DNA 一次配列からのプロモーター活性予測

平成 23 年 8 月

東京大学大学院　新領域創成科学研究科
メディカルゲノム専攻
ゲノム制御医科学分野

入　江　　拓　磨

目次

1. 序論

1.1 転写制御とヒトゲノム

遺伝子の発現制御は多くの生命現象において重要な制御段階であり，その理解は生命現象を明らかにする上で重要であるといえる[1]. 遺伝子の発現制御は転写・翻訳など多段階の制御ステップによって構成されており，中でも転写開始の制御は最初の制御段階であるため主要な制御ステップであるといえる. 遺伝子の転写開始の制御を担っているのは，プロモーターやエンハンサーとよばれるゲノム上の転写調節領域である. これら転写調節領域はDNAの配列情報として記述されているため，遺伝子機能の理解をする上でもプロモーター領域の配列情報の解析は重要である.

遺伝子の転写はRNAポリメラーゼIIにより行われる. 転写開始は遺伝子上流へRNAポリメラーゼIIのリクルート，遺伝子上流における転写開始前複合体の形成[2-3]によってなされる. この効率は転写因子とよばれる遺伝子によって，*cis*因子や転写因子結合配列（transcription factor binding site；以下TFBS）と呼ばれるDNAの制御エレメントに塩基配列特異的に結合することにより正負に制御が行われる. そのため，網羅的なプロモーター領域の配列情報解析によって，ゲノムの転写制御機構の全体像が明らかになることが期待される.

米国を中心にヒトゲノム全体の解明を目指してヒトゲノムプロジェクトが提案され，2003年4月にヒトゲノム配列の完全解読が宣言された[2]. その成果として，全ヒトゲノム配列は約30億塩基対により構成され，約22,000遺伝子により構成されていることなどがわかってきた. そしてヒトゲノムの配列情報を自由に扱うことが出来るようになり，プロモーター領域などの機能性領域の解析も進むことが期待される.

しかしながらヒトゲノムプロジェクトの成果により，ゲノムを塩基配列とした一次配列情報として利用することが可能になったが，ゲノムの配列情報だけでは生物学的な情報を抽出することは非常に困難であった. 遺伝子のエクソン・イントロン構造やクロマチン構造，プロモーター等のゲノム上の機能エレメントの予測である. 米国ではENCODE（Encyclopedia of DNA Elements）プロジェクト，国内ではゲノムネットワークプロジェクトが行われてきた. ENCODEプロジェクトでは. ヒトゲノム1%の領域をターゲットトして遺伝子領域，遺伝子上流のプロモーター配列，クロマチン構造，ゲノムの複製などゲノムの全体像解明の向けた実験的・情報学的な技術論・方法論の開発を目指した[3]. ゲノムネットワークプロジェクトでは，転写制御を中心とした生体内のネットワークの網羅的な探査を目的とし，転写因子の転写制御領域への結合，転写因子を中心としたタンパク質相互作用についての網羅的な情報な創出し，分化・発生や疾患などの生命機能の解明を目指した. これらの成果により，ゲノムレベルでの技術や情報の蓄積が進んだ.

1.2　完全長 cDNA とプロモーター領域

上述のENCODEプロジェクトやゲノムネットワークプロジェクトを補完する位置付けとして，完全長cDNAが行われてきた．当研究室ではオリゴキャップ法[4]により作成された5’端の情報を含んだcDNAであるヒト完全長cDNAプロジェクト(FLJプロジェクト)が行われた[5]．これまでのEST解析の主な問題点として，合成されたcDNAの中には逆転写が完全に行われていないものや，分解されたmRNA由来のものが含まれている事があった．それを解決する手法として，mRNAの5’端に存在するキャップ構造特異的にRNAオリゴを置換し，それをタグ配列として用いることで全長のcDNAを得ることが可能になった．したがって完全長cDNAはmRNAの5’端を有している．mRNAの転写開始点（Transcription Start Site, TSS）の決定が可能になった．ゲノム配列情報とTSS情報を組み合わせることで，転写開始点の近傍のゲノム領域をプロモーター領域として決定し，配列解析を取得し解析することも可能になった[6-7]．この情報は，転写開始点データベースDBTSS（ http://dbtss.hgc.jp/)として公開されており[8]，プロモーターのデータベース化により，ヒトプロモーター領域の網羅的な配列情報解析が可能になった．また完全長cDNAを用いた解析によって遺伝子間領域から数多くの転写産物があるということ[9]や，複数の転写開始点をもつ遺伝子の例[10]も見つかってきており，従来の転写制御像よりも遥かに複雑な制御が細胞内で行われていると考えられるようになってきた．

