr12	+	54788046	93	0.49	GCTCCTTTCTG	heterogeneous nuclear ribonucleoprotein A1 (HNRPA1)
NM_002295	chr3	+	38681430	32	0.81	CTGTCTTTTCC	laminin receptor 1 (ribosomal protein SA, 67kDa) (LAMR1)
NM_002520	chr5	+	171517398	22	0.77	CGTCCTTTCCC	nucleophosmin (nucleolar phosphoprotein B23, numatrin) (NPM1)
NM_002568	chr8	-	101804353	15	0.62	CTTCCCCTTCT	poly(A) binding protein, cytoplasmic 1 (PABPC1)
NM_003295	chr13	-	39901802	247	0.78	CGGCCTTTTCC	tumor protein, translationally-controlled 1 (TPT1)

表 4-3　翻訳伸張因子(elongation factor)と TOP 遺伝子予測結果

翻訳伸張因子のうちの TOP 遺伝子。各列は、NM id: RefSeq ID、TOP?: TOP 遺伝子（Yes）・TOP として検出されなかった（No）、: DBTSS に登録されている 5'端配列、definition: 遺伝子名を示す。

NM id	TOP?	#clones	definition
NM_001402	Yes	12067	elongation factor 1 alpha 1 (EEF1A1)
NM_001958	Yes	8	elongation factor 1 alpha 2 (EEF1A2)
NM_001959	Yes	29	elongation factor 1 beta 2 (EEF1B2), transcript variant 1
NM_021121	Yes	29	elongation factor 1 beta 2 (EEF1B2), transcript variant 2
NM_032378	Yes	21	elongation factor 1 delta (guanine nucleotide exchange protein) (EEF1D), transcript variant 1
NM_001960	Yes	21	elongation factor 1 delta (guanine nucleotide exchange protein) (EEF1D), transcript variant 2
NM_004280	No	8	elongation factor 1 epsilon 1 (EEF1E1)
NM_001404	Yes	1107	elongation factor 1 gamma (EEF1G)
NM_001961	Yes	157	elongation factor 2 (EEF2)

4.3.2 転写開始点 113,875 カ所からの検索

113,875 の転写開始点全てに関して、上記の 3 条件により検索した。その結果 793 転写開始点(511 Loci)が TOP 遺伝子候補として抽出された（補足表 1)。リボソーム遺伝子以外の TOP 遺伝子とされている遺伝子 8 種は、本研究の手法によって全て検出された(表 5-2)。

NCBI の RefSeq 遺伝子にはスプライシングバリアント 6 遺伝子を含む 84 遺伝子のリボソームタンパク質が登録されている。このうち、今回の手法で検出されたのは 81 種であった（補足表 1)。検出されなかった 3 種は ribosomal protein S4 Y-linked 2, ribosomal protein L27, ribosomal protein S9 であり、DBTSS に登録されている 5'端配列がそれぞれ 0,9,2 と少なかった。

RefSeq に登録されている翻訳伸張因子（eukaryotic elongation factor: EEF）は、9 遺伝子あった。このうち、EEF1E1 以外の 8 遺伝子は TOP 遺伝子として検出

表 4-4　翻訳開始因子(initiation factor)と TOP 遺伝子予測

翻訳開始因子のうちの TOP 遺伝子予測結果。各列は、NM id: RefSeq ID、TOP?: TOP 遺伝子（Yes)・目で見て TOP 遺伝子と判断（manual)、TOP として検出されなかった（No)、: DBTSS に登録されている 5'端配列、definition: 遺伝子名を示す。

NM id	TOP?	#clones	definition
NM_001412	NO	46	initiation factor 1A (EIF1A)
NM_004681	NO	5	initiation factor 1A, Y chromosome (EIF1AY)
NM_004094	NO	26	initiation factor 2, subunit 1 alpha, 35kDa (EIF2S1)
NM_003908	Yes	108	initiation factor 2, subunit 2 beta, 38kDa (EIF2S2)
NM_001415	Yes	39	initiation factor 2, subunit 3 gamma, 52kDa (EIF2S3)
NM_032025	NO	36	initiation factor 2A eIF2a (eIF2a)
NM_004836	NO	7	initiation factor 2-alpha kinase 3 (EIF2AK3)
NM_001414	NO	33	initiation factor 2B, subunit 1 alpha, 26kDa (EIF2B1)
NM_014239	NO	2	initiation factor 2B, subunit 2 beta, 39kDa (EIF2B2)
NM_020365	NO	30	initiation factor 2B, subunit 3 gamma, 58kDa (EIF2B3)
NM_015636	NO	4	initiation factor 2B, subunit 4 delta, 67kDa (EIF2B4), transcript variant 1
NM_012199	NO	9	initiation factor 2C, 1 (EIF2C1)
NM_012154	NO	0	initiation factor 2C, 2 (EIF2C2)
NM_013234	Yes	41	initiation factor 3 subunit k (eIF3k)
NM_003758	Manual	31	initiation factor 3, subunit 1 alpha, 35kDa (EIF3S1)
NM_003750	Yes	71	initiation factor 3, subunit 10 theta, 150/170kDa (EIF3S10)
NM_003757	Yes	60	initiation factor 3, subunit 2 beta, 36kDa (EIF3S2)
NM_003756	Yes	58	initiation factor 3, subunit 3 gamma, 40kDa (EIF3S3)
NM_003755	Manual	19	initiation factor 3, subunit 4 delta, 44kDa (EIF3S4)
NM_003754	Yes	70	initiation factor 3, subunit 5 epsilon, 47kDa (EIF3S5)
NM_001568	Yes	75	initiation factor 3, subunit 6 48kDa (EIF3S6)
NM_016091	Yes	689	initiation factor 3, subunit 6 interacting protein (EIF3S6IP)
NM_003753	Yes	101	initiation factor 3, subunit 7 zeta, 66/67kDa (EIF3S7)
NM_003752	Yes	297	initiation factor 3, subunit 8, 110kDa (EIF3S8)
NM_003751	NO	22	initiation factor 3, subunit 9 eta, 116kDa (EIF3S9)
NM_004953	NO	54	initiation factor 4 gamma, 1 (EIF4G1)
NM_001418	NO	647	initiation factor 4 gamma, 2 (EIF4G2)
NM_003760	NO	20	initiation factor 4 gamma, 3 (EIF4G3)
NM_001416	Yes	258	initiation factor 4A, isoform 1 (EIF4A1)
NM_001967	Yes	397	initiation factor 4A, isoform 2 (EIF4A2)
NM_001417	Yes	277	initiation factor 4B (EIF4B)
NM_001968	NO	44	initiation factor 4E (EIF4E)
NM_004095	NO	3	initiation factor 4E binding protein 1 (EIF4EBP1)
NM_004096	NO	17	initiation factor 4E binding protein 2 (EIF4EBP2)
NM_003732	NO	1	initiation factor 4E binding protein 3 (EIF4EBP3)
NM_019843	NO	7	initiation factor 4E nuclear import factor 1 (EIF4ENIF1)
NM_004846	NO	7	initiation factor 4E-like 3 (EIF4EL3)
NM_001969	NO	74	initiation factor 5 (EIF5)
NM_001970	NO	28	initiation factor 5A (EIF5A)
NM_020390	NO	0	initiation factor 5A2 (EIF5A2)

された(表 5-3)。EEF1E1 の 5'端配列は DBTSS 内に 8 配列あった。この転写開始点付近の配列を目で調べたが、pyrmidine に富む配列は見つからなかった (図 4-2 A)。

翻訳開始因子（eukaryotic initiation factor: EIF）に関しては、表 4-4 にまとめた。EIF2,EIF3,EIF4A,EIF4B 等が TOP 遺伝子として検出されたが、EIF4E, EIF4G

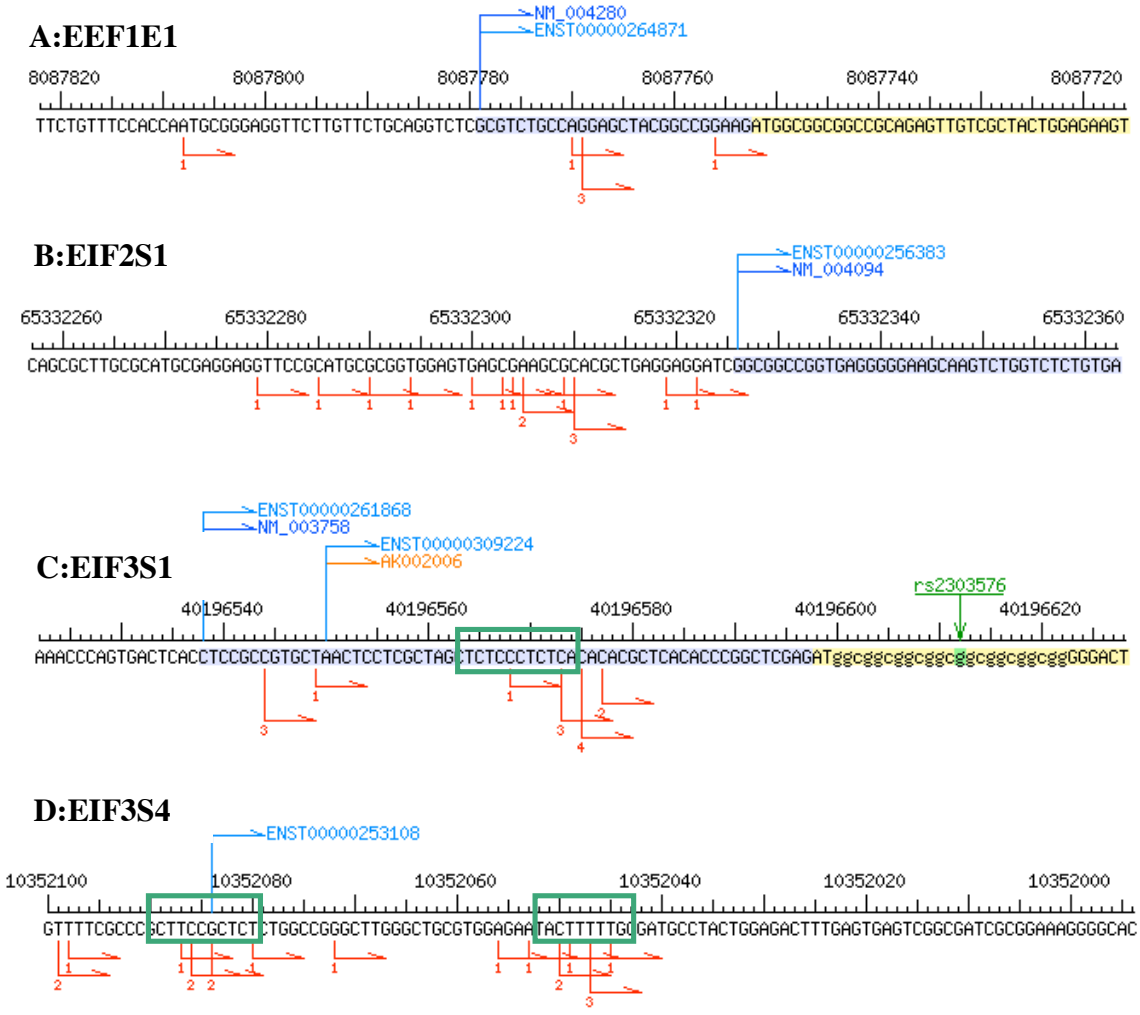


図 4-2　manual 探索の例

一部の遺伝子に関しては、TOP 遺伝子の判定を DBTSS の web page を利用し、目で見て行った。A,B:TOP 遺伝子ではないと判断される例。C,D:TOP 遺伝子の可能性が考えられる例。TOP の可能性が考えられそうな pyrimidine rich な配列を緑で囲んだ。

は TOP 遺伝子では無いと判断された。また翻訳開始因子 3 (EIE3)の 12 遺伝子のうち 9 遺伝子は TOP 遺伝子として検出された。目で確認した結果、さらにあと二つ (EIF3S1, EIF3S4) が TOP 遺伝子でありそうなことがわかった(図 4-2 C,D)。しかし残りの EIF3S9 は、pyrimidine に富む配列が見つからず、TOP 遺伝子ではないと判断した。

4.3.3 予測されたヒト TOP 遺伝子の組織特異性

TOP 遺伝子は、ribosomal protein や翻訳開始因子や翻訳伸張因子など、ユビ

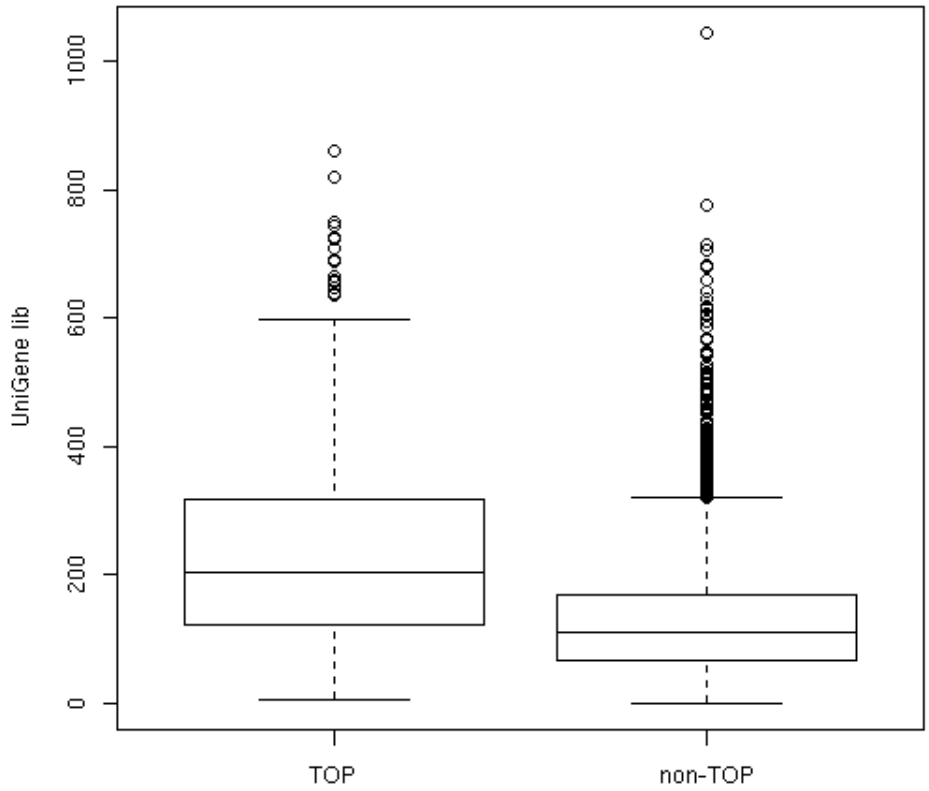


図 4-3　TOP 遺伝子と非 TOP 遺伝子との組織特異性比較

TOP 遺伝子として検出されたもの(TOP)と TOP ではないと判断されたもの（non-TOP）の箱ひげ図による発現比較。縦軸は、5 章で詳しく説明するある遺伝子の UniGene のソース数を示す。箱ひげ図は、小さい方から 1/４の位置の値を第 1 四分位、大きい方から 1/４の位置の値を第 3 四分位としたときに、この間に含まれる値の範囲を長方形で示したものである。第 1 四分位と第 3 四分位の差 1.5 倍した線をヒゲとして、第１四分位の下、第 3 四分位の上に伸ばす。それより外れた値は、〇で示してある。

キタスに発現している遺伝子に多いと考えられる。そこで、5 章で詳しく述べるが、組織特異性の指標として UniGene の library 数を TOP 遺伝子と予測された 511 遺伝子と残りの 10,755 遺伝子とで比較した(図 4-3)。その結果、TOP 遺伝子と予測した遺伝子群の方が、ユビキタスに発現していた。Wilcox test の結果 $P<10^{-100}$ で有為差が認められた。

4.3.4 マウスの転写開始点の TOP 遺伝子検索

DBTSS ver.3 には、マウスの転写開始点も登録されている。この情報を利用してマウスでも 4.3.1 で述べた条件で検索した。ただし、マウスの転写開始点は、２章で述べたように原因不明であるがアーティファクトと考えられる揺ら

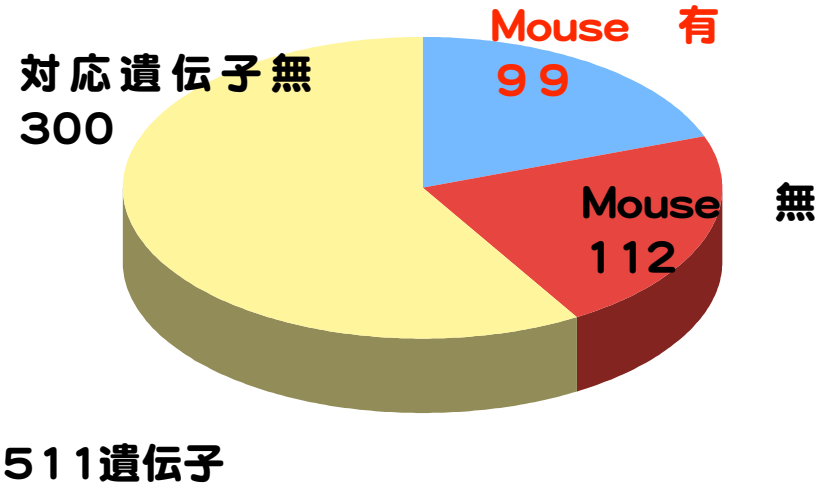


図 4-4　human-mouse のオーソログ遺伝子比較

TOP 遺伝子と検出したもののうち、マウスにオーソログ遺伝子がどのくらいあったかを示した。211 遺伝子は、マウスにオーソログ遺伝子があった。そのうち、マウスで TOP 遺伝子と検出された 1405 遺伝子内に含まれていたものを青色、マウスでは、検出できなかったものを赤で示した。オーソログ関係が得られなかったものは、黄色で示した。

ぎがあるため、転写開始点から-5~+5 の範囲内を探索し、一つでも陽性であればその遺伝子は陽性とした。その結果 1405 遺伝子が陽性と判断された。

4.3.5 検出されたヒト・マウス TOP 遺伝子のオーソログ遺伝子比較

NCBI の homologene の情報を利用して、TOP 遺伝子としたヒト 511 遺伝子とマウス 1405 遺伝子間でオーソログ関係を取ると、211 遺伝子に関してオーソログ関係が得られた。このうち、99 遺伝子が、ヒト、マウス共に TOP 遺伝子と予測された(図 4-4)。この 99 遺伝子に関して、Gene ontology や文献により生物学的な機能を調べたところ、20 遺伝子がリボソームタンパク質、6 遺伝子が翻訳に関わるものであった。さらに、5 遺伝子（chaperonin containing TCP1 subunit の 3,4,8, heat shock 105kDa/110kDa protein1, t-complex 1）がシャペロンに関わる物であるものであった (図 4-5)。