これまで得られた転写開始点情報は各細胞・組織の寄せ集めの形になっており，細胞種・組織ごとのトランスクリプトームを反映していなかった．近年，超並列的にDNA配列解読が可能な“次世代型シークエンサー”が登場し，この問題を解決出来るようになったきている．次世代型シークエンサーでは一度に並列的にDNA配列解読が可能であり，細胞種や組織ごとのTSS情報約1000万箇所を一度の実験で決定出来るようになった．興味のある細胞の転写開始点情報を大規模に得ることができプロモーター領域の解析や発現量の解析が可能になった[11]．

1.3　次世代型シークエンサー

これまでのDNA配列決定法には，Frederic Sangerによって開発されたサンガー・ジデオキシ法が主要な方法として用いられていた[12]．ヒトゲノムプロジェクトにおいても，サンガー・ジデオキシ法が主に用いられてきたものの，全ゲノム配列解読のためには，時間もかかり，コストも莫大なものであった．したがって，ゲノム配列を高速に，大量に，そして低コストに決定する方法がヒトの個人ゲノムの解読や*de novo*のゲノム配列解読などこれからのゲノム研究においては必要不可欠であり，そのような背景から，近年，サンガー法に変わる手法を用いたDNAの塩基配列決定を行う”次世代型シークエンサー”と呼ばれる機器が幾つか登場してきた．Roche/GS FLX (454とも呼ばれる)[13]，Illumina/Genome Analyzer(Solexaとも呼ばれる) [14]，Life Technologies/SOLiDなど各社から製品化されており，それぞれ異なる方法によって配列決定

を実現している[15]. 次世代型シークエンサーのスペックを表1-1で示した. 454ではパイロシークエンス法を採用しておりで並行して配列決定できるサンプルが100万本で約500塩基の配列を決定することが可能である. Illumina GAでは合成シークエンス法を採用しており, 並行して約1億リード, 150塩基を解読することが可能である. SOLiDではリガーゼシークエンス法を採用しており, 並行して約1億リード, 約50塩基を解読することができる. 一回の実験で数10億塩基の配列を決定が可能になってきた. 次世代シークエンサーの使用用途としては, 個人ゲノムや癌ゲノムなどのre-sequencingや新しい生物種のゲノム解読が挙げられる. またSolexaやSOLiDでは並行して数億本ものDNA配列を決定することが可能なことから, 読まれた"タグ配列"の数を計測することによって, 細胞腫・組織レベルでのトランスクリプトーム解析（RNA-seq [16-18], TSS-seq[11]）が可能になった. その他にも, 転写因子とゲノムDNAを架橋結合したサンプルを転写因子に特異的な抗体で濃縮するクロマチン免疫沈降(chromatin immunoprecipitation, ChIP)したサンプルを次世代型シークエンサーで配列を決定し転写因子の結合領域を決定するChIP-Seq法 [19-20], ゲノムDNAをmicrococcalヌクレアーゼによりDNAを消化してヌクレオソームを回収し, 巻き付いているDNA配列を次世代型シークエンサーで解析して, クロマチン構造を解析するNucleosome-Seq法[21], bisulfite処理によるゲノム中のCpGサイトのメチル化領域を決定するmethyl-seq法[22]など様々な応用例が報告されている. 次世代型シークエンサーを用いた応用した研究の一例を表1-2で示した[23-30]. DNA配列を基にした分子生物学の実験手法には応用可能であり, これらの生物学的情報を同一プラットフォームで得ることが可能になってきた.

表1-1 次世代型シークエンサーのスペック比較

Platform	template preparation	chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)
Roche/454 GS FLX Titanium	emulsion PCR	pyrosequencing	330	0.35	0.45	500,000
Illumina/Solexa GAII	solid-phase amplification	reversible terminator	75-100	4, 9	18-35	540,000
Life/APG SOLiD3	emulsion PCR	sequencing by ligation	50	7, 14	30-50	595,000

Metzker., M. L. Nature Reveiws Genetics 11, 31-46(2010)より改変

表1-2 次世代型シークエンサーを用いた実験の応用例

カテゴリー		参考文献
Genomic resequncing	個人ゲノム解読	14, 24
Metagenomic sequencing	メタゲノム解析	25
Transcriptome sequencing (RNA seq, TSS seq)	転写産物のプロファイル	16, 17, 18, 26, 27
bisulfite sequencing (methly seq)	メチル化シトシンのパターン	22
Chromatin immunoprecipitation sequencing (ChIP seq)	転写因子, 修飾ヒストンの結合位置	28, 63
Nuclease fragmentation and sequencing (Nucleosome seq)	ヌクレオソーム構造	29
Small RNA sequencing	microRNAのプロファイル	30

Shendure., J. & Ji., H. L Nat Biotechnol 26, 1135-1145 (2008)より改変

1.4　転写制御情報のモデル化

全ゲノム配列情報解読(1.1), 完全長cDNAによるプロモーター配列の同定(1.2)による配列情報の拡充に加え, マイクロアレイなどの発達によって網羅的な遺伝子発現解析技術も進歩してきた. これらの網羅的な遺伝子発現情報とそれを司るプロモーター配列情報を統合することで転写制御の包括的な理解が期待される. 実際にプロモーター配列からの転写制御の予測が報告されている. すなわち遺伝子発現制御の*in silico*モデルの構築を行うことで遺伝子の転写制御を理解する試みが幾つか行われている. これまでに酵母などの下等真核生物において重回帰分析[31], Motif Expression Decomposition(MED)[32-33], ベイジアンネットワークを用いた確率モデル[34], 熱力学モデル[35-36]などの方法が提案されている. いずれの予測モデルもプロモーター配列中の既知のTFBSを説明変数として転写の説明を試みている. Bussemakerらはマイクロアレイによって得られた遺伝子の転写応答情報をTFBSによって説明する線形和モデルである[31]. NguyenとD'haeseleerはMEDと呼ばれる計算方法で転写応答性予測を行った[32]. この手法も基本となっているのはTFBSを説明変数とした線形和モデルである. BeerとTavazoieはベイジアンネットワークによる確率モデルを用いて酵母の転写応答性のパターン予測法を提案している[34]. Gertzらは数種類の既知のTFBSを含むオリゴDNAをランダムに結合させた人工的に合成したプロモーター断片のプロモーター活性情報を基に転写因子-DNA間, 転写因子-転写因子間の熱力学的パラメーターを推定することで転写活性の予測を試みている[35-36]. 各モデルの精度については, 数値の単純な比較はできないが, BeerとTavazoieのベイジアンネットワークモデルでは酵母の2,587遺伝子の発現変化について相関係数0.51の予測精度, NguyenとD'haeseleerのMEDモデルでは酵母の5,719遺伝子の発現変化について相関係数0.52の予測精度であった. Gertzらの熱力学的モデルでは数百種類の人工プロモーターの活性について相関係数0.66の予測精度であった.

1.5　本研究について

本研究では，体系的なルシフェラーゼアッセイの活性情報を定量的なプロモーター活性情報として用いることで，DNA配列からプロモーター活性の予測モデルの構築を試みた．

DNA配列からの転写活性化能を予測する試みは特に酵母を用いたシステムで成果を挙げている．しかしながらその予測精度は依然として不十分であり，特にヒトやマウスなどの高等真核生物における転写活性化能予測モデルに適用している例[37-39]はあるが，精度の高い予測をすることは困難であった．またプロモーター活性能の“絶対値”を予測する試みはほとんど行われてきていなかった．モデル構築の難しさとして，これまでの転写応答の予測モデルでは発現データとしてマイクロアレイを用いた発現情報を利用していたことが挙げられる．マイクロアレイは発現レベルの変化（相対的変化）を計測するもので，絶対量を計測するものではないことが絶対値予測を困難にしている理由として挙げられる．また細胞内のmRNAの量は幾つかの制御段階を経た生成物であるためmRNAの発現量を決定する因子には図1-1に示すようにプロモーター配列の特徴だけではなく，CpGアイランドのメチル化の状態，ヒストン修飾やヌクレオソーム構造などのクロマチン状態，mRNAの合成速度，mRNAの分解速度など様々な要因が関与していると考えられる[40]．そのためマイクロアレイはDNA配列情報に内在されているプロモーター活性化能の直接の指標とはならないと考えられる．したがってDNA配列で記述されていると考えられる内在的なプロモーター活性のモデル化を行う上では，DNA配列とプロモーター活性の間の関係のみの，要素還元的な実験的アプローチによる遺伝子発現情報が必要であると考えられた．

そこで本研究ではプロモーター活性情報として，体系的なルシフェラーゼレポーター遺伝子アッセイを用いた．レポーター遺伝子アッセイは目的のDNA断片の転写活性化能をルシフェラーゼなどの酵素活性をプロモーター活性として検出する手法である．この情報を利用することによりDNA配列情報をプロモーター活性との間の関係としたモデルを構築できると期待した．すなわち図1-1で示す”naked DNA”である．レポーターアッセイは一過的なプラスミドのトランスフェクションで行うため，CpG islandのメチル化状態やクロマチン構造，mRNAの半減期などの影響も除くことができると期待される．まず体系的なルシフェラーゼアッセイによる定量的プロモーター活性情報を用いることで，プロモーター活性の絶対値を指標としたモデル化を試みた．