4.4 考察

本研究により、ヒトの 511 遺伝子が TOP 遺伝子である可能性が示唆された。

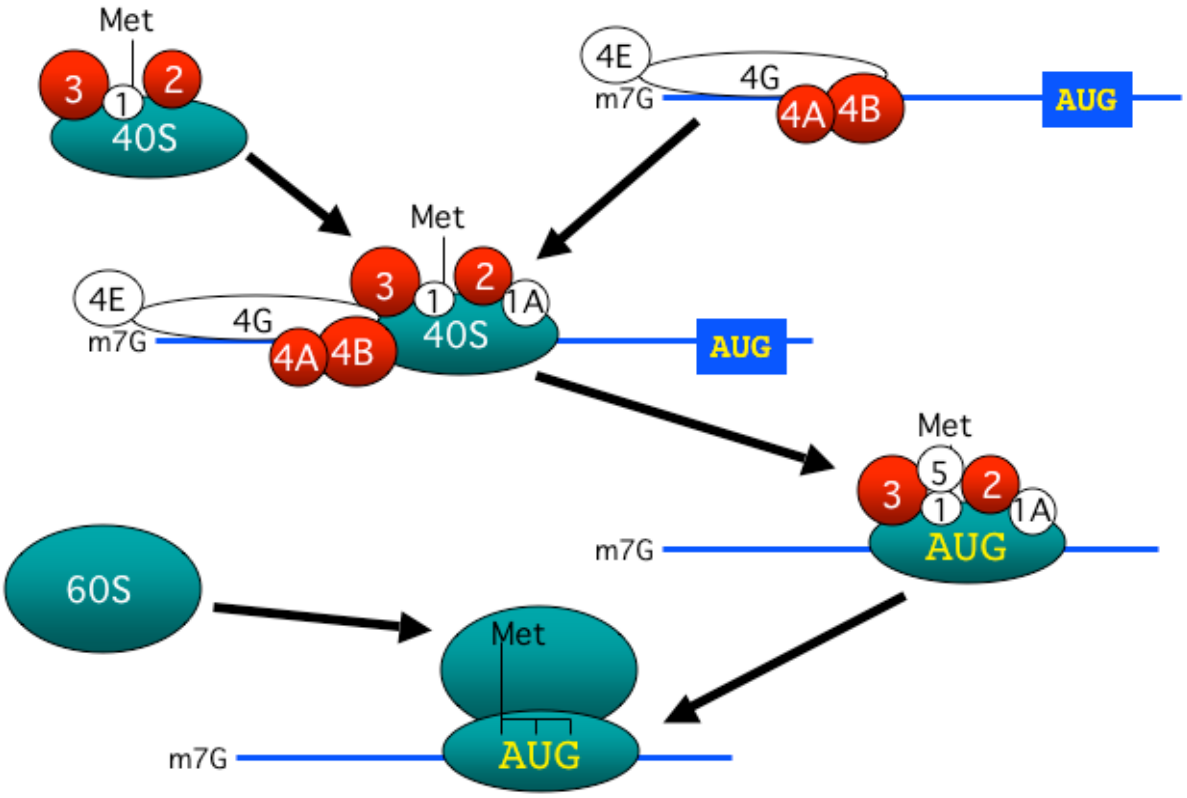


図 4-6　翻訳開始に関わる TOP 遺伝子

これは、511 遺伝子には、現在までに TOP 遺伝子として知られている 88 種のうち 85 種まで含む。従って、今回私の用いた手法は、擬陽性が少ないことが考えられる。ただし、今回の手法は、転写開始点があって初めて検出できる手法である。例えば、検出できなかった 3 つの ribosomal protein は、DBTSS に登録されている 5'端配列が 0,2,9 個と少なかった。このように、5'端の情報が不足している場合は、今回の手法では検出不可能である。

本手法での有効性をさらに検討するために、511 遺伝子がマウスのオーソログ遺伝子でどれだけ TOP 遺伝子であると予測されているかを調べた、オーソログ関係が取れた 211 遺伝子のうち 99 遺伝子(47%)はヒトでもマウスでも TOP 遺伝子として検出された。従って 511 遺伝子のうち、47%程度の約 240 遺伝子は、真に TOP 遺伝子であることが示唆される。しかし、211 遺伝子のうち、マウスで TOP 遺伝子として検出されなかった 112 遺伝子全てが擬陽性であるとは限らない。マウスの DBTSS 内に TOP 配列部分の 5'端配列が登録されていないので、本手法では検出されなかったことも考えられるためである。

今回予測した TOP 遺伝子群を細かく見ていくと，いくつか興味深いことがわかった。従来まで TOP 遺伝子として知られていたのは，翻訳伸張因子(eukaryotic elongation factor: EEF)である EEF1A1、EEF1B2、EEF2 があった。しかし、それらも含め翻訳開始因子 9 種のうち、EEF1E1 を除く 8 種が翻訳開始因子であることが示唆された。また、従来知られていなかった、いくつかの翻訳開始因子(eukaryotic initiation factor: EIF)も TOP 遺伝子であることが示唆された。EIF はいくつものサブクラスに大別されているが，そのうち翻訳課程においてリボソームと直接結合するような EIF2、EIF3、EIF4A、EIF4B は TOP 候補として挙げられるのに対し，CAP の認識にかかわる EIF5E 等直接リボソームと結合しないものは，TOP 遺伝子ではない可能性が考えられた(図 4-6)。また，EIF3 はサブユニット 12 個のうち 11 個が TOP 遺伝子であると推測された（表 4-4)。このように，同一のクラスに属する遺伝子が同時に抽出できたことは，この検出法の有効性を支持していると考えられる。

検出された TOP 遺伝子群には、シャペロンやタンパク質輸送等、翻訳と関わりの深い役割を担っている遺伝子を含んでいた。また、TOP 遺伝子は、従来考えられていた翻訳関係の遺伝子にあるばかりでなく、一部の膜タンパク質、転写因子等、より広範囲に存在する可能性が示唆された。

図 4-3 より、今回予測した TOP 遺伝子は、ユビキタスに発現している遺伝子に富むことがわかった。翻訳は、全ての細胞で不可欠な作業であり、ユビキタスに発現していることは容易に想像できる。しかし、翻訳に関わる遺伝子以外にも TOP と考えられるものがあり、その mRNA の発現がユビキタスに観察されてた。このことは、例えば饑餓状態の時に、TOP 構造を持つ遺伝子の発現を一斉に止めることで、全体の翻訳レベルだけでなく、翻訳装置そのものの数を減らし、生物学的にコストのかかる翻訳を停止する機構があることが考えられる。

今回，TOP 遺伝子候補を網羅的に予測することができたが，これらは本当に翻訳制御を受ける一連のタンパク質をコードしているのであろうか。これについては、これらの mRNA を適当な細胞で発現させて、polysome や monosome の状態を観察する必要がある。今後、完全に TOP 遺伝子であることを実証するには、この実験が不可欠である。ただし、今回 TOP 遺伝子と予測した遺伝子は、翻訳とは関係ない DNA の-4 まで保存されていた mRNA 領域だけでなく、非転写領域も保存されていたことは、単に翻訳だけでなく転写でも制御を受けてい

る可能性がある。実際、EEF1A に関しては、転写レベルでも制御を受ける可能性についての報告がある(Shibui-Nihei et al., 2003)。こうした事実をふまえると、TOP 遺伝子は翻訳制御というよりだけでなく、転写制御にも関わっている可能性が強い。この可能性は、２章の解析の結果の TOP は Initiator と同じ位置に存在していると言うことからも裏付けられる。今後、翻訳制御だけでなく転写制御も視野に入れた実験的な解析が、TOP 遺伝子の生物学的意義を明らかにする上で欠かせないであろう。

5. CpG island と遺伝子発現の組織特異性との関係

5.1 序論

ほ乳類のゲノムは、GC 含量によって isochore という単位に分画できる。Isochore は重い部分(heavy isochore)と軽い部分(light isochore)に分けられる。少し前の報告では、重い Isochore はクロマチンの凝集度が低くユビキタスに発現している遺伝子が多いとされ、軽い Isochore は組織特異的な遺伝子群が多いとされていた(Bernardi, 1995)。しかし近年、Isochore と遺伝子の組織特異性に相関が見られないという報告が相次いだ(Goncalves et al., 2000; Ponger et al., 2001)。さらに、遺伝子の GC 含量と遺伝子発現はあまり関係がないという報告もある(D'Onofrio, 2002)。また、ヒトの遺伝子では翻訳領域コドンの三番目の文字（3rd コドン）において、GC 含量の低い遺伝子は組織特異的に発現するが、マウスにおいては差は見られなかったという報告もされた(Vinogradov, 2003)。

一方、promoter 領域の CpG island と遺伝子発現の相関が考えられてきた。一般にプロモータ領域に CpG island がある遺伝子はユビキタスに発現しており、そうでない遺伝子は組織特異的に発現していると言われている。これは次のように説明されている。ユビキタスな発現をしている遺伝子のプロモータにある CpG の C の部分はメチル化されていないが、それ以外の CpG はゲノム上でメチル化されている。メチル化されている C は T に変化しやすく、その領域には T やその相補鎖である A が蓄積するが、ユビキタスな発現をしている遺伝子のプロモータ領域ではこの置換が起こりにくい。結果として、ユビキタスな遺伝子のプロモータ領域に CpG が蓄積したためであると説明されている(Gardiner-Garden and M., 1987; Larsen et al., 1992)。しかし、ヒトゲノムが決定された時、CpG island とプロモータとの関係を見いだすことは出来ていない(Lander et al., 2001)。従って、大規模にプロモータ領域の CpG islands と組織特異性を見た報告はない。

このように GC 含量・CpG island と遺伝子発現の組織特異性について一貫性のない結果が生じている要因は、二つあると考えられる。１つは、遺伝子の転写開始点が正確にわかっていないため、プロモータ領域の CpG island を正確に定義できないためである。そしてもう１つは、組織特異性を表す適切な指標が存在しないためである。本研究の３章においてヒトプロモータ領域の CpG island を探索し、全体の 76.0%の遺伝子のプロモータ領域に CpG island があることがわかった。以下、第５章では、ヒトだけでなく、DBTSS ver.3 で使用できるよ

うになったマウスの転写開始点情報も用いて CpG islands とプロモータ及び組織特異性との関係を調べた。まず、遺伝子群をプロモータ領域に CpG island がある遺伝子（CpG+遺伝子群）とない遺伝子（CpG-遺伝子群）に分類した。さらに、遺伝子の組織特異性を見積もるために、NCBI の UniGene を用いてその遺伝子が何種類の細胞・組織から cDNA が得られているかを指標にした。その結果、CpG のある遺伝子群（ヒト 6600,マウス 2948 遺伝子）は、ユビキタスな発現が見られるのに対し、CpG islands のない遺伝子群（ヒト 2619,マウス 1830 遺伝子）では組織特異的に発現することが示唆された。

5.2 材料と方法

5.2.1 代表 TSS の選別

データセットは、DBTSS ver.3 から、ヒト、マウスの転写開始点情報（ヒト 11234 遺伝子、マウス 7534 遺伝子）を元にした。一つの遺伝子に複数の転写開始点が存在する場合があることは、２章で述べた。本研究では簡便化して考えるために、１つの遺伝子から１つの代表転写開始点を以下のように設定した。University of Carifornia Santa Cruz(UCSC)の UCSC Genome browser には、

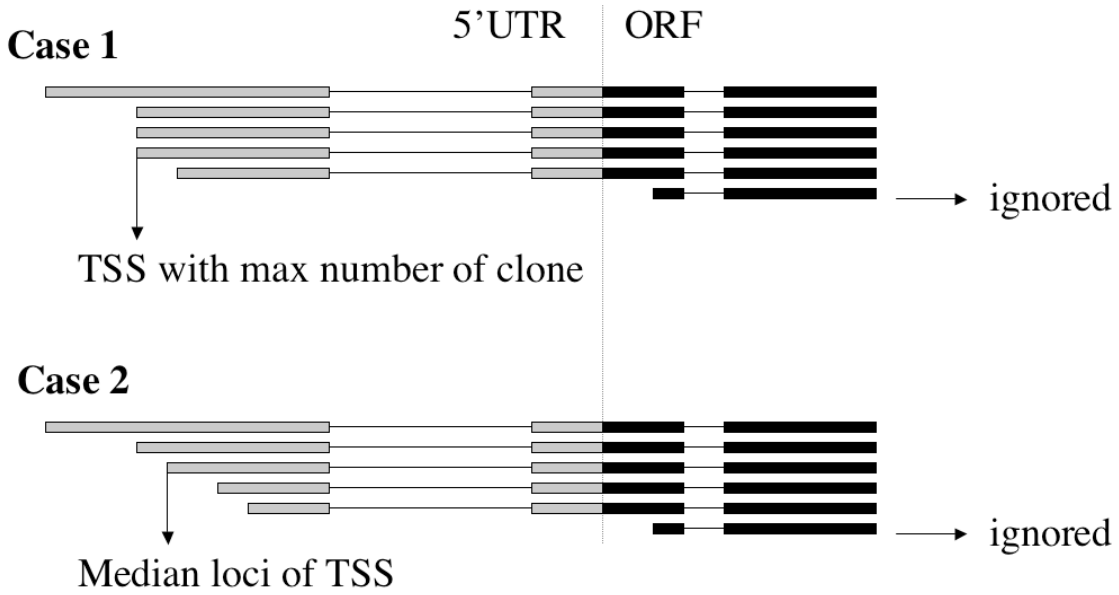


図 5-1 代表転写開始点の設定

Case1: 5'端配列による最頻値の転写開始点情報がただ一つ決まれば、それを代表転写開始点とした。Case2:Case1 で最頻値を示す 5'端配列が複数あった場合、あるいは、全てが一つの 5'端配列だった場合には、翻訳開始点から上流のもののうち中央値を取った。もし、偶数だった場合には、より 5'端方向に伸びているものを採用した。

refGene.txt という RefSeq の cDNA をゲノムにマップしたデータがある。これには、RefSeq cDNA の Open Reading Frame(ORF)のゲノム上での位置情報がある。これを利用して翻訳開始点を求め、一つの遺伝子内で翻訳開始点情報より下流に転写開始点がマッピングされているものは、その遺伝子の翻訳情報を正確に反映していないと判断して全て除去した。その後、もし最頻値の転写開始点情報があれば、それを代表転写開始点と定義した。最頻値が複数ある場合、あるいは全ての転写開始点においてサポートする 5'端配列が一つしかない場合には、翻訳開始点より上流の転写開始点のうち中央値を選び出した。このとき、最頻値の転写開始点が偶数である場合は、より 5'端に延びている方を中央値とした（図 5-1）

5.2.2 転写開始点付近の GC 含量と CpG score の計算

全ての遺伝子の代表転写開始点に対して、上流-2000 下流+2000 を取り出した。ゲノム配列未決定領域のギャップがある場合は、GC 含量の値が大きく崩れるので、取り出した配列に未決定文字の N が 5 文字以上ある配列は、解析対象から外した。

これらの配列の GC 含量(G+C/200)と CpG score((CG/G*C)*200)を window サイズ 200 塩基で 1 塩基ずつ移動させながら求めた。転写開始点からの相対位置が同じ場所の window の値を全ての遺伝子についてヒト・マウスについてそれぞれ求め、その平均値を求めた。

5.2.3 UniGene から遺伝子発現部位数の推定

組織特異性を見るためには、その遺伝子が何種類の細胞・組織から cDNA が得られているかを指標にすればよい。本研究では、その指標として一つの遺伝子に登録されている cDNA がいくつのライブラリに由来するかを UniGene を用いて推定した。

UniGene は、GenBank に登録されている配列をクラスタリングしたものである。この中にある LID、すなわちその cDNA が得られた libarary の ID を取り出し、この ID の種類を数えた。さらに、NCBI の LocusLink にある loc2ug を利用して、対応する LocusLink ID を求め、その LocusLink ID に対応する RefSeq 遺伝子を loc2ref によって求めた。この作業により、UniGene library 数を RefSeq 遺伝子と対応付けができ、その遺伝子の組織特異性の指標とした（図 5-2）

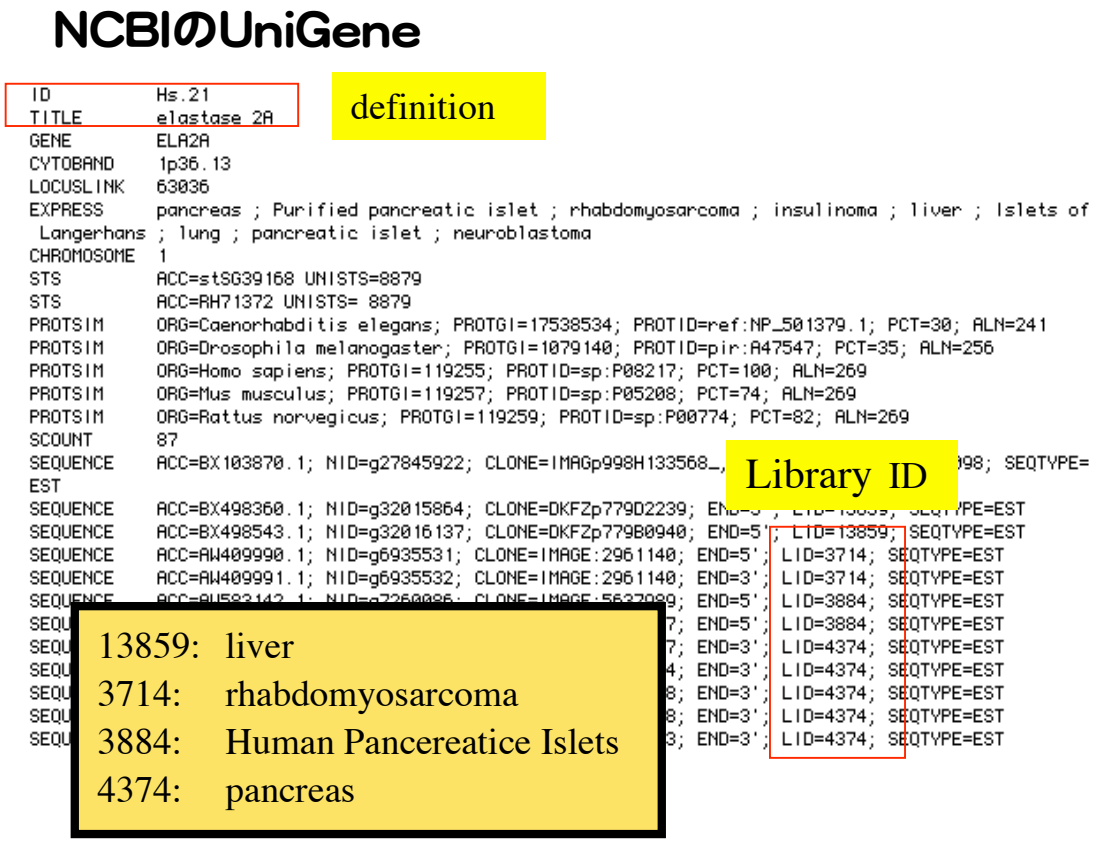


図 5-2　UniGene による組織特異性の指標

一つの遺伝子に登録されている LID 数を数えた。この例では、Hs.21 に分類される cDNA 配列は、11 個あるが、そのソースは 4 種である。この４というソース数を組織特異性の指標とした。

5.2.4 組織特異性を比較する領域の定義

全ての遺伝子に対して、代表転写開始点から-100~+100 領域をプロモータ領域（promoter）と定義した。また、対照としてエクソン領域（spliced mRNA)、エクソン＋イントロン領域(Transcribed DNA)をそれぞれ定義した。さらに、翻訳領域のコドンの三番目を 3rd コドンと定義した (図 5-3)。

5.2.5 ヒト・マウスのオーソログの同定

NCBI の LocusLink にある homologene は、生物種間のオーソログ遺伝子の対応表である。これを利用して、ヒトとマウスのオーソログ遺伝子を決定した。

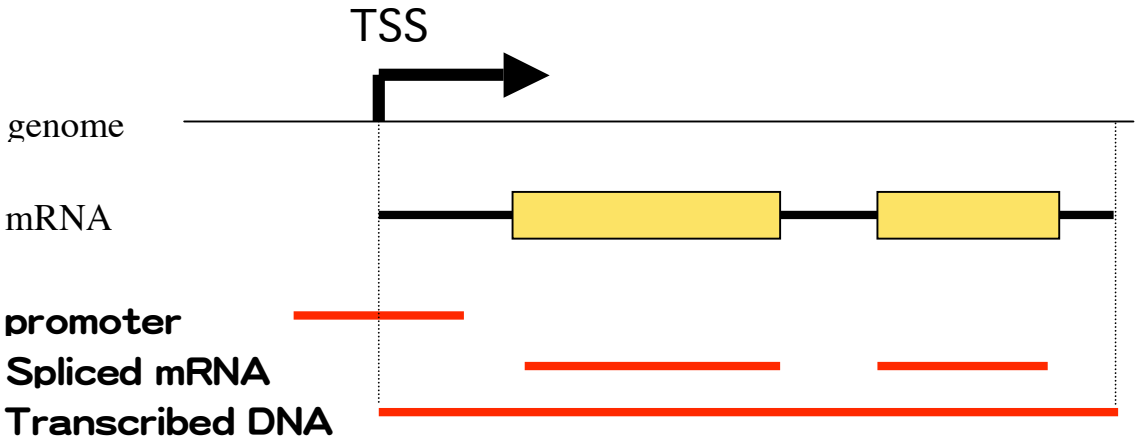


図 5-3　GC 含量を解析した領域

エクソンや、TSS に対して GC 含量を計算した領域を図示した。Promoter:TSS から-100~+100、spliced mRNA:exon 部分、Transcribed DNA: exon+intron。

5.2.6 chromosome band の濃さと発現量

ヒトに関しては、UCSC　Genome browser に cytoBand.txt という情報がテキスト形式で存在する。これは、chromosome band の濃さとゲノム上の位置をおおよそ対応づけることができる(Furey and Haussler, 2003)。各遺伝子を、その転写開始点がそれぞれ band のどの濃さに属するかによって 0,25,50,75,100 の 5 段階に分類した。

5.2.7 GO annotation

NCBI の LocusLink にある loc2go のテーブルを使い、ヒト 6505 遺伝子とマウス 2777 遺伝子に関して、Gene Ontology(Harris et al., 2004)の単語を割り振ることができた。GO は階層構造になっており、ここで得られた単語は下位階層に属する単語が多く、細かすぎて分類に向いていない。そこで、これを、EBI の GO slim を用いて上位階層の単語に置き換えた。各 GO 単語についてプロモータに CpG　islands のある群と無い群での出現頻度を、全体の CpG　islands の有無（ヒト CpG+ 6600,CpG- 2948: マウス CpG+2619,CpG-1830）を元に超幾何分布を仮定して p-value を求め、有意性を検討した。

5.2.8 TATA の検索

TATA の検索には TRANSFAC 7.2 にある位置特異的スコア行列を利用した。TRANSFAC 7.2 にある二つの TATA 用の行列（V$_TATA_01、V$_TATA_C）を使い、モチーフ検索プログラムである MATCH を使って転写開始点から-35～+25 の範囲のみに検索を行った。閾値の設定は、比較的 False negative を排除できるとされている minFN72.lib を用いた。どちらか一方のスコア行列で陽性と判断された場合、TATA 有と判定した。

5.2.9 転写開始点の揺らぎ

代表転写開始点から+100~-100 内にマップされた転写開始点を全て抜き出した。それらのゲノム上の座標から標準偏差を式 3-1 により求めて、これを転写開始点の揺らぎと定義した。

5.3 結果

5.3.1 プロモータ領域の GC 含量と CpG score

プロモータ領域のどの辺りに CpG islands が有るのかを推定するために、-2000~+2000 の領域の GC 含量と CpG score を求めた。GC 含量はヒト-

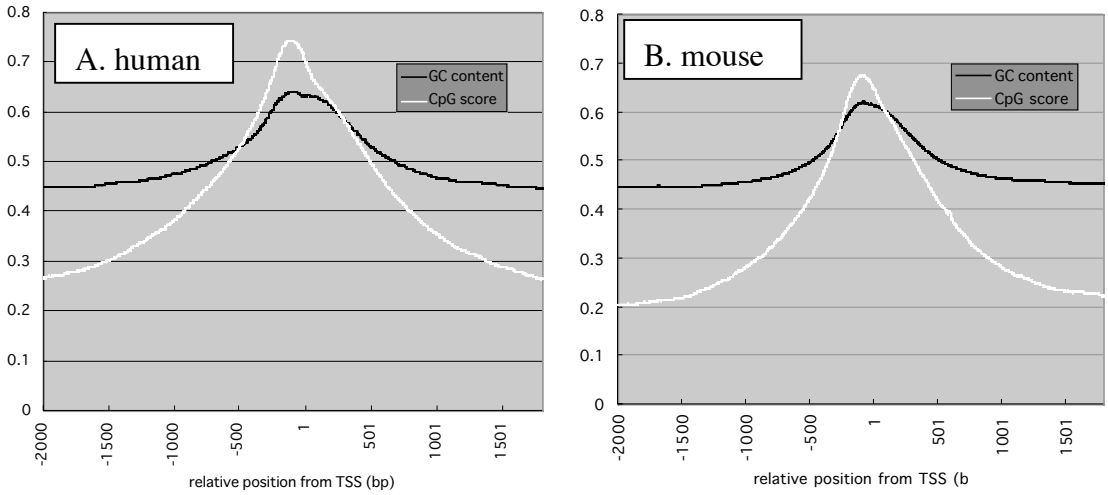


図 5-4　–2000~+2000 の GC 含量と CpG score

全ての遺伝子の上流配列を 1 塩基ずつずらして GC 含量と CpG score を 200 塩基の window size ごとに計算した。各ポジションでの平均値を human(A)、mouse(B)で示した。例えば 1 は、+1~+200 までの範囲を意味する。

、

106~+94(0.64) ,マウスは-83:+117(0.62)で最大になった。CpG score に関しては、ヒト-118:+82(0.74) ,マウスは-87~+113(0.67)で最大になった（図 5-4)。つまり、ヒト・マウスとも転写開始点付近で両者のスコアが最大になった。そこで、転写開始点から-100~+100 の 200 塩基長をプロモータ領域と定義し、以下の解析を行った。

5.3.2 GC 含量と組織特異性の分布

図 5-3 で分類したゲノム上での遺伝子領域の中で、どの部分が GC 含量と組織特異性が関係しているのか調べるために、mRNA（exon 領域のみ)、DNA（exon+intron)、5.3.1 で定義したプロモータ領域（-100~+100)、3rd position の GC conotent と組織特異性の関係をそれぞれ調べた。DNA（図 5-5 B,F）, mRNA（図 5-5 C,G）ともにヒトでもマウスでも GC 含量と UniGene ソース数との関係は観察されなかった。しかし、プロモータ領域では、組織特異性の低い、すなわちユビキタスに発現している遺伝子群では、GC 含量が高い傾向にあった（図 5-5　A,B)。この傾向は、ヒト・マウスで同様に認められ、相関係数も高い傾向にあった。また、3rd コドンに関しては、ヒトでは全体的に分布していたが、マウスでは組織特異性が高い遺伝子は GC 含量がやや高かった（図 5-5 D,H)。

5.3.3 プロモータ領域に CpG islands がある遺伝子の選定

プロモータ領域の CpG score と組織特異性に相関が有ることがわかったので、遺伝子をプロモータに CpG islands が有る遺伝子群（CpG+遺伝子群）と無い遺伝子 (CpG-遺伝子群) に分類した (図 5-6)。CpG islands の定義は Gardinar-Garden らの定義(GC 含量　>=　0.5、CpG score　>= 0.6)にしたがった。その結果、ヒトでプロモータ領域に CpG　islands がある遺伝子（CpG+）6620、無い遺伝子（CpG-）2948 に分類された。マウスでは CpG　islands がある遺伝子 2619、無い遺伝子が 1830 であった。

5.3.4 組織特異性とプロモータ領域の CpG islands

プロモータ領域に CpG islands が有る遺伝子と無い遺伝子に分け、組織特異性を UniGene のライブラリソース数で比較した (図 5-7)。ヒト・マウスともに CpG +とCpG-の群で顕著な差が認められた。Wilcox test によりヒト・マウス共に p

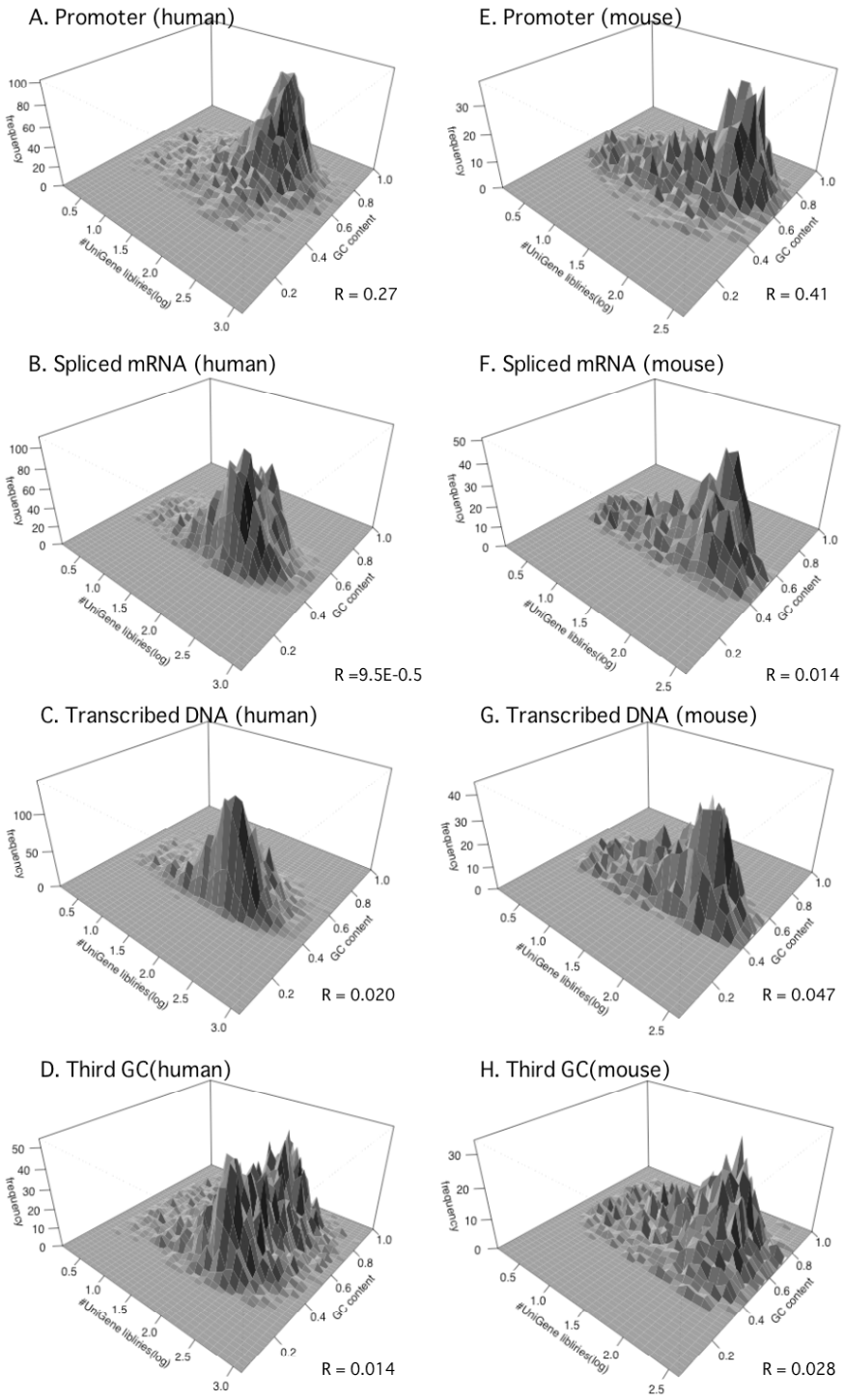


図 5-5　各領域の GC contentと発現の組織特異性

GC contentと Unigenelibrary数を x , y軸に z 軸にその属性を持つ遺伝子の数を表示している。左側の列がヒト、右側の列がマウスの結果である。転写開始点から-100:+100の promoter(A,E)、ｅｘｏｎすなわち spliced mRNA(B,F)、exon+intronすなわち Transcribed DNA(C,G)、third GC(D,H)に分けた。各図の横には、相関係数（R）を示した。

、

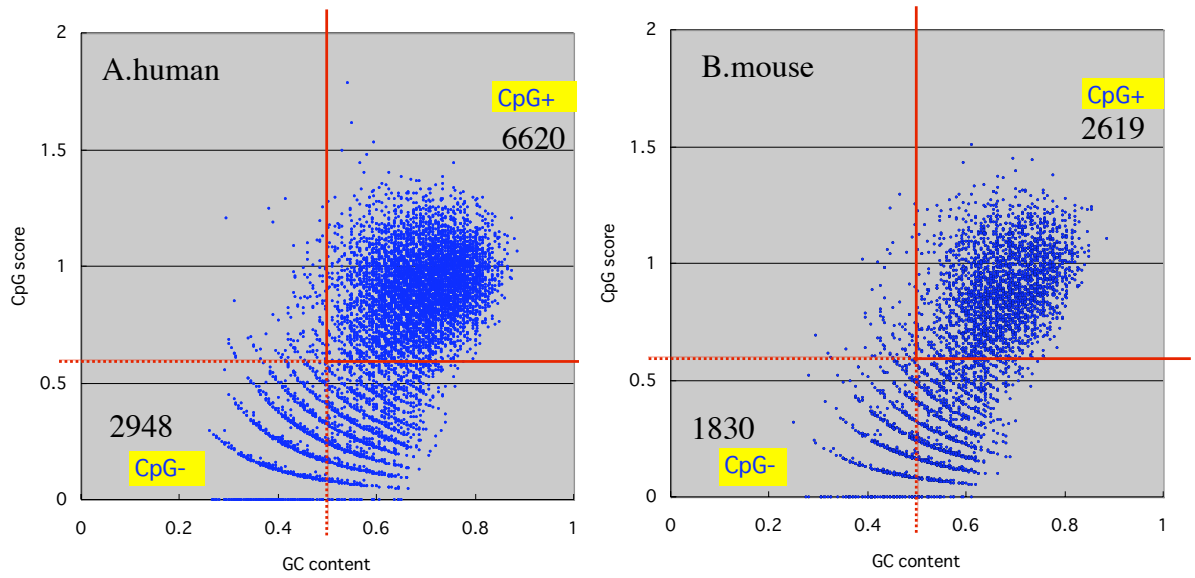


図 5-6　–100~+100 の CpG score と GC 含量の散布図

GC 含量 >= 0.5 CpG score >= 0.6 を CpG+遺伝子、それ以外を CpG-遺伝子と定義した。A:human、B:mouse。

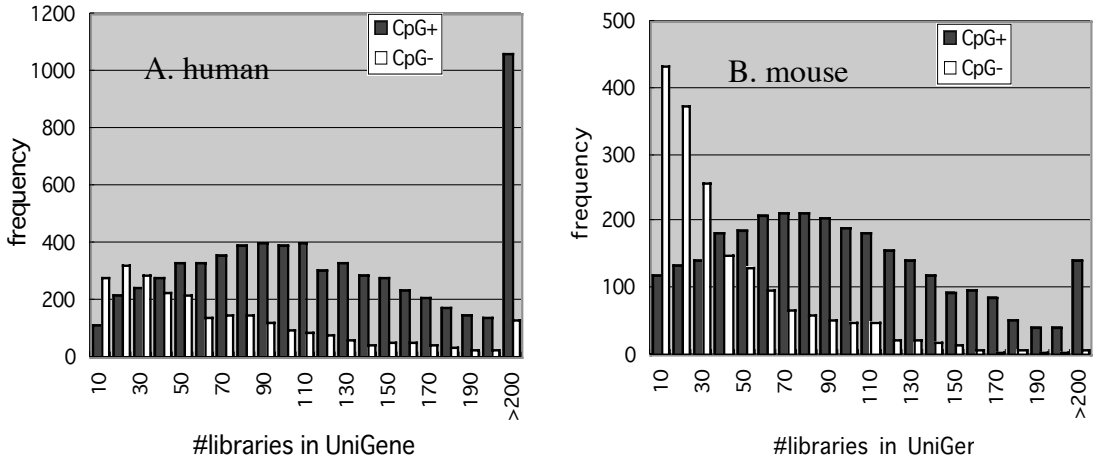


図 5-7　CpG+遺伝子と CpG-遺伝子の発現量の差

A:9219 human 遺伝子、B:4778 mouse 遺伝子。

$<10^{-100}$ と有意差があることが確認された。

5.3.5 マウス・ヒトのオーソログ遺伝子同士の比較

今回用いたデータセットの中で、ヒトとマウスのオーソログ遺伝子は 1472

あった。まず、そのオーソログ遺伝子内でプロモータ領域に CpG islands があるか無いかを比較した（表 5-1）。このうち、ヒトとマウスで同じカテゴリーに分類されたのは、87%(1283)であった。