完全長cDNAと次世代型シークエンサーを組み合わせた”TSS-seq”が開発された[11]．TSS-seq法の利点としては一回の実験で，1)転写開始点情報を数千万単位で得ることができること，2) 配列の出現回数の頻度情報をmRNAのコピー数の絶対値として利用することができることが挙げられる．したがって転写開始点情報からのプロモーター領域の決定と転写活性化能情報を同時に取得することが可能である．本研究ではTSS-seq法による発現情報を細胞内の*in vivo*な転写活性化能として利用し，構築したプロモーター活性予測モデルと*in vivo*の転写の関連についての解析も行った．さらに次世代型シークエンサーより得られる，RNA polymerase IIの結合領域情報としてChIP-Seq，ゲノムのヌクレオソーム構造として

Nucleosome Seq（micrococcal nuclease-digested genomic DNA sequencing）の情報の関連について解析を行い，DNA配列情報とmRNAとの間の情報として，ヒトゲノム中のプロモーター活性予測値の全体像とRNAポリメラーゼIIの結合箇所やクロマチン構造との比較を行うことでプロモーター活性予測モデルの評価を試みた．

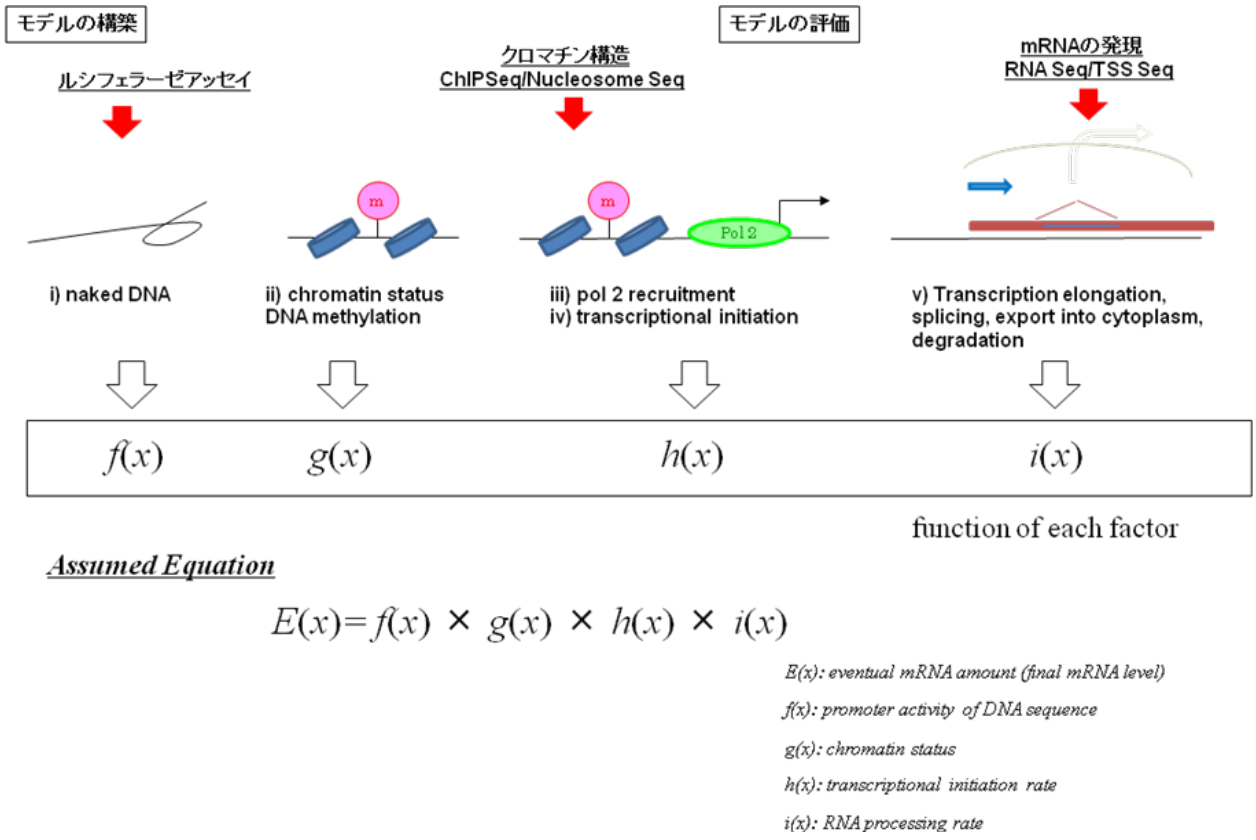


図1-1 転写活性化能の概略図

転写は多段階の制御を受けているため，転写産物の量を推定するためにはそれぞれの制御段階の関数が必要になってくる．本研究では，図中の$f(x)$で表している，DNA一次配列に含まれているプロモーター活性のモデル化を試みた．そのため裸のDNAとプロモーター活性のみの関係とした要素還元的なアプローチをとる必要があると考えられる．本研究ではその目的として，体系的なルシフェラーゼアッセイデータを出発材料とした．またmRNAの合成過程では様々な制御段階が存在しているため，$g(x)$，$h(x)$，$i(x)$などの変数も必要であるといえる．本研究では，構築したプロモーター活性予測モデルの評価を行う目的としてmRNAの発現量，Pol IIの結合，ヌクレオソーム構造との比較を行った．

2. 材料及び方法

2.1 細胞培養

本研究では，ヒト培養細胞であるHuman embryonic kidney (HEK) 293細胞(ATCC number: CRL-1573)を用いた．HEK293細胞は，10% Fetal bovine serum (FBS)，0.584mg/l L-グルタミン酸(GIBCO)，0.15% 炭酸水素ナトリウム，60mg// カナマイシン(GIBCO)を含んだ9.5g// Dulbecco's modified Eagle's medium (DMEM:Nissui)中で37℃，5%CO_2インキュベーター中で培養を行った．

2.2 ヒト遺伝子プロモーター領域のクローニング

2.2.1 遺伝子のプロモーター領域の同定