続いて、オーソログ同士で、UniGene による組織特異性の比較を行った。この結果、R=0.72 と相関関係が認められた（図 5-8）。また、ヒト・マウス共に CpG+ と CpG-に関して同じクラスに分類された遺伝子群（CpG+:980、CpG- 303）の組織特異性を、ヒトとマウスのライブラリ数の和で比較した(図 5-9)。こちらも図 5-7 でヒト・マウスだけで比較した場合と同様、Wilcox test で $p<10^{-100}$ で有為差が認められた。

表 5-1　オーソログ遺伝子間の CpG island に関する保存

-		human	
		CpG+	CpG-
mouse	CpG+	980	84
	CpG-	105	303

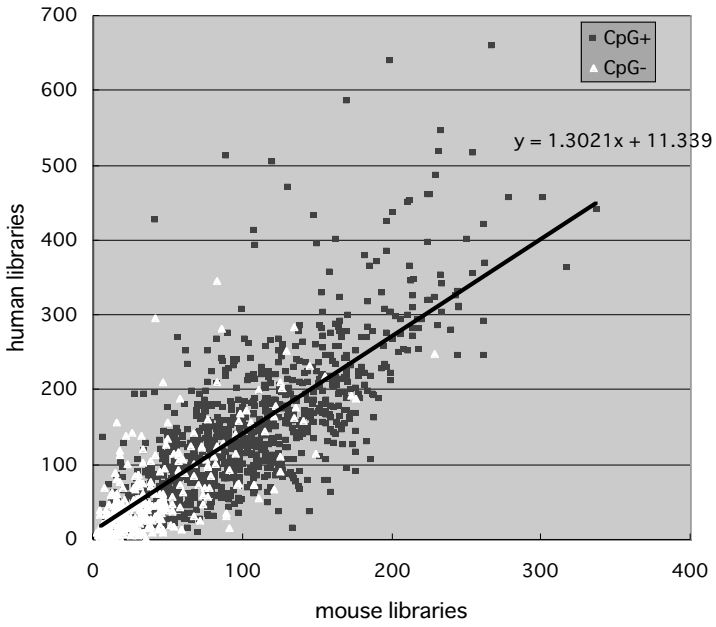


図 5-8　human と mouse の Unigene ライブラリ数の比較

横軸に mouse のライブラリ数、縦軸に対応する human のライブラリ数を 1283 オーソログ遺伝子に対してプロットした。

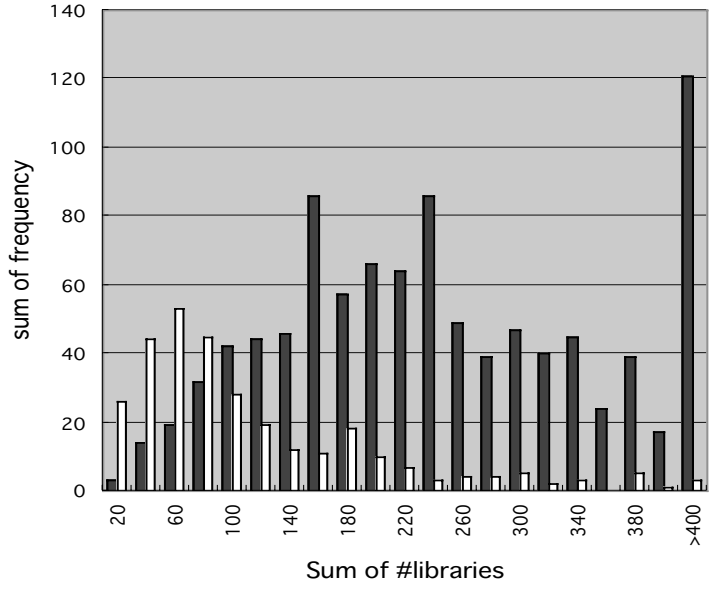


図 5-9　オーソログ遺伝子のみの CpG+、CpG-発現比較

横軸に human と mouse のライブラリ数の和、縦軸に human と mouse の頻度の和を、1283 オーソログ遺伝子に対して CpG+（黒）、CpG-（白）の棒グラフに分けて比較した。

5.3.6 Chromosome band と発現量の関係

ヒトに関しては、ゲノム上での Chromosome band 濃さの大まかな位置がわかる。これを利用して、band の濃さと遺伝子発現の組織特異性を CpG+遺伝子と CpG-遺伝子に分類して比較した(図 5-10)。CpG+遺伝子では、chromosome band の色が濃くなると UniGene の libarary 数が減少することが Spearman rank test で確認された($p<10^{-9}$)。しかし、CpG-遺伝子ではその傾向は見られなかった（$p<0.13$）。ただし、CpG+と CpG-が最も接近した band の色 100 の所でも、UniGene のソース数の分布に両者で差が見られた（Wilcox test: $p<10^{-11}$）

5.3.7 GO アノーテション

CpG+遺伝子群では、enzyme, metabolism といったユビキタスに発現する遺伝子に関わるような GO term が有為に認められた。一方、CpG-遺伝子群では、cell communication、extracelur, physiological process, signal transductio という、組織特異的発現する遺伝子に関連する GO term が顕著であった(表 5-2)。

図 5-2　CpG+遺伝子と CpG-遺伝子の Gene ontology

	human					mouse				
GO term	CpG+	ratio (%)	CpG-	ratio (%)	p-value	CpG+	ratio (%)	CpG-	ratio (%)	p-value
B:cell communication	813	12.3	516	19.8	3.30E-19 **	223	11.7	260	21.3	4.26E-13 **
B:cell growth and/or maintenance	1151	17.5	457	17.5	0.493	431	22.5	214	17.5	3.94E-04 **
B:metabolism	2512	38.1	774	29.6	7.54E-15 **	787	41.1	367	30.1	1.77E-10 **
B:death	170	2.6	86	3.3	0.037 *	65	3.4	31	2.5	0.104
B:developmental processes	381	5.8	234	9.0	5.05E-08 **	149	7.8	123	10.1	0.016 *
B:physiological processes	500	7.6	471	18.0	3.73E-45 **	115	6.0	203	16.6	2.66E-21 **
C:cell	3221	48.9	1197	45.8	4.71E-03 **	1213	63.4	672	55.0	1.90E-06 **
C:extracellular	163	2.5	262	10.0	1.56E-48 **	276	14.4	438	35.9	1.76E-43 **
C:unlocalized	31	0.5	9	0.3	0.263	7	0.4	2	0.2	0.240
M:chaperone	93	1.4	16	0.6	5.57E-04 **	39	2.0	6	0.5	1.46E-04 **
M:enzyme	1650	25.0	539	20.6	3.83E-06 **	528	27.6	299	24.5	0.029 *
M:enzyme regulator	171	2.6	97	3.7	2.91E-03 **	43	2.2	45	3.7	0.012 *
M:ligand binding or carrier	2317	35.2	935	35.8	0.284	862	45.1	530	43.4	0.192
M:motor	51	0.8	12	0.5	0.062	8	0.4	3	0.2	0.317
M:signal transducer	387	5.9	307	11.8	1.29E-20 **	125	6.5	155	12.7	4.36E-09 **
M:structural molecule	232	3.5	81	3.1	0.176	96	5.0	53	4.3	0.217
M:transcription regulator	385	5.8	130	5.0	0.057	127	6.6	66	5.4	0.092
M:transporter	556	8.4	229	8.8	0.315	205	10.7	114	9.3	0.118

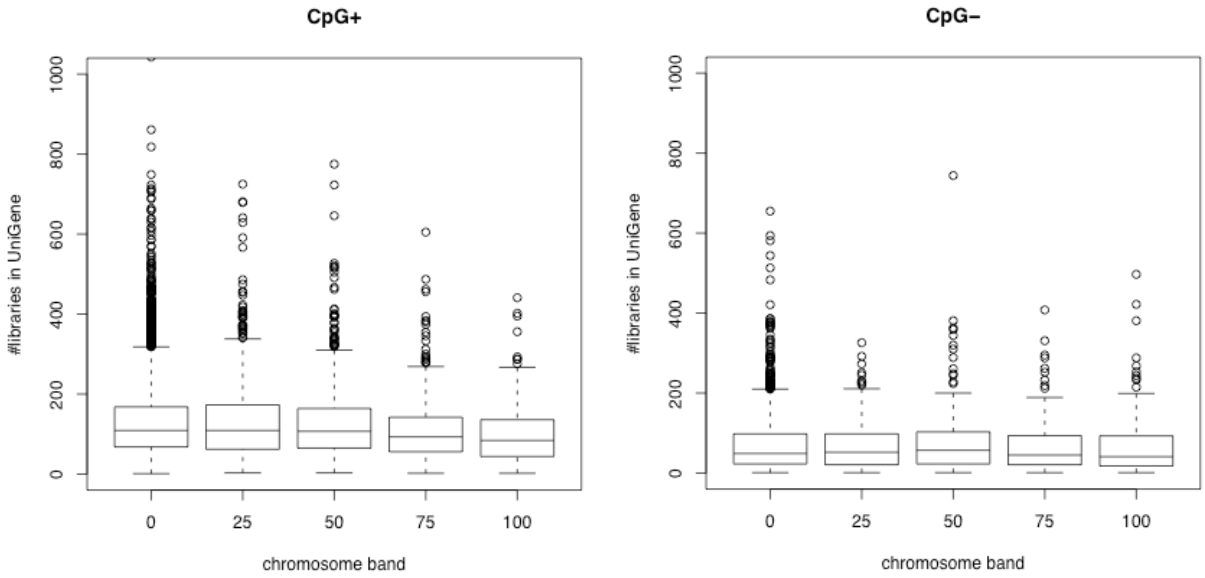


図 5-10　ヒト遺伝子の染色体 band の濃さと発現量

遺伝子を UCSC genome annotation に従って 5 つのクラス（0, 25, 50, 75, 100）に分けた。これを、CpG+遺伝子と CpG-遺伝子に分け、UniGene のライブラリ数の比較を箱ひげ図で示した。

、

5.3.8 TATA のある遺伝子と CpG islands がある遺伝子の揺らぎの差

ヒト 9,219、マウス 4,778 のプロモータ領域のうち、位置特異的スコア行列によって TATA-box があると判断された遺伝子は、ヒトで 1,092，マウス 874 あった。これを、プロモータ領域に CpG islands があるかどうかを加味して分類した結果を表 5-3 に示した。TATA-box があって CpG islands が無い遺伝子群（TATA+・CpG-: ヒト 576、マウス 524）と TATA-box が無くて CpG islands がある遺伝子群（TATA-・CpG+: ヒト 6084、マウス 2600）間で、転写開始点の揺らぎを標準偏差によって求めた。TATA+・CpG+の遺伝子群（ヒト 516、マウス 348）は、TATA-box、CpG islands のどちらの影響を受けているのかわからないため、解析からはずした。ヒト、マウスともに TATA があり CpG islands がない遺伝子群では、転写開始点がそろっている傾向が観察された。一方、CpG islands のみがある遺伝子群では、転写開始点のゆらぎが観察され、14~18 塩基のところに分布のピークがあった（表 5-3、図 5-11）。

表 5-3　TATA と CpG island 有無集計

-	human			mouse		
	TATA+	TATA-	total	TATA+	TATA-	total
CpG+	516	6084	6600	348	2600	2948
CpG-	576	2043	2619	526	1304	1830
total	1092	8127	9219	874	3904	4778

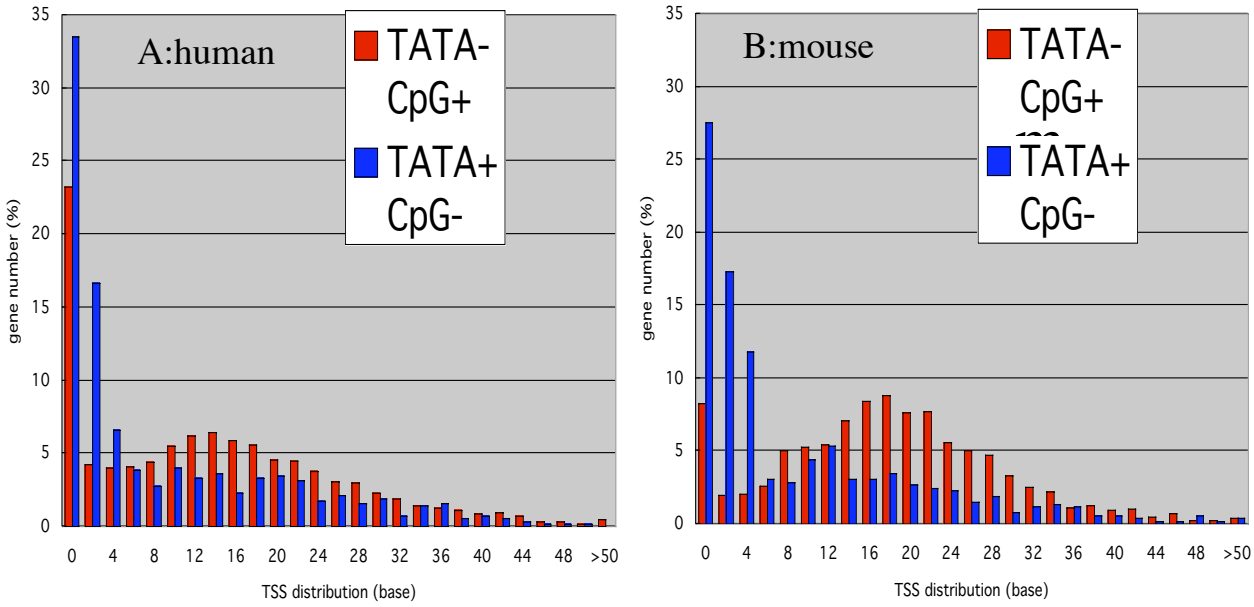


図 5-11　TATA と CpG 島を持つ遺伝子の TSS の揺らぎの差

プロモータ領域に CpG 島を持つ遺伝子と TATA を持つ遺伝子ごとに、転写開始点の揺らぎを見た。代表転写開始点から 100 塩基以内の転写開始点のみを対象とした。A:human、B:mouse

、

5.3.9 解析結果の公開

解析した結果は、全て、DBTSS のページから download のリンクをたどることによって入手できる。ftp://ftp.hgc.jp/pub/hgc/db/dbtss/Yamashita_et_alから、ftp 経由で直接ダウンロードすることも可能である。ダウンロード可能なデータ形式とその例は表 5-4 に示した。

5.4　考察

本研究の目的は、プロモータ領域に CpG island はあるのか、あるとしたら、

表 5-4　ダウンロードサイトのデータ形式

Contents of the files	example
NM_ID	NM_003532
splicing variants NM_ID	NM_003532
Chromosome	chr6
Chromosome Strand (0= forward/1= reverse)	0
Position of NM_ TSS	26281966
GC content of (−100:+100)	0.355
CpG score of (−100:+100)	0.325203252
CpG island (Y= Yes/ N= No)	N
LocusLink ID	8353
UniGene ID	Hs.143522
Number of sequences in UniGene	8
Number of libraries in UniGene	7
library IDs	4768,205,4665,5610,7085,186,1042,
each Number of clones	1,1,2,1,1,1,1,

組織特異性と関連があるのかを明確にすることである。まず、図 5-4 でプロモータ領域に顕著に CpG islanad が存在することがわかった。したがって、プロモータ領域に CpG があるという初期の観察(Gardiner-Garden et al, 1987; Larsen et al., 1992)は正しいことが示唆される。続いて、遺伝子発現の組織特異性と CpG islands には相関があるのか検討した。図 5-5 を見るとプロモータ領域のみの分布が偏っており相関（図 5-5 A,E:ヒト 0.27、マウス 0.41）が認められる。しかし、スプライシング後の mRNA レベル（図 5-5: B,F ヒト 9.5E-0.5、マウス 0.014)、DNA レベル（図 5-5 ： C,G ヒト 0.020、マウス 0.047)、では、GC 含量と組織特異性には相関は見いだされない。これは明らかに、プロモータ領域のみの GC 含量が発現の組織特異性と関係があるためであると考えられる。この傾向は CpG score で見た時も同様であった。なお、3rd ポジションの GC 含量と組織特異性の関係は、ヒトとマウスで差があるという報告(Vinogradov, 2003)は、今回の私のデータでも観察された（図 5-5 D,H)。

次ぎに、図 5-6 によりプロモータ領域に CpG がある遺伝子群（CpG+）とない遺伝子群（CpG-）に分けた。プロモータ領域の CpG islands を調べた結果、

ヒトで CpG がある遺伝子は、6600（71.6%）マウスで 2948（61.7%）見つかった。またこの分類を行った後に組織特異性を見てみると、明らかに CpG+の遺伝子群と CpG-の遺伝子群では分布が異なり、CpG-の遺伝子群では組織特異性が高い（図 5-7）。この傾向は、ヒト・マウスに同様に認められ、高い有為差（$p<10^{-100}$）を示した。さらに、表 5-2 の Gene ontology 単語 による解析においても、CpG+群には housekeeping 遺伝子を示唆する単語が、CpG-群には組織特異性を示唆する単語が濃縮されてきていた。これらの結果は、プロモータ領域の CpG islands と組織特異性に関係があることを強く示唆する。

ヒトに関して染色体バンド別に組織特異性を比較すると、最も組織特異性の差が小さいバンドの濃さ 100 においても、CpG+遺伝子群と CpG-遺伝子群で有為差（10^{-11}）が検出された。従って、バンドの濃さよりも CpG islands の有無の方が発現に強く影響すると考えられる。しかし、CpG+遺伝子群では、バンドが濃くなるに連れて発現が下がっていく傾向が確認されたが（$p<10^{-9}$）、CpG-群では観察されなかったこと（P<0.13）は、さらなる検討が必要であろう。

ヒト 9219 遺伝子、マウス 4778 遺伝子のうち、オーソログであったものは 1472 遺伝子あった。このうち 1283 遺伝子（87.2%）が、CpG island の有無に関して挙動が一致した（表 5-1）。したがって、ヒトとマウスの CpG islands を持つ遺伝子群は保存されていることが示唆される。また、オーソログ同士で組織特異性の検討をした結果、両者は高い相関（R=0.72）を示した。従って CpG islands を介した発現様式も、両生物種で保存されていると考えられる。

では、CpG island のある遺伝子はなぜユビキタスに発現するのだろうか。1 つの仮説として考えられるのは、CpG islands 自身が、単にメチル化を受けていないから T が蓄積せずに生じた(Gardiner-Garden and M., 1987; Larsen et al., 1992) ものではなく、積極的に転写に関わっている可能性である。図 5-11 は、CpG island のみがある遺伝子群と TATA-box のみある遺伝子群に分け、転写開始点の標準偏差のバラつきを見たものである。これによると、CpG islands のみを持つ遺伝子は、転写開始点の揺らぎが大きい。これは、CpG island がある遺伝子は、一つのプロモータ内でより多くの塩基から転写を始めることが考えられる。つまり、CpG のある遺伝子と TATA-box を持つ遺伝子の制御機構は、異なる可能性が考えられる。

本研究に関していくつかの問題点を述べる必要がある。まず、代表転写開始点の設定の仕方である。今回最頻値か中央値を代表転写開始点と設定した。し

かし、alternative promoter がある遺伝子などは、その遺伝子の代表をとってきていることにはならない。特に UniGene の発現情報と合わせた時に、異なる転写開始点に基づいた発現情報を取ってきているというミスを含む可能性がある。続いて CpG islands の定義である。CpG islands の定義に関しては、Takai ら(Takai and Jones, 2002)が異なる指標（GC>0.55、CpG>0.65）を提案している。今回、この基準でも CpG の分類を試みたが、最終的な発現量との相関を見た時に、分離能が悪かった。この CpG island の定義を、例えは組織特異性の分離を最大にして最適化するようなことは、今後の研究として考えられる。最後に、UniGene を組織特異性の指標として使うことの問題である。多量の規格化された EST 配列を発現量の指標に使うことは、BodyMap という手法で知られている(Okubo et al., 1992)。しかし、UniGene は GeneBank にある配列 cDNA 配列を機械的にクラスタリングしたものである。配列決定された数は、ライブラリーによって大きく異なり、そのバイアスを含んでいることも十分に考えられる。しかしこれにもかかわらず、CpG+の遺伝子群と CpG-の遺伝子群で組織特異性の差が十分に識別できたので、本研究の手法が有効であったと考えられる。

6　謝辞

本研究を行うに当たって、指導教官の小笠原直毅教授に感謝致します。そして、貴重な 5'端配列を提供してくださった、東大医科研ゲノム構造解析の菅野純夫教授、鈴木穣助教授に感謝いたします。その他、共同研究者の Edgar Wingender, Alexander Kel には、TRANSFAC の使用について様々なアドバイスをいただきました。DBTSS の構築に当たっては、様々なわがままを聞いてくださった、株式会社ダイナコムの若栗浩幸様のご尽力なしには語れません。また、松原謙一先生を初めとする奈良先端科学技術大学院大学大正製薬ゲノム機能解析講座の皆様には、分子生物学の基礎を含め貴重な指導を賜りました。特に私をバイオインフォマティックスの分野に進むことを勧めてくださった加藤菊也教授には改めて謝意を申し上げます。また、貴重な研究の場を与えてくれた、現在東大新領域創成科学の高木利久教授を初めとする、東大医科研ゲノムデータベース分野・機能解析イン・シリコの皆様に感謝いたします。最後になりましたがバイオインフォマティックスの基礎からの指導、全ての研究においての有用な助言、また時には叱咤激励までも頂いた機能解析イン・シリコの中井謙太教授に厚く御礼申し上げます。

7. 参考文献

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F.*, et al*. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Ayoubi, T. A., and Van De Ven, W. J. (1996). Regulation of gene expression by alternative promoters. Faseb J *10*, 453-460.

Bajic, V. B., Tan, S. L., Suzuki, Y., and Sugano, S. (2004). Promoter prediction analysis on the whole human genome. Nat Biotechnol *22*, 1467-1473.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank: update. Nucleic Acids Res *32*, D23-26.

Bernardi, G. (1995). The human genome: organization and evolutionary history. Annu Rev Genet *29*, 445-476.

Burke, T. W., and Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev *10*, 711-724.

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M.*, et al*. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics *37*, 327-336.

Carninci, P., Westover, A., Nishiyama, Y., Ohsumi, T., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., Schneider, C., and Hayashizaki, Y. (1997). High efficiency selection of full-length cDNA by improved biotinylated cap trapper. DNA Res *4*, 61-66.

CESC (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. Science *282*, 2012-2018.

D'Onofrio, G. (2002). Expression patterns and gene distribution in the human genome. Gene *300*, 155-160.

Dreyfuss, G., Matunis, M. J., Pinol-Roma, S., and Burd, C. G. (1993). hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem *62*, 289-321.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., and et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res *8*, 967-974.

Ford, C. L., Randal-Whitis, L., and Ellis, S. R. (1999). Yeast proteins related to the p40/laminin receptor precursor are required for 20S ribosomal RNA processing and the maturation of 40S ribosomal subunits. Cancer Res *59*, 704-710.

Furey, T. S., and Haussler, D. (2003). Integration of the cytogenetic map with the draft human genome sequence. Hum Mol Genet *12*, 1037-1044.

Gachet, Y., Tournier, S., Lee, M., Lazaris-Karatzas, A., Poulton, T., and Bommer, U. A. (1999). The growth-related, translationally controlled protein P23 has properties of a tubulin binding protein and associates transiently with microtubules during the cell cycle. J Cell Sci *112*, 1257-1271.

Gardiner-Garden, M., and M., F. (1987). CpG Islands in Vertebrate Genomes. J Mol Biol *196*, 261-282.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., *et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Goncalves, I., Duret, L., and Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. Genome Res *10*, 672-678.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al.* (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res *32*, D258-261.

Huang, X. Q., and Miller, W. (1991). A space-efficient algorithm for local similarities. Adv Appl Math *12*, 337-357.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., *et al.* (2005). Ensembl 2005. Nucleic Acids Res *33*, D447-453.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al.* (2002). The Ensembl genome database project. Nucleic Acids Res *30*, 38-41.

IHGSC (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., *et al.* (2004). Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol *2*, e162. Epub 2004 Apr 2020.

Izaurralde, E., Jarmolowski, A., Beisel, C., Mattaj, I. W., Dreyfuss, G., and Fischer, U. (1997). A role for the M9 transport signal of hnRNP A1 in mRNA nuclear export. J Cell Biol *137*, 27-35.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al.* (2001). Functional annotation of a full-length mouse cDNA collection. Nature *409*, 685-690.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kraus, R. J., Murray, E. E., Wiley, S. R., Zink, N. M., Loritz, K., Gelembiuk, G. W., and Mertz, J. E. (1996). Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. Nucleic Acids Res *24*, 1531-1539.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev *12*, 34-44.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

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.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095-1107.

Loreni, F., and Amaldi, F. (1997). Translational control of terminal oligopyrimidine mRNAs requires a specific regulator. FEBS Lett *416*, 239-242.

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. (2000). NCBI's LocusLink and RefSeq. Nucleic Acids Res *28*, 126-128.

Maruyama, K., and Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene *138*, 171-174.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.* (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res *31*, 374-378.

Meyuhas, O. (2000a). Synthesis of the translational apparatus is regulated at the translational level. Eur J Biochem *267*, 6321-6330.

Meyuhas, O., Avni, D., and Hornstein, E. (1996). Translational control of Ribosomal Protein mRNA in Eukaryotes. (Cold Spring Harbor, New York., Cold Spring Harbor Laboratory Press).

Meyuhas, O., Avni, D., Shama, S. (1996). Translational control of ribosomal protein mRNAs in eukaryotes. (Cold Spring Harbor, New York., Cold Spring Harbor Laboratory Press).

Meyuhas, O. H., E. (2000b). Translational control of TOP mRNAs. (Cold Spring Harbor, New York., Cold Spring Harbor Laboratory Press).

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.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara, K. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nat Genet *2*, 173-179.

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. Epub 2003 Dec 2021.

Ponger, L., Duret, L., and Mouchiroud, D. (2001). Determinants of CpG islands:

expression in early embryo and isochore structure. Genome Res *11*, 1854-1860.

Praz, V., Perier, R., Bonnard, C., and Bucher, P. (2002). The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res *30*, 322-324.

Pruitt, K. D., and Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res *29*, 137-140.

Schmidt, M. C., Kao, C. C., Pei, R., and Berk, A. J. (1989). Yeast TATA-box transcription factor gene. Proc Natl Acad Sci U S A *86*, 7785-7789.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Shibui-Nihei, A., Ohmori, Y., Yoshida, K., Imai, J., Oosuga, I., Iidaka, M., Suzuki, Y., Mizushima-Sugano, J., Yoshitomo-Nakagawa, K., and Sugano, S. (2003). 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*, 137-145.

Stolovich, M., Tang, H., Hornstein, E., Levy, G., Cohen, R., Bae, S. S., Birnbaum, M. J., and Meyuhas, O. (2002). Transduction of growth or mitogenic signals into translational activation of TOP mRNAs is fully reliant on the phosphatidylinositol 3-kinase-mediated pathway but requires neither S6K1 nor rpS6 phosphorylation. Mol Cell Biol *22*, 8101-8113.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S.*, et al.* (2001a). Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites. EMBO Rep *2*, 388-393.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y.*, et al.* (2001b). Identification and characterization

of the potential promoter regions of 1031 kinds of human genes. Genome Res *11*, 677-684.

Suzuki, Y., Yamashita, R., Nakai, K., and Sugano, S. (2002). DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res *30*, 328-331.

Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K. (2004). DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. Nucleic Acids Res *32*, D78-81.

Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S. (1997). Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. Gene *200*, 149-156.

Takai, D., and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740-3745.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tohgo, A., Takasawa, S., Munakata, H., Yonekura, H., Hayashi, N., and Okamoto, H. (1994). Structural determination and characterization of a 40 kDa protein isolated from rat 40 S ribosomal subunit. FEBS Lett *340*, 133-138.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vinogradov, A. E. (2003). Isochores and tissue-specificity. Nucleic Acids Res *31*, 5212-5220.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P*., et al*. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I., and Schacherer, F. (2000). TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res *28*, 316-319.

Yoshihama, M., Uechi, T., Asakawa, S., Kawasaki, K., Kato, S., Higa, S., Maeda, N., Minoshima, S., Tanaka, T., Shimizu, N., and Kenmochi, N. (2002). The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. Genome Res *12*, 379-390.

Zatsepina, O. V., Rousselet, A., Chan, P. K., Olson, M. O., Jordan, E. G., and Bornens, M. (1999). The nucleolar phosphoprotein B23 redistributes in part to the spindle poles during mitosis. J Cell Sci *112*, 455-466.

補足表 1　検出された TOP 遺伝子

NM_ID	chr	strand	TSS	-4~+7	definition
NM_130781	chr5	-	177573224	CGCCCTCTAGC	(RAB24)
NM_000859	chr5	+	74869542	GCTCCTTCCGC	3-hydroxy-3-methylglutaryl-Coenzyme A reductase (HMGCR)
NM_002150	chr12	+	121808730	AGGCCTCTAGT	4-hydroxyphenylpyruvate dioxygenase (HPD)
NM_012255	chr20	+	21231983	CCGTCTCTTTG	5'-3' exoribonuclease 2 (XRN2)
NM_016630	chr15	-	58381536	CGGCCTCCCGC	acid cluster protein 33 (ACP33)
NM_006401	chr9	+	92489156	CCCCCTTTTCC	acidic (leucine-rich) nuclear phosphoprotein 32 family, member B (ANP32B)
NM_006409	chr7	+	97458096	CGCTCCCTCTG	actin related protein 2/3 complex, subunit 1A, 41kDa (ARPC1A)
NM_003664	chr5	-	77827099	GAACCTTTTGG	adaptor-related protein complex 3, beta 1 subunit (AP3B1)
NM_012095	chr10	-	75013983	CGGCCTTCTCG	adaptor-related protein complex 3, mu 1 subunit (AP3M1)
NM_004722	chr7	+	98233841	CGTTCTTTTGT	adaptor-related protein complex 4, mu 1 subunit (AP4M1)
NM_000034	chr16	+	30399651	GTTCCTCTCGG	aldolase A, fructose-bisphosphate (ALDOA)
NM_016608	chrX	+	97776670	CGTCCTTCTAA	ALEX1 protein (ALEX1)
NM_000014	chr12	-	9416514	GCTCCTTCTTT	alpha-2-macroglobulin (A2M)
NM_012103	chr2	-	74969619	CCGCCTTCCCA	ancient ubiquitous protein 1 (AUP1)
NM_017664	chr13	-	105954539	CGTTCCTTTGT	ankyrin repeat domain 10 (ANKRD10)
NM_001641	chr14	+	14710851	CCTTCTTTGTG	APEX nuclease (multifunctional DNA repair enzyme) 1 (APEX1), transcript variant 1
NM_000384	chr2	-	21321002	AGTTCTCTGTA	apolipoprotein B (including Ag(x) antigen) (APOB)
NM_005736	chr10	-	103496069	GTTCCTTCCCC	ARP1 actin-related protein 1 homolog A, centractin alpha (yeast) (ACTR1A)
NM_014062	chr16	-	70209266	TCCCCTCTCAC	ART-4 protein (ART-4)
NM_004539	chr18	-	55263668	CGCTCTCTGAT	asparaginyl-tRNA synthetase (NARS)
NM_001349	chr2	-	135040840	CGATCTTTCTG	aspartyl-tRNA synthetase (DARS)
NM_005176	chr12	-	54182190	CTGTCTTCTCT	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit c (subunit 9), isoform 2 (ATP5G2)
NM_006476	chr11	+	119784135	GGTCCTTCCGG	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit g (ATP5L)
NM_001183	chrX	+	150041293	CCCCCTCCTCA	ATPase, H+ transporting, lysosomal interacting protein 1 (ATP6IP1)
NM_000701	chr1	+	115812292	GATTCTTTGTT	ATPase, Na+/K+ transporting, alpha 1 polypeptide (ATP1A1)
NM_000702	chr1	+	155817812	CTTTCTCTGTC	ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide (ATP1A2)
NM_001207	chr5	+	73027645	CTCCCTTTAGC	basic transcription factor 3 (BTF3)
NM_001731	chr12	-	92999052	GCATCTCTTCG	B-cell translocation gene 1, anti-proliferative (BTG1)
NM_032667	chr11	-	64050491	CCTCCTTTCCT	Bernardinelli-Seip congenital lipodystrophy 2 (seipin) (BSCL2)
NM_001711	chrX	+	149215361	CTCCCTCTCTC	biglycan (BGN)
NM_000386	chr17	-	30525413	TTTCCTTTTTC	bleomycin hydrolase (BLMH)

NM_007236	chr15	+	34419331	CTTCCTTCCCT	calcium binding protein P22 (CHP)
NM_001745	chr5	+	134588490	CGGCCTCTAGT	calcium modulating ligand (CAMLG)
NM_001744	chr5	+	111005835	CTCTCTCTCGC	calcium/calmodulin-dependent protein kinase IV (CAMK4)
NM_001746	chr5	+	180070067	AGGCCTCTTGG	calnexin (CANX)
NM_005186	chr11	+	66631010	CGCTCTTCCTG	calpain 1, (mu/I) large subunit (CAPN1)
NM_000070	chr15	+	35547525	CACTCTCTTTC	calpain 3, (p94) (CAPN3), transcript variant 1
NM_006371	chr3	+	32434964	CTTCCTTTCGC	cartilage associated protein (CRTAP)
NM_001892	chr5	-	149525447	ATCCCTTTCCC	casein kinase 1, alpha 1 (CSNK1A1)
NM_001774	chr19	+	50206849	CTTTCTTTCTC	CD37 antigen (CD37)
NM_000610	chr11	+	35929036	TACTCTTTTTG	CD44 antigen (homing function and Indian blood group system) (CD44)
NM_000560	chr1	+	110282327	TCTCCTTTTAC	CD53 antigen (CD53)
NM_001251	chr17	+	8222669	CCTCCTTTCCA	CD68 antigen (CD68)
NM_032025	chr3	+	151149518	GTTTCTCTTTC	CDA02 protein (CDA02)
NM_016315	chr2	+	187840629	GCTTCTTCTGG	CED-6 protein (CED-6)
NM_031942	chr2	+	172880763	GCTCCTCCTGC	cell division cycle associated 7 (CDCA7), transcript variant 1
NM_015416	chr12	-	51648578	CAACCTCTTCT	cervical cancer 1 protooncogene (DKFZP586A011)
NM_015938	chr3	+	161898606	TCTTCTCTGTG	CGI-07 protein (CGI-07)
NM_016057	chr12	+	54832424	GTTTCTTTTGC	CGI-120 protein (COPZ1)
NM_016079	chr2	-	87084631	CCTCCTTTTCC	CGI-149 protein (CGI-149)
NM_015958	chr1	-	100672414	GGGCCTTTTCT	CGI-30 protein (CGI-30)
NM_015380	chr22	+	40968136	CCGCCTTCTGC	CGI-51 protein (CGI-51)
NM_016025	chr16	+	21619681	CAGCCTTTGCC	CGI-81 protein (DREV1)
NM_016039	chr14	+	46252240	CGCCCTCTCGC	CGI-99 protein (CGI-99)
NM_005998	chr1	-	152073263	GGTTCTCTCTC	chaperonin containing TCP1, subunit 3 (gamma) (CCT3)
NM_006430	chr2	-	62301964	CCCCCTTCTCC	chaperonin containing TCP1, subunit 4 (delta) (CCT4)
NM_006585	chr21	-	27106435	CTTCCTCCGCG	chaperonin containing TCP1, subunit 8 (theta) (CCT8)
NM_001293	chr11	-	78887870	CTGCCTCTTCC	chloride channel, nucleotide-sensitive, 1A (CLNS1A)
NM_004385	chr5	+	82994535	GAGCCTTTCTG	chondroitin sulfate proteoglycan 2 (versican) (CSPG2)
NM_005441	chr21	+	34418162	GTGCCTCTGAC	chromatin assembly factor 1, subunit B (p60) (CHAF1B)
NM_015039	chr1	-	178815816	CTTCCTTTCTC	chromosome 1 open reading frame 15 (C1orf15), transcript variant 1
NM_006820	chr1	+	77985429	CTTTCTTTCCT	chromosome 1 open reading frame 29 (C1orf29)
NM_022067	chr14	-	71740943	CCTTCTCTAAG	chromosome 14 open reading frame 133 (C14orf133)
NM_004894	chr14	-	98203242	TGACCTTTCCG	chromosome 14 open reading frame 2 (C14orf2)
NM_021944	chr14	-	17266653	CCTCCTTTTTG	chromosome 14 open reading frame 93 (C14orf93)
NM_017815	chr14	-	17213600	GGGCCTTTCAG	chromosome 14 open reading frame 94 (C14orf94)
NM_016304	chr15	-	48536196	CTTCCTCTCAA	chromosome 15 open reading frame 15 (C15orf15)

NM_019025	chr20	+	4100357	GGTTCTTTCTG	chromosome 20 open reading frame 16 (C20orf16), transcript variant 5
NM_033542	chr20	+	43680224	GTTTCTTTCCT	chromosome 20 open reading frame 169 (C20orf169)
NM_006462	chr20	+	336975	CTCCCCTTGCG	chromosome 20 open reading frame 18 (C20orf18), transcript variant 1
NM_003678	chr22	-	26645830	CCTTCCTTACT	chromosome 22 open reading frame 19 (C22orf19)
NM_016947	chr6_random	+	7564019	AGTTCTTTTTG	chromosome 6 open reading frame 48 (C6orf48)
NM_015524	chr6	-	111949942	TGTTCTTCTAC	chromosome 6 open reading frame 5 (C6orf5)
NM_001269	chr1	+	27791648	CTCTCCTTTTT	chromosome condensation 1 (CHC1)
NM_004077	chr12	+	56603052	CCTCCTTTCAA	citrate synthase (CS), nuclear gene encoding mitochondrial protein
NM_004859	chr17	+	60206323	TCTTCTTTAGG	clathrin, heavy polypeptide (Hc) (CLTC)
NM_016207	chr2	+	9615175	CTTCCTTTTTT	cleavage and polyadenylation specific factor 3, 73kDa (CPSF3)
NM_007006	chr16	-	56536893	CCTCCTCTTGC	cleavage and polyadenylation specific factor 5, 25 kDa (CPSF5)
NM_015526	chr19	-	36972731	CTCCCTCTCCG	CLIP-170-related protein (CLIPR-59)
NM_001280	chr19	+	1338623	CCCCCCCTCAC	cold inducible RNA binding protein (CIRBP)
NM_003653	chr17	-	18543759	CTGCCTTCGCC	COP9 constitutive photomorphogenic homolog subunit 3 (Arabidopsis) (COPS3)
NM_005776	chr14	-	48704124	GCTCCTCCTCC	cornichon homolog (Drosophila) (CNIH)
NM_004376	chr10	-	100725526	GCTTCTCTTTT	COX15 homolog, cytochrome c oxidase assembly protein (yeast) (COX15), nuclear gene encoding mitochondrial protein, transcript variant 2
NM_020990	chr15	+	36782012	GGGCCTCCCTC	creatine kinase, mitochondrial 1 (ubiquitous) (CKMT1), nuclear gene encoding mitochondrial protein
NM_005190	chr6	-	100033245	CTTCCTTTCGC	cyclin C (CCNC)
NM_004060	chr5	+	163457909	CGGCCCCTTCG	cyclin G1 (CCNG1)
NM_006835	chr4	-	78187322	CTTCCCCTCCC	cyclin I (CCNI)
NM_000075	chr12	-	58289288	GCCTCTCTAGC	cyclin-dependent kinase 4 (CDK4), transcript variant 1
NM_001861	chr16	+	86866441	TGCTCTCTTCC	cytochrome c oxidase subunit IV isoform 1 (COX4I1), nuclear gene encoding mitochondrial protein
NM_004374	chr8	-	100979784	TTTTCCTTTAG	cytochrome c oxidase subunit VIc (COX6C), nuclear gene encoding mitochondrial protein
NM_004718	chr2	-	42757101	GGTCCTTCTCT	cytochrome c oxidase subunit VIIa polypeptide 2 like (COX7A2L), nuclear gene encoding mitochondrial protein
NM_001867	chr5	+	86101790	CTTTCTTTTCA	cytochrome c oxidase subunit VIIc (COX7C), nuclear gene encoding mitochondrial protein
NM_021227	chr4	+	109892876	CGGCCCTTGCT	DC2 protein (DC2)
NM_004728	chr10	+	69802661	ACCTCTTCCTC	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 21 (DDX21)
NM_004632	chr1	+	151406485	GACCCTTTTTT	death associated protein 3 (DAP3), nuclear gene encoding mitochondrial protein, transcript variant 2
NM_006360	chr11_random	-	589613	GTTCCCTTTTC	dendritic cell protein (GA17)
NM_001927	chr2	+	219004475	CTGTCTCCCCT	desmin (DES)
NM_006870	chr20	+	17498764	GGGTCTCTCGG	destrin (actin depolymerizing factor) (DSTN)
NM_001386	chr8	+	26878171	CTCTCTCTTTT	dihydropyrimidinase-like 2 (DPYSL2)

NM_024740	chr11	-	113254381	AATTCTTTTTT	disrupted in bipolar disorder 1 (DIBD1)
NM_015412	chr3	-	112309040	CCGCCTTTCGT	DKFZP434F2021 protein (DKFZP434F2021)
NM_015654	chr17	-	75830363	CACCCTTTCTG	DKFZP564C103 protein (DKFZP564C103)
NM_015469	chr9	+	99253502	GCTCCTTTCCA	DKFZp564D177 protein (DKFZp564D177)
NM_015388	chr6	-	43481084	TCTCCTTTTTG	DKFZP566C243 protein (DKFZP566C243)
NM_015528	chr17	+	5181886	CGTCCCCTTTC	DKFZP566H073 protein (DKFZP566H073)
NM_018431	chr20	+	52780516	CCTCCTTCTCG	docking protein 5 (DOK5)
NM_005216	chr1	-	20023374	GGGTCCTTCGG	dolichyl-diphosphooligosaccharide-protein glycosyltransferase (DDOST)
NM_004417	chr5	-	172898616	TGCCCTTTCTG	dual specificity phosphatase 1 (DUSP1)
NM_018234	chr2	+	117902167	CCGCCTTCGCC	dudulin 2 (TSAP6)
NM_006400	chr12	-	58084146	CGCTCCCTTTG	dynactin 2 (p50) (DCTN2)
NM_001378	chr2	+	171085458	AGTTCTTCTCG	dynein, cytoplasmic, intermediate polypeptide 2 (DNCI2)
NM_004433	chr1	+	197413557	CTCCCTCCAGG	E74-like factor 3 (ets domain transcription factor, epithelial-specific) (ELF3)
NM_014390	chr7	+	125769837	GCGTCTCTTTC	EBNA-2 co-activator (100kD) (p100)
NM_014210	chr17	-	31555579	TATCCTTTTTT	ecotropic viral integration site 2A (EVI2A)
NM_006495	chr17	-	31548011	TTTCCTTTCTT	ecotropic viral integration site 2B (EVI2B)
NM_004105	chr2	-	56337144	TCTCCTCCTCC	EGF-containing fibulin-like extracellular matrix protein 1 (EFEMP1), transcript variant 1
NM_000501	chr7	+	72082672	CTCCCTCCCTC	elastin (supravalvular aortic stenosis, Williams-Beuren syndrome) (ELN)
NM_018255	chr18	+	33529277	GCGTCTCTTGT	elongator protein 2 (ELP2)
NM_001397	chr1	-	20583769	AGCTCTTCTTC	endothelin converting enzyme 1 (ECE1)
NM_001975	chr12	-	7035474	CTTTCTCCTTC	enolase 2, (gamma, neuronal) (ENO2)
NM_016262	chr6	-	112431580	AGCTCTCTAGC	epsilon-tubulin (TUBE)
NM_018538	chr1	+	42286354	TGCCCTCTTCC	erythroblast membrane-associated protein (ERMAP)
NM_001402	chr6	-	74197456	CGTTCTTTTTC	eukaryotic translation elongation factor 1 alpha 1 (EEF1A1)
NM_001958	chr20	-	61967856	CAGTCCCTCTG	eukaryotic translation elongation factor 1 alpha 2 (EEF1A2)
NM_001959	chr2	+	205749235	GGTCCTTTTTC	eukaryotic translation elongation factor 1 beta 2 (EEF1B2), transcript variant 1
NM_001960	chr8	-	144907582	CCCTCCCTTTC	eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein) (EEF1D), transcript variant 2
NM_001404	chr7	-	131063238	CAGCCTTTCTT	eukaryotic translation elongation factor 1 gamma (EEF1G)
NM_001961	chr19	-	4054739	CGTTCTCTTCC	eukaryotic translation elongation factor 2 (EEF2)
NM_003908	chr20	-	32418717	TTTCCTTTCGC	eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa (EIF2S2)
NM_001415	chrX	+	22719864	CTTCCTCTTTT	eukaryotic translation initiation factor 2, subunit 3 gamma, 52kDa (EIF2S3)
NM_013234	chr19	+	39501051	CCACCTCTTCC	eukaryotic translation initiation factor 3 subunit k (eIF3k)
NM_003750	chr10	-	120084013	CTTCCTTTCCG	eukaryotic translation initiation factor 3, subunit 10 theta, 150/170kDa (EIF3S10)
NM_003757	chr1	+	31675668	AAACCTTTTCC	eukaryotic translation initiation factor 3, subunit 2 beta, 36kDa (EIF3S2)

NM_003756	chr8	-	117829246	GTTTCTCTTTC	eukaryotic translation initiation factor 3, subunit 3 gamma, 40kDa (EIF3S3)
NM_003754	chr11	+	8469321	CCTTCTTTCTC	eukaryotic translation initiation factor 3, subunit 5 epsilon, 47kDa (EIF3S5)
NM_001568	chr8	-	109328278	CTCCCTTTTCT	eukaryotic translation initiation factor 3, subunit 6 48kDa (EIF3S6)
NM_016091	chr22	+	34859957	CGCTCTTTCCG	eukaryotic translation initiation factor 3, subunit 6 interacting protein (EIF3S6IP)
NM_003753	chr22	-	33539744	TTTCCTCTTTT	eukaryotic translation initiation factor 3, subunit 7 zeta, 66/67kDa (EIF3S7)
NM_003752	chr16	-	28573508	CCTTCTCTCTC	eukaryotic translation initiation factor 3, subunit 8, 110kDa (EIF3S8)
NM_001967	chr3	+	187758448	CTGTCTTTTCA	eukaryotic translation initiation factor 4A, isoform 2 (EIF4A2)
NM_001417	chr12	+	53313785	CGTTCTCTTTC	eukaryotic translation initiation factor 4B (EIF4B)
NM_004629	chr9	-	35249382	CCACCCTTTCT	Fanconi anemia, complementation group G (FANCG)
NM_003902	chr1	-	77391212	TTTTCTTTCTT	far upstream element (FUSE) binding protein 1 (FUBP1)
NM_004462	chr8	+	11645862	CTGCCTTTATG	farnesyl-diphosphate farnesyltransferase 1 (FDFT1)
NM_004458	chrX	-	106123242	GCTCCTCCTCG	fatty-acid-Coenzyme A ligase, long-chain 4 (FACL4), transcript variant 1
NM_002032	chr11	-	63310721	CGTTCTTCGCC	ferritin, heavy polypeptide 1 (FTH1)
NM_001436	chr19	-	40728175	CGCTCTTTTCC	fibrillarin (FBL)
NM_001458	chr7	+	126947530	CCGCCCTTCCC	filamin C, gamma (actin binding protein 280) (FLNC)
NM_001997	chr11	-	66571214	CTTCCTCTTTC	Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed (fox derived); ribosomal protein S30 (FAU)
NM_021996	chr9	-	127596159	CTTCCTCTTTT	Forssman glycolipid synthetase (FS)
NM_005087	chr3	+	181474166	CGGCCTTTGCG	fragile X mental retardation, autosomal homolog 1 (FXR1)
NM_001494	chr10	-	5819683	AGTTCTTCTCT	GDP dissociation inhibitor 2 (GDI2)
NM_004128	chr13	+	39681121	GTTCCTCTTTT	general transcription factor IIF, polypeptide 2 (30kD subunit) (GTF2F2)
NM_001517	chr6_random	+	6639830	CCTTCTCTTCT	general transcription factor IIH, polypeptide 4 (52kD subunit) (GTF2H4)
NM_020194	chr2	+	226911060	GCGCCTTTCGC	GL004 protein (GL004)
NM_015710	chr19	+	48639690	CTTCCTTTGAC	glioma tumor suppressor candidate region gene 2 (GLTSCR2)
NM_000175	chr19	+	35305070	CTTCCTCCTCG	glucose phosphate isomerase (GPI)
NM_000817	chr2	+	170214891	CTCTCTTCTCC	glutamate decarboxylase 1 (brain, 67kDa) (GAD1), transcript variant GAD67
NM_015532	chr15	+	51045977	CGTTCTTCCGG	glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A (GRINL1A)
NM_005051	chr3	+	48503778	GTTTCTTTTAG	glutaminyl-tRNA synthetase (QARS)
NM_004446	chr1	-	216063239	CTTCCTTTCGC	glutamyl-prolyl-tRNA synthetase (EPRS)
NM_002085	chr19	+	1173352	CCGCCTTTGCC	glutathione peroxidase 4 (phospholipid hydroperoxidase) (GPX4)
NM_002046	chr12	+	6618146	CGCTCTCTGCT	glyceraldehyde-3-phosphate dehydrogenase (GAPD)
NM_002047	chr7	+	30276786	CACCCTCTCTG	glycyl-tRNA synthetase (GARS)
NM_004484	chrX	-	129965361	ACGTCTCTTGC	glypican 3 (GPC3)
NM_005708	chr13	+	88266243	CCTCCTTTCTC	glypican 6 (GPC6)
NM_000405	chr5	+	151245149	TTTCCTTTTTG	GM2 ganglioside activator protein (GM2A)
NM_030799	chr5	-	144132796	CGTTCTTTGGC	golgi membrane protein SB140 (SMAP-5)

NM_002087	chr17	+	44605054	GTGCCTTCTGC	granulin (GRN)
NM_014394	chr10	+	85121029	CGTCCTTTCGA	growth hormone inducible transmembrane protein (GHITM)
NM_006496	chr1	+	109020577	GGTTCTTCTGG	guanine nucleotide binding protein (G protein), alpha inhibiting activity polypeptide 3 (GNAI3)
NM_006098	chr5	+	181694846	CTCTCTTTCAC	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1 (GNB2L1)
NM_006644	chr13	-	25722668	TCCCCTTTTGG	heat shock 105kDa/110kDa protein 1 (HSPH1)
NM_002156	chr2	-	197077221	CTGTCCCTCAC	heat shock 60kDa protein 1 (chaperonin) (HSPD1)
NM_006597	chr11	-	124444707	GGCCCTTTATG	heat shock 70kDa protein 8 (HSPA8), transcript variant 1
NM_007355	chr6	+	44211252	AGCTCTCTCGA	heat shock 90kDa protein 1, beta (HSPCB)
NM_002136	chr12	+	54788046	GCTCCTTTCTG	heterogeneous nuclear ribonucleoprotein A1 (HNRPA1), transcript variant 1
NM_005463	chr4	-	83644397	GGATCTCTTCC	heterogeneous nuclear ribonucleoprotein D-like (HNRPDL), transcript variant 1
NM_004966	chr10	-	43373094	CGTCCTTCCGG	heterogeneous nuclear ribonucleoprotein F (HNRPF)
NM_000521	chr5	+	74213874	CTTCCTCTGAT	hexosaminidase B (beta polypeptide) (HEXB)
NM_002131	chr6	+	34201065	CGCTCTTTTTA	high mobility group AT-hook 1 (HMGA1), transcript variant 2
NM_002143	chr1	+	32307024	CCGCCTTCCCT	hippocalcin (HPCA)
NM_006895	chr2	+	137029690	CTGTCTTTCTC	histamine N-methyltransferase (HNMT)
NM_005340	chr5	-	130948219	GAGCCTCTCCT	histidine triad nucleotide binding protein 1 (HINT1)
NM_033445	chr1	-	224384167	TGCCCTCTTGT	histone 3, H2a (HIST3H2A)
NM_001527	chr6	-	114315197	CGGCCTCCTGA	histone deacetylase 2 (HDAC2)
NM_004640	chr6	-	31563651	TTCCCTCCTTC	HLA-B associated transcript 1 (BAT1), transcript variant 1
NM_019059	chr7	-	22504749	CCTCCTTTCCC	homolog of Tom7 (S. cerevisiae) (TOM7)
NM_021814	chr6	-	53210143	CTTCCTCTTCC	homolog of yeast long chain polyunsaturated fatty acid elongation enzyme 2 (HELO1)
NM_016096	chr8	-	102285963	CCTTCCTTTCC	HSPC038 protein (LOC51123)
NM_007312	chr3	+	49815845	GCTCCTTCCTC	hyaluronoglucosaminidase 1 (HYAL1), transcript variant 1
NM_000182	chr2	-	26560063	TGTCCTCTTCA	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit (HADHA)
NM_000183	chr2	+	26560368	CCGCCCCTTGG	hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), beta subunit (HADHB)
NM_002153	chr16	+	83029544	CTCCCTTCTTG	hydroxysteroid (17-beta) dehydrogenase 2 (HSD17B2)
NM_138425	chr12	-	7245194	TTTCCTTTCCG	hypothetical protein BC009925 (LOC113246)
NM_015702	chr2	+	148848063	CTTCCTTTGCC	hypothetical protein CL25022 (CL25022)
NM_152834	chr2	-	667406	CCTCCTCTGTG	hypothetical protein DKFZp434C1714 (DKFZp434C1714)
NM_031307	chr11	-	127285840	CTTCCTTTCTC	hypothetical protein FKSG32 (FKSG32)
NM_024573	chr6	+	151668890	CCGCCTCTGTT	hypothetical protein FLJ12910 (FLJ12910)

NM_139076	chr4	-	84700058	CGTCCTCTTGT	hypothetical protein FLJ13614 (FLJ13614)
NM_023079	chr17	+	49485471	GTACCTTTACA	hypothetical protein FLJ13855 (FLJ13855)
NM_024699	chr8	-	82681708	CCGCCCCTTAC	hypothetical protein FLJ14007 (FLJ14007)
NM_032795	chr11	-	127594290	CCGCCCTTCCT	hypothetical protein FLJ14494 (FLJ14494)
NM_017691	chr15	+	64281907	GGGTCTCTTTG	hypothetical protein FLJ20156 (FLJ20156)
NM_017706	chr5	+	140627521	CAGTCCTTCTC	hypothetical protein FLJ20195 (FLJ20195)
NM_017749	chr11	-	47493306	CCGTCTTCTTG	hypothetical protein FLJ20294 (FLJ20294)
NM_017836	chr3	-	125834003	CCGCCTCTTTC	hypothetical protein FLJ20473 (FLJ20473)
NM_017875	chr3	+	38658354	TAGCCTTCTTC	hypothetical protein FLJ20551 (FLJ20551)
NM_024612	chr17	+	60092544	TCGTCTTTCCC	hypothetical protein FLJ22060 (FLJ22060)
NM_024717	chr5	-	94881889	CAGCCTCTTTT	hypothetical protein FLJ22344 (FLJ22344)
NM_032226	chr9	+	37290013	GTCCCTCTACG	hypothetical protein FLJ22611 (FLJ22611)
NM_024715	chr5	+	134724313	CTTCCTCCGGC	hypothetical protein FLJ22625 (FLJ22625)
NM_024897	chr1	-	151983063	CTTCCTCCATC	hypothetical protein FLJ22672 (FLJ22672)
NM_025109	chr17	-	36757083	TGTTCTTCCCG	hypothetical protein FLJ22865 (FLJ22865)
NM_022761	chr11	+	113262420	GAACCTTTTTT	hypothetical protein FLJ23499 (FLJ23499)
NM_152316	chr11	+	31119497	CGTCCTCTCAG	hypothetical protein FLJ38968 (FLJ38968)
NM_153689	chr2	+	199500659	CGGCCTCTGAC	hypothetical protein FLJ38973 (FLJ38973)
NM_032492	chr3	+	9872161	AGTTCTCTTCA	hypothetical protein GL009 (GL009)
NM_016406	chr1	+	156856030	GTTTCTCTTGC	hypothetical protein HSPC155 (HSPC155)
NM_016467	chr2	-	189331378	TTTTCTCTGGC	hypothetical protein LOC51240 (LOC51240)
NM_020313	chr16	-	57540913	CCTCCTCTCGC	hypothetical protein LOC57019 (LOC57019)
NM_145049	chr5	+	159199647	CGGTCTCTCAG	hypothetical protein MGC10067 (MGC10067)
NM_032750	chr3	-	51261340	CGGCCTCTTCC	hypothetical protein MGC15429 (MGC15429)
NM_024106	chr19	-	9872130	CGTTCCTTTTG	hypothetical protein MGC2663 (MGC2663)
NM_152581	chrX	+	13716989	CACCCTTCTCT	hypothetical protein MGC26706 (MGC26706)
NM_024069	chr19	+	19060565	CGCCCTTTCCT	hypothetical protein MGC2749 (MGC2749)
NM_032313	chr4	-	57840079	CCGCCCCTTTG	hypothetical protein MGC3232 (MGC3232)
NM_152291	chr4	+	71386422	CTTTCTCTTCT	hypothetical protein MGC34772 (MGC34772)
NM_152350	chr17	+	17453017	TGACCTTTTCA	hypothetical protein MGC40157 (MGC40157)
NM_032376	chr17	-	44269530	CTGCCCTTTCC	hypothetical protein MGC4251 (MGC4251)
NM_152289	chr19	-	9954717	CCATCTTTTCC	hypothetical protein MGC45408 (MGC45408)
NM_024116	chr11	-	94982371	AACCCTTTTCT	hypothetical protein MGC5306 (MGC5306)
NM_145058	chr12	+	123915534	CCTCCTTTTCC	hypothetical protein MGC7036 (MGC7036)
NM_152705	chr13	+	22175991	CCTCCTCCCTC	hypothetical protein MGC9850 (MGC9850)
NM_013341	chr2	-	173774408	CCTCCTTCCTC	hypothetical protein PTD004 (PTD004)

NM_018096	chr17	-	35505131	GGCTCTTTCTC	hypothetical protein similar to beta-transducin family (FLJ10458)
NM_021242	chrX	+	36950979	GGGCCTTTTAT	hypothetical protein STRAIT11499 (STRAIT11499)
NM_014886	chr5	+	74295942	GACTCTTTCCT	hypothetical protein YR-29 (YR-29)
NM_001551	chrX	+	66831100	GTTCCTCTCTC	immunoglobulin (CD79A) binding protein 1 (IGBP1)
NM_024658	chr14	-	18445323	TCCCCTTTTCG	importin 4 (IPO4)
NM_006391	chr11	+	9976504	CTTTCCTTTCG	importin 7 (IPO7)
NM_002216	chr10	+	7709605	TGTTCCTTTGA	inter-alpha (globulin) inhibitor, H2 polypeptide (ITIH2)
NM_002199	chr4	-	186075578	TCCTCTCCTTG	interferon regulatory factor 2 (IRF2)
NM_000576	chr2	-	111515593	AAACCTCTTCG	interleukin 1, beta (IL1B)
NM_004515	chr1	-	149374353	ACGCCTCTTCA	interleukin enhancer binding factor 2, 45kDa (ILF2)
NM_002271	chr13	+	93016105	CCTTCTCTCTC	karyopherin (importin) beta 3 (KPNB3)
NM_006801	chr19	-	49262934	CTCCCTCTTCC	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 1 (KDELR1)
NM_006559	chr1	+	31467142	CTCTCTCTCGC	KH domain containing, RNA binding, signal transduction associated 1 (KHDRBS1)
NM_015342	chr5	+	66397220	GCGCCTTTTCT	KIAA0073 protein (KIAA0073)
NM_014773	chr5	+	141890794	ATCTCCCTTGT	KIAA0141 gene product (KIAA0141)
NM_014812	chr1	-	238665097	CGGTCTTTGCC	KIAA0470 gene product (KIAA0470)
NM_014790	chr5	-	147743812	CTCCCTCCTTT	KIAA0555 gene product (KIAA0555)
NM_005327	chr4	+	109232140	GGGTCTCCTCG	L-3-hydroxyacyl-Coenzyme A dehydrogenase, short chain (HADHSC)
NM_002295	chr3	+	38681430	CTGTCTTTTCC	laminin receptor 1 (ribosomal protein SA, 67kDa) (LAMR1)
NM_002290	chr6	-	112474225	CTGTCTTTTCA	laminin, alpha 4 (LAMA4)
NM_002291	chr7	-	106127306	TTCCCTTCTTT	laminin, beta 1 (LAMB1)
NM_021070	chr11	-	66995725	AGGTCTTTCCG	latent transforming growth factor beta binding protein 3 (LTBP3)
NM_002305	chr22	+	34686178	ATCTCTCTCGG	lectin, galactoside-binding, soluble, 1 (galectin 1) (LGALS1)
NM_005570	chr18	-	57000514	CCTCCTCCGCG	lectin, mannose-binding, 1 (LMAN1)
NM_016015	chr16	+	25222412	TACCCTCTTCT	leucine carboxyl methyltransferase 1 (LCMT1)
NM_133259	chr2	-	44391872	TGTCCTTCTGG	leucine-rich PPR-motif containing (LRPPRC)
NM_000895	chr12	-	96793058	CCTCCTCTTCT	leukotriene A4 hydrolase (LTA4H)
NM_006893	chr1	-	202496037	GGCCCTTTTCG	ligatin (LGTN)
NM_013236	chr22	+	42703043	CCGTCTCCTCC	like mouse brain protein E46 (E46L)
NM_016022	chr1	-	145983688	TCCCCTCTTCG	likely ortholog of C. elegans anterior pharynx defective 1A (APH-1A)
NM_014056	chr3	-	42015735	AATTCTTTCTC	likely ortholog of mouse hypoxia induced gene 1 (HIG1)
NM_015972	chr13	+	22176029	CGGTCCTTGCT	likely ortholog of mouse RNA polymerase 1-3 (16 kDa subunit) (RPAC2)
NM_022490	chr9	+	37655387	ACGCCTTTTCC	likely ortholog of mouse RNA polymerase I associated factor, 53 kD (PAF53)
NM_021932	chr11	-	99728	AGTTCTTTGAC	likely ortholog of mouse synembryn (RIC-8)

NM_004862	chr16	-	11833544	CGGCCCTTTTC	lipopolysaccharide-induced TNF factor (LITAF)
NM_000527	chr19	+	11463045	CTTCCTTTGCC	low density lipoprotein receptor (familial hypercholesterolemia) (LDLR)
NM_006330	chr8	-	54953641	CTTCCTTCCGC	lysophospholipase I (LYPLA1)
NM_006120	chr6_random	-	8616984	CACCCTCTCGG	major histocompatibility complex, class II, DM alpha (HLA-DMA)
NM_033554	chr6_random	-	8737164	CTGCCTCCACT	major histocompatibility complex, class II, DP alpha 1 (HLA-DPA1)
NM_002121	chr6	+	33040215	AGTCCTTCTTT	major histocompatibility complex, class II, DP beta 1 (HLA-DPB1)
NM_019111	chr10	-	53496952	TTTTCTTTTAT	major histocompatibility complex, class II, DR alpha (HLA-DRA)
NM_013446	chr7	-	138449704	GGGCCTTTGCT	makorin, ring finger protein, 1 (MKRN1)
NM_005917	chr2	+	64007781	GGGTCTTTTTC	malate dehydrogenase 1, NAD (soluble) (MDH1)
NM_007230	chr9	+	131671377	GGGCCCCTTGG	mannosidase, alpha, class 1B, member 1 (MAN1B1)
NM_002372	chr5	+	109467058	GTTCCTTTCCC	mannosidase, alpha, class 2A, member 1 (MAN2A1)
NM_000528	chr19	-	13000287	CGGCCTTTCCA	mannosidase, alpha, class 2B, member 1 (MAN2B1)
NM_018834	chr5	+	139284851	CTGCCCTTTCC	matrin 3 (MATR3)
NM_006739	chr22	+	32492167	CCGCCTCTTGT	MCM5 minichromosome maintenance deficient 5, cell division cycle 46 (S. cerevisiae) (MCM5)
NM_005898	chr11	+	34751747	TGTCCTTCCTC	membrane component, chromosome 11, surface marker 1 (M11S1)
NM_006838	chr12	+	96231604	CATTCCCTCGC	methionyl aminopeptidase 2 (METAP2)
NM_006636	chr2	+	74638530	TTCCCTCCCGG	methylene tetrahydrofolate dehydrogenase (NAD+ dependent), methenyltetrahydrofolate cyclohydrolase (MTHFD2)
NM_005909	chr5	+	71636666	AATCCTTTCTC	microtubule-associated protein 1B (MAP1B), transcript variant 1
NM_002375	chr3	-	47384798	GGCTCCCTTCT	microtubule-associated protein 4 (MAP4), transcript variant 1
NM_145255	chr17	-	48389437	CATTCTTCCGG	mitochondrial ribosomal protein L10 (MRPL10), nuclear gene encoding mitochondrial protein, transcript variant 1
NM_007208	chr3	-	131962861	CTTTCTTTCCG	mitochondrial ribosomal protein L3 (MRPL3), nuclear gene encoding mitochondrial protein
NM_016503	chr2	+	98250402	CTTCCCTTCGG	mitochondrial ribosomal protein L30 (MRPL30), nuclear gene encoding mitochondrial protein, transcript variant 2
NM_032351	chr17	+	38446349	CGGCCTTTGCG	mitochondrial ribosomal protein L45 (MRPL45), nuclear gene encoding mitochondrial protein
NM_053050	chr2	-	74912595	AGTTCTTCCGG	mitochondrial ribosomal protein L53 (MRPL53), nuclear gene encoding mitochondrial protein
NM_015084	chr5	-	71849442	GTTCCTTTTGG	mitochondrial ribosomal protein S27 (MRPS27), nuclear gene encoding mitochondrial protein
NM_016071	chr7	-	138985286	GGGCCTTCCGG	mitochondrial ribosomal protein S33 (MRPS33), nuclear gene encoding mitochondrial protein, transcript variant 1
NM_002453	chr2	-	55682549	CATTCTTCCGG	mitochondrial translational initiation factor 2 (MTIF2), nuclear gene encoding mitochondrial protein
NM_021038	chr3	+	152830486	CAGTCTTTTCA	muscleblind-like (Drosophila) (MBNL1)
NM_000249	chr3	+	36306934	GGCTCTTCTGG	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1)

NM_004315	chr8	-	18004512	GGCTCTTCTTT	N-acylsphingosine amidohydrolase (acid ceramidase) 1 (ASAH1)
NM_002492	chr3	+	180165292	CCTTCTTCCTC	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 16kDa (NDUFB5)
NM_007103	chr11	+	69070703	GCGTCTCTATC	NADH dehydrogenase (ubiquinone) flavoprotein 1, 51kDa (NDUFV1)
NM_005594	chr12	-	57343767	CGGTCTTTCTG	nascent-polypeptide-associated complex alpha polypeptide (NACA)
NM_000271	chr18	-	20898016	CTTCCTTCCTG	Niemann-Pick disease, type C1 (NPC1)
NM_006432	chr14	-	68777435	GCTTCTTTCCC	Niemann-Pick disease, type C2 (NPC2)
NM_003634	chr22	-	26673177	GGGCCTTCCTG	nipsnap homolog 1 (C. elegans) (NIPSNAP1)
NM_007363	chrX	+	67980817	CGCTCTTTTCT	non-POU domain containing, octamer-binding (NONO)
NM_002482	chr1	+	45050068	GGGTCTCTAAT	nuclear autoantigenic sperm protein (histone-binding) (NASP), transcript variant 2
NM_005437	chr10	+	50433289	CTGCCTTTGGG	nuclear receptor coactivator 4 (NCOA4)
NM_005693	chr11	+	48156964	CAGTCCTTTTG	nuclear receptor subfamily 1, group H, member 3 (NR1H3)
NM_014223	chr1	+	40161052	GGGCCTCTGCA	nuclear transcription factor Y, gamma (NFYC)
NM_006184	chr19	+	49771736	CGCCCTCTGCG	nucleobindin 1 (NUCB1)
NM_005381	chr2	-	231060975	CAGTCTTTCGC	nucleolin (NCL)
NM_002520	chr5	+	171517398	CGTCCTTTCCC	nucleophosmin (nucleolar phosphoprotein B23, numatrin) (NPM1)
NM_018230	chr1	-	225338268	CCATCTCTTCC	nucleoporin 133kDa (NUP133)
NM_005387	chr11	-	4084632	GGCCCTCTGCG	nucleoporin 98kDa (NUP98), transcript variant 3
NM_004537	chr12	+	75989120	GGGTCTTTTTT	nucleosome assembly protein 1-like 1 (NAP1L1), transcript variant 2
NM_014142	chr10	-	12201455	CATCCTTTTAG	nudix (nucleoside diphosphate linked moiety X)-type motif 5 (NUDT5)
NM_015853	chr11	-	64022182	CTTTCTTCTCG	ORF (LOC51035)
NM_004153	chr1	-	51759106	CCTTCTTTTCA	origin recognition complex, subunit 1-like (yeast) (ORC1L)
NM_015878	chr8	-	103944583	TTTCCTTTTTT	ornithine decarboxylase antizyme inhibitor (OAZIN), transcript variant 1
NM_005015	chr14	+	17023198	AGTCCTCTTCC	oxidase (cytochrome c) assembly 1-like (OXA1L)
NM_022121	chr6	-	138275711	CGGCCTCTTCG	p53-induced protein PIGPC1 (PIGPC1)
NM_015640	chr1	-	66827422	CCCCCTCTCTC	PAI-1 mRNA-binding protein (PAI-RBP1)
NM_000278	chr10	+	101739582	CTCCCTTTTCT	paired box gene 2 (PAX2), transcript variant b
NM_006229	chr10	+	117594223	ACTCCTTTCCC	pancreatic lipase-related protein 1 (PNLIPRP1)
NM_020992	chr10	-	96284416	CGCTCTTTCTC	PDZ and LIM domain 1 (elfin) (PDLIM1)
NM_004279	chr7	+	101421598	CTTCCTTCTAG	peptidase (mitochondrial processing) beta (PMPCB)
NM_006567	chr6	+	5246736	GGGCCTCTGGG	phenylalanine-tRNA synthetase (FARS1), nuclear gene encoding mitochondrial protein
NM_018323	chr4	+	25353491	GTTCCCTTTTC	phosphatidylinositol 4-kinase type-II beta (PI4K2B)
NM_004563	chr14	+	18350785	CCTCCTTTTTA	phosphoenolpyruvate carboxykinase 2 (mitochondrial) (PCK2)
NM_000289	chr12	+	48412332	CCCCCTTTCCC	phosphofructokinase, muscle (PFKM)
NM_002631	chr1	+	10302766	GGGTCTTTCCC	phosphogluconate dehydrogenase (PGD)
NM_002663	chr17	+	5055955	TGCTCTCTTGG	phospholipase D2 (PLD2)

NM_002766	chr17	-	77342011	TTGCCTCTGGC	phosphoribosyl pyrophosphate synthetase-associated protein 1 (PRPSAP1)
NM_014759	chr8	-	22442250	CGTTCTTTCTC	phytanoyl-CoA hydroxylase interacting protein (PHYHIP)
NM_016518	chr17	+	29269522	CTGTCTTTGCT	pipecolic acid oxidase (PIPOX)
NM_002568	chr8	-	101804353	CTTCCCCTTCT	poly(A) binding protein, cytoplasmic 1 (PABPC1)
NM_005016	chr12	+	53958442	CCGCCCTTCCC	poly(rC) binding protein 2 (PCBP2), transcript variant 1
NM_013284	chr7	-	43768278	TTCCCTCTGCG	polymerase (DNA directed), mu (POLM)
NM_006115	chr22	-	19598704	CGTTCTTTCCT	preferentially expressed antigen in melanoma (PRAME)
NM_002624	chr12	-	53815347	CTTCCTCTTCG	prefoldin 5 (PFDN5), transcript variant 1
NM_006667	chrX	+	115359967	TGACCTTTCTG	progesterone receptor membrane component 1 (PGRMC1)
NM_005040	chr11	-	84150307	CCTCCTTTTCG	prolylcarboxypeptidase (angiotensinase C) (PRCP)
NM_002786	chr11	-	15289524	ACTTCTCTGTA	proteasome (prosome, macropain) subunit, alpha type, 1 (PSMA1), transcript variant 2
NM_002797	chr14	-	17291410	AGTTCTTTCTG	proteasome (prosome, macropain) subunit, beta type, 5 (PSMB5)
NM_015897	chr19	+	4087711	GGCCCTTCTTG	protein inhibitor of activated STAT protein PIASy (PIASY)
NM_002743	chr19	+	11769285	CTTTCTTTCTG	protein kinase C substrate 80K-H (PRKCSH)
NM_005400	chr2	+	46469715	CGTCCTTCCAG	protein kinase C, epsilon (PRKCE)
NM_002731	chr1	+	83840437	TTTTCTTTGCT	protein kinase, cAMP-dependent, catalytic, beta (PRKACB)
NM_002707	chr2	-	27725002	GCGCCTTTCAC	protein phosphatase 1G (formerly 2C), magnesium-dependent, gamma isoform (PPM1G)
NM_014225	chr19	+	53054635	CCTTCTTCTCC	protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), alpha isoform (PPP2R1A)
NM_013336	chr3	+	127954721	GTGTCTCTCGG	protein transport protein SEC61 alpha subunit isoform 1 (SEC61A1)
NM_002828	chr18	-	12970087	CGCTCTCCCCG	protein tyrosine phosphatase, non-receptor type 2 (PTPN2), transcript variant 1
NM_000533	chrX	+	99940828	AGCCCTTTTCA	proteolipid protein 1 (Pelizaeus-Merzbacher disease, spastic paraplegia 2, uncomplicated) (PLP1)
NM_002668	chrX	+	47269160	CCCCCTTCCCG	proteolipid protein 2 (colonic epithelium-enriched) (PLP2)
NM_004103	chr8	+	27688345	AGCCCTTTTAC	PTK2B protein tyrosine kinase 2 beta (PTK2B), transcript variant 2
NM_004582	chr1	+	75186983	CTCTCCTTTCC	Rab geranylgeranyltransferase, beta subunit (RABGGTB)
NM_002866	chr19	-	18706855	CTCCCTTTGCA	RAB3A, member RAS oncogene family (RAB3A)
NM_019034	chr12	+	121969315	CGACCTCTTGG	ras homolog gene family, member F (in filopodia) (ARHF)
NM_001665	chr11	-	4128043	CTTCCTTCTCG	ras homolog gene family, member G (rho G) (ARHG)
NM_007273	chr12	+	6980605	CTTTCTTTCGC	repressor of estrogen receptor activity (REA)
NM_003979	chr12	+	13206192	TCCTCTTTTCC	retinoic acid induced 3 (RAI3)
NM_001033	chr11	+	4381889	CGCCCCTTTGT	ribonucleotide reductase M1 polypeptide (RRM1)
NM_002950	chr3	-	128566544	TGCTCTTCCCG	ribophorin I (RPN1)
NM_006013	chrX	+	150007389	GCGCCTCTTTC	ribosomal protein L10 (RPL10)
NM_007104	chr6	+	35432597	GTCTCTTTTCC	ribosomal protein L10a (RPL10A)

NM_000975	chr1	+	23089282	CTTTCTCTTCC	ribosomal protein L11 (RPL11)
NM_000976	chr9	-	121860223	CGGCCTCTCGG	ribosomal protein L12 (RPL12)
NM_000977	chr16	+	90614348	CTTCCTTTCCG	ribosomal protein L13 (RPL13), transcript variant 1
NM_012423	chr19	+	50359002	CCTCCTTTTCC	ribosomal protein L13a (RPL13A)
NM_003973	chr3	+	39727112	CTTCCTTCTCG	ribosomal protein L14 (RPL14)
NM_002948	chr3	+	23641155	CTTTCCTTTCC	ribosomal protein L15 (RPL15)
NM_000985	chr18	-	46818521	CTTCCTCTTTC	ribosomal protein L17 (RPL17)
NM_000979	chr19	-	49490561	CGTTCTCTCTT	ribosomal protein L18 (RPL18)
NM_000980	chr19	+	18362738	CTTCCTTTTGC	ribosomal protein L18a (RPL18A)
NM_000981	chr17	+	39354385	CTTTCCTTTCG	ribosomal protein L19 (RPL19)
NM_000982	chr7	+	19686691	TTGCCCTTTCG	ribosomal protein L21 (RPL21)
NM_000983	chr1	-	6068571	CTCCCTTTCTA	ribosomal protein L22 (RPL22)
NM_000978	chr17	-	39007795	CTTCCTTTTTT	ribosomal protein L23 (RPL23)
NM_000984	chr17	+	28946868	GACCCTTTTCA	ribosomal protein L23a (RPL23A)
NM_000986	chr3	-	100835248	CTTTCTTTTCG	ribosomal protein L24 (RPL24)
NM_000987	chr17	-	9380108	AGTTCTCTTCC	ribosomal protein L26 (RPL26)
NM_000990	chr11	+	9150892	CTTCCTTTTTC	ribosomal protein L27a (RPL27A)
NM_000991	chr19	+	56321905	TCCTCTTTCCG	ribosomal protein L28 (RPL28)
NM_000992	chr3	-	51283185	CAGCCCCTTTC	ribosomal protein L29 (RPL29)
NM_000967	chr22	-	36330158	CGGCCTCTACC	ribosomal protein L3 (RPL3)
NM_000989	chr8	-	99131761	CTTTCCTTTCT	ribosomal protein L30 (RPL30)
NM_000993	chr2	+	100071483	CTTCCTTTCCA	ribosomal protein L31 (RPL31)
NM_000994	chr3	-	12822921	CCGTCCCTTCT	ribosomal protein L32 (RPL32)
NM_000995	chr4	+	109862813	CTTCCTCTTCC	ribosomal protein L34 (RPL34), transcript variant 1
NM_007209	chr9	-	119270786	CTTCCTCTTTC	ribosomal protein L35 (RPL35)
NM_015414	chr19	+	5758909	AGCCCTTCCGC	ribosomal protein L36 (RPL36), transcript variant 2
NM_021029	chrX	+	97617140	CTTTCTTTCCG	ribosomal protein L36A (RPL36A)
NM_001001	chr14	-	43883318	CTTCCCTTTCC	ribosomal protein L36a-like (RPL36AL)
NM_000997	chr5	-	41927498	CGGTCTTTCTG	ribosomal protein L37 (RPL37)
NM_000998	chr2	+	216084945	CTTCCTTTCTG	ribosomal protein L37a (RPL37A)
NM_000999	chr17	+	75257904	CGTCCTTTTCC	ribosomal protein L38 (RPL38)
NM_001000	chrX	-	115911188	CCTCCTCTTCC	ribosomal protein L39 (RPL39)
NM_000968	chr15	-	59896199	CTTCCTTTTCC	ribosomal protein L4 (RPL4)
NM_021104	chr12	-	56872911	CTTTCTCTCGG	ribosomal protein L41 (RPL41)
NM_000969	chr1	+	92503591	GGCCCTTTTCC	ribosomal protein L5 (RPL5)
NM_000970	chr12	-	112324592	AATTCTCTTTC	ribosomal protein L6 (RPL6)

NM_000971	chr8	-	74146410	CTTCCTCTTTT	ribosomal protein L7 (RPL7)
NM_000972	chr14	-	64149282	AGCTCTCTCCT	ribosomal protein L7a (RPL7A)
NM_000973	chr8	-	146054949	TTTCCTCTTTC	ribosomal protein L8 (RPL8), transcript variant 1
NM_000661	chr4	-	39634946	CGTTCTTTCTT	ribosomal protein L9 (RPL9)
NM_001014	chr6	-	34390261	CCTTCCTTTCC	ribosomal protein S10 (RPS10)
NM_001015	chr19	+	50367828	CTTTCTTTTTT	ribosomal protein S11 (RPS11)
NM_001016	chr6	+	132982826	AGGCCTCTTTC	ribosomal protein S12 (RPS12)
NM_001017	chr11	-	18042975	CTCTCCTTTCG	ribosomal protein S13 (RPS13)
NM_005617	chr5	-	150426438	CTCTCTCTTTC	ribosomal protein S14 (RPS14)
NM_001018	chr19	+	1507698	CGATCTCTTCT	ribosomal protein S15 (RPS15)
NM_001019	chr16	-	18231920	CGTCCTCTTTC	ribosomal protein S15a (RPS15A)
NM_001020	chr19	-	40317744	TTTCCTTTTCC	ribosomal protein S16 (RPS16)
NM_001021	chr15	-	76142596	TTTCCTCTTTT	ribosomal protein S17 (RPS17)
NM_022551	chr6	+	33236302	CTCTCTCTTCC	ribosomal protein S18 (RPS18)
NM_001022	chr19	+	42756041	TTCCCTTTCCC	ribosomal protein S19 (RPS19)
NM_002952	chr16	-	2041908	CGTTCTTCTTT	ribosomal protein S2 (RPS2)
NM_001023	chr8	-	56926074	CCACCCCTTTC	ribosomal protein S20 (RPS20)
NM_001024	chr20	+	60685687	CTTCCTTTCTC	ribosomal protein S21 (RPS21)
NM_001025	chr5	-	81801096	TTCTCTCTTTC	ribosomal protein S23 (RPS23)
NM_001026	chr10	+	79003250	GGTTCTCTTTT	ribosomal protein S24 (RPS24), transcript variant 2
NM_001028	chr11	-	120400848	CTTCCTTTTTG	ribosomal protein S25 (RPS25)
NM_001029	chr4	-	114416697	GCTCCTCTCTC	ribosomal protein S26 (RPS26)
NM_001030	chr1	+	149694319	GCTCCTTTCCG	ribosomal protein S27 (metallopanstimulin 1) (RPS27)
NM_002954	chr2	+	55646009	CTTCCTTTTCG	ribosomal protein S27a (RPS27A)
NM_001031	chr19	+	8476429	ACTCCTCTCCG	ribosomal protein S28 (RPS28)
NM_001032	chr14	-	43849063	CTTCCTTTTAC	ribosomal protein S29 (RPS29)
NM_001005	chr11	+	76649600	CTTCCTTTCCT	ribosomal protein S3 (RPS3)
NM_001006	chr4	+	152409374	CGCCCTTTTGG	ribosomal protein S3A (RPS3A)
NM_001007	chrX	-	68731855	GGTCCTCTTTC	ribosomal protein S4, X-linked (RPS4X)
NM_001008	chrY	+	2617394	GATTCTCTTCC	ribosomal protein S4, Y-linked (RPS4Y)
NM_001009	chr19	+	59352607	CGGCCTCTTCC	ribosomal protein S5 (RPS5)
NM_001010	chr9	-	19569196	GGCCCTCTTTT	ribosomal protein S6 (RPS6)
NM_001011	chr2	+	3265224	GGGTCTCTTCC	ribosomal protein S7 (RPS7)
NM_001012	chr1	+	44241669	GTTTCTCTTTC	ribosomal protein S8 (RPS8)
NM_001004	chr11	-	735497	CTTCCTTTTCC	ribosomal protein, large P2 (RPLP2)
NM_001002	chr12	-	120165589	CCTTCTCTCGC	ribosomal protein, large, P0 (RPLP0), transcript variant 1

NM_001003	chr15	+	62843704	AGCCCCTTTCC	ribosomal protein, large, P1 (RPLP1)
NM_002938	chr4	+	2337829	GCTTCTTCTCC	ring finger protein 4 (RNF4)
NM_006743	chrX	+	46693526	TACTCTTTATC	RNA binding motif protein 3 (RBM3)
NM_016090	chr11	+	115783421	CGACCTTTTGG	RNA binding motif protein 7 (RBM7)
NM_015169	chr8	+	67280841	CTTTCTTTTCC	RRS1 ribosome biogenesis regulator homolog (S. cerevisiae) (RRS1)
NM_003729	chr1	+	99912878	GCTTCTTCCGC	RTC domain containing 1 (RTCD1)
NM_005622	chr16	+	20784104	CCCTCTTCTTT	SA hypertension-associated homolog (rat) (SAH)
NM_000231	chr13	+	17735076	AGCCCTTTCTC	sarcoglycan, gamma (35kDa dystrophin-associated glycoprotein) (SGCG)
NM_006063	chr2	+	168907877	CTGCCTTTTTA	sarcomeric muscle protein (SARCOSIN)
NM_019887	chr12	+	122649705	CTTCCTTTTCA	second mitochondria-derived activator of caspase (SMAC)
NM_015129	chrX	-	115812754	CTTTCTCTTTG	septin 6 (SEPT6), transcript variant II
NM_000295	chr14	-	88671653	CTGTCTCCTCA	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 (SERPINA1)
NM_000624	chr14	+	88864477	AGCCCTCTGCC	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5 (SERPINA5)
NM_004568	chr6	-	2956894	CTCCCTTCGCG	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 6 (SERPINB6)
NM_001235	chr11	+	76812208	AGGTCTTTGGC	serine (or cysteine) proteinase inhibitor, clade H (heat shock protein 47), member 2 (SERPINH2)
NM_015966	chr20	+	33848259	TCCCCTTTCCG	serologically defined breast cancer antigen 84 (SDBCAG84)
NM_003022	chrX	+	77419216	CCTTCTCTGCC	SH3 domain binding glutamic acid-rich protein like (SH3BGRL)
NM_006456	chr17	-	77573954	CTCCCTTCTGC	sialyltransferase (STHM)
NM_006947	chr4	+	57295368	CCGCCCCTCGT	signal recognition particle 72kDa (SRP72)
NM_003145	chr1	-	151755974	CGGTCTTTCGG	signal sequence receptor, beta (translocon-associated protein beta) (SSR2)
NM_004596	chr19	+	41648832	AGTTCTCTCCG	small nuclear ribonucleoprotein polypeptide A (SNRPA)
NM_006516	chr1	-	42428105	CGCTCTCTGGC	solute carrier family 2 (facilitated glucose transporter), member 1 (SLC2A1)
NM_001042	chr17	+	7914596	GCGTCTTTTCC	solute carrier family 2 (facilitated glucose transporter), member 4 (SLC2A4)
NM_002635	chr12	+	99491254	CGGCCTCTGTG	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3 (SLC25A3)
NM_003982	chr14	-	17071944	CTTCCCTTTTA	solute carrier family 7 (cationic amino acid transporter, y+ system), member 7 (SLC7A7)
NM_014748	chr2	+	27686061	CCGCCTTCCCA	sorting nexin 17 (SNX17)
NM_014426	chr20	-	17897437	CGGTCTTTCTC	sorting nexin 5 (SNX5), transcript variant 2
NM_004684	chr4	-	88752557	TTCCCTTTTGG	SPARC-like 1 (mast9, hevin) (SPARCL1)
NM_005470	chr10	-	26866445	CTGTCTCTTTA	spectrin SH3 domain binding protein 1 (SSH3BP1)
NM_012426	chr16	+	70978128	GTGCCTTTTTC	splicing factor 3b, subunit 3, 130kDa (SF3B3)
NM_005850	chr1	-	145642095	GGATCTCTTTC	splicing factor 3b, subunit 4, 49kDa (SF3B4)

NM_006804	chr17	+	39791164	TCTTCTTCCGC	START domain containing 3 (STARD3)
NM_014445	chr3	-	151149177	CTTCCTCTTTC	stress-associated endoplasmic reticulum protein 1 (SERP1)
NM_007265	chr10	-	74031185	CTTTCTCTCAG	suppressor of S. cerevisiae gcr2 (HSGT1)
NM_006754	chr7	-	104236443	TGCCCTTCCTC	synaptophysin-like protein (SYPL)
NM_003081	chr20	+	10147546	CTTTCCTTCCC	synaptosomal-associated protein, 25kDa (SNAP25), transcript variant 1
NM_005486	chr17	+	55459610	GGCCCTCTGGC	target of myb1-like 1 (chicken) (TOM1L1)
NM_030752	chr6	-	160084267	GCCTCTTTTTC	t-complex 1 (TCP1)
NM_021238	chr12	-	31502805	CTATCTTTCTA	TERA protein (TERA)
NM_003217	chr12	+	50212845	TTTCCTTTTTG	testis enhanced gene transcript (TEGT)
NM_004786	chr18	-	54208513	ATCCCTCCCCG	thioredoxin-like, 32kDa (TXNL)
NM_003191	chr5	+	33887230	GCGCCTTTCGA	threonyl-tRNA synthetase (TARS)
NM_021103	chr2	+	85426844	CGCTCTTTTGT	thymosin, beta 10 (TMSB10)
NM_004236	chr15	-	42493764	CCCCCTCCCGG	thyroid receptor interacting protein 15 (TRIP15)
NM_053000	chr5	+	111942278	CTGTCCTTTCT	TIGA1 (TIGA1)
NM_006287	chr2	-	187102254	CTCCCTCTTTG	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor) (TFPI)
NM_014350	chr5	+	119002837	CCTCCTTTTCT	TNF-induced protein (GG2-1)
NM_003205	chr15	+	50626135	CCTCCTCCGTG	transcription factor 12 (HTF4, helix-loop-helix transcription factors 4) (TCF12)
NM_003199	chr18	-	52973528	TTCCCTCTCTT	transcription factor 4 (TCF4)
NM_006022	chr13	-	38995105	ATCTCTTTTCC	transforming growth factor beta-stimulated protein TSC-22 (TSC22)
NM_014765	chr1	-	230745626	CGGCCTTTCTG	translocase of outer mitochondrial membrane 20 (yeast) homolog (KIAA0016)
NM_020243	chr22	+	35692486	CCTCCTTTCCG	translocase of outer mitochondrial membrane 22 homolog (yeast) (TOMM22)
NM_004616	chr12	-	71503755	CTTTCTCTTTC	transmembrane 4 superfamily member 3 (TM4SF3)
NM_004800	chr13	+	94557089	AGTTCTTCCTT	transmembrane 9 superfamily member 2 (TM9SF2)
NM_016456	chr1	-	196582020	GGGTCTTTTGC	transmembrane protein 9 (TMEM9)
NM_015271	chr4	+	154567113	CAGTCTTTTCA	tripartite motif-containing 2 (TRIM2)
NM_007177	chr3	-	57931760	AGCCCTCCTTG	TU3A protein (TU3A)
NM_003295	chr13	-	39901802	CGGCCTTTTCC	tumor protein, translationally-controlled 1 (TPT1)
NM_003595	chr22	-	23682188	CCCCCTTCCCC	tyrosylprotein sulfotransferase 2 (TPST2)
NM_003366	chr16	+	21973351	CGGCCTCCGCC	ubiquinol-cytochrome c reductase core protein II (UQCRC2)
NM_003333	chr19	+	19074649	TCTTCTTTTTC	ubiquitin A-52 residue ribosomal protein fusion product 1 (UBA52)
NM_014256	chr19	+	18298242	AACTCTTTCTT	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 3 (B3GNT3)
NM_006759	chr2	+	64255196	TTACCTTTTCC	UDP-glucose pyrophosphorylase 2 (UGP2)
NM_003715	chr4	+	76839943	TCCCCTTTTGC	vesicle docking protein p115 (VDP)
NM_003380	chr10	+	17234838	AGTCCTCTGCC	vimentin (VIM)

NM_003375	chr10	+	76073979	GTCTCCTTCAC	voltage-dependent anion channel 2 (VDAC2)
NM_017883	chrX	+	46716528	AGTTCTTTCTG	WD repeat domain 13 (WDR13)
NM_018031	chr3	+	48192625	CTCCCTTCTTA	WD repeat domain 6 (WDR6), transcript variant 1
NM_004804	chr2	+	95386098	GAGCCTCTGTC	WD40 protein Ciao1 (CIAO1)
NM_022170	chr7	+	72228938	GTTCCTCTCGG	Williams-Beuren syndrome chromosome region 1 (WBSCR1), transcript variant 1
NM_006523	chr10	-	110916890	CCTCCTTCGCG	X-prolyl aminopeptidase (aminopeptidase P) 1, soluble (XPNPEP1)
NM_133476	chr12	+	6908210	CCCCCTTTTCG	zinc finger protein 384 (ZNF384)
NM_006626	chr9	-	117322151	TTTTCTTCCAG	zinc finger protein with interaction domain (ZID)