各遺伝子のプロモーター領域の配列情報は以下のような手順で決定した． HEK293細胞由来のmRNAを，オリゴキャッピング法を利用し，完全長cDNAライブラリーを作成した[4]．作成されたcDNAライブラリーより，ランダムに選んだ12,504種類のクローンについて5'末端側の配列を確認したOne-pass配列を取得した．取得した配列をヒトゲノム配列(hg_18, UCSC genome browser；http://genome.ucsc.edu/)に配列解析ソフトウェアであるBLAT[41]およびSIM4[42]を用いてマッピングを行った．2,170遺伝子についての転写開始点の同定，転写開始点を基準として上流1.0kbから下流0.2kbの領域をプロモーター領域としてヒトゲノム配列から抽出した．

2.2.2 遺伝子のプロモーター領域へのプライマー作成

各遺伝子のプロモーター領域に対して，転写開始点を0としたとき-1000から-900の領域に5'-プライマーを，0から+200の領域に3'-プライマーを設計した．プライマー設計にはPRIMER3(http://frodo.wi.mit.edu/primer3/)を利用した．クローニングにはGatewey Cloning System(Invitrogen)を用いるため，必要となるattB1・attB2組み換えサイトを付加するために5'-primerの5'端にattB1サイトの一部である5'-AAAAAGCAGGT-3'を，3'-primerの5'端にattB2サイトの一部である5'-AGAAAGCTGGGT-3'を付加したプライマーとして作成した．

2.2.3 遺伝子のプロモーター領域の増幅

プロモーター領域のクローニングの概略を図2-1に示した.

PCRによりヒトのゲノムDNAから目的の各遺伝子のプロモーター領域の増幅を行った. 50ngのhuman genomic DNA(Clontech)をテンプレートとし, 作成した5'-primer, 3'-primerを10pmolずつ用いてKOD-plus PCR kit(TOYOBO)を用い, 35サイクル(94℃, 1分;58℃, 1分;68℃, 2分)の条件でPCRを行った.

増幅した各遺伝子のプロモーター領域に対し, 完全なattB1サイトおよびattB2サイトを付加するために, 5'-プライマーにattB1配列(5'-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3')を, 3'-プライマーにattB2配列(5'-GGGGACCACTTTGTACAAGAAAGCTGGGT -3')を用い, PCRのテンプレートとして増幅した各遺伝子のプロモーター領域溶液2μlを用いて再びPCRを, 20サイクル(94℃, 15秒;55℃, 30秒;68℃, 2分)の反応条件で行った. PCRにはKOD-plus PCR kitを利用した.

PCRにより増幅された, attB1・attB2サイトが付加されたPCR産物に対し, ポリエチレングリコール沈殿処理により精製, 濃縮を行った. 濃縮されたプロモーター領域溶液の吸光度をSPECTRA max PLUS 384 (Molecular Devices)を用い測定し, 滅菌水で45ng/μlに調製した.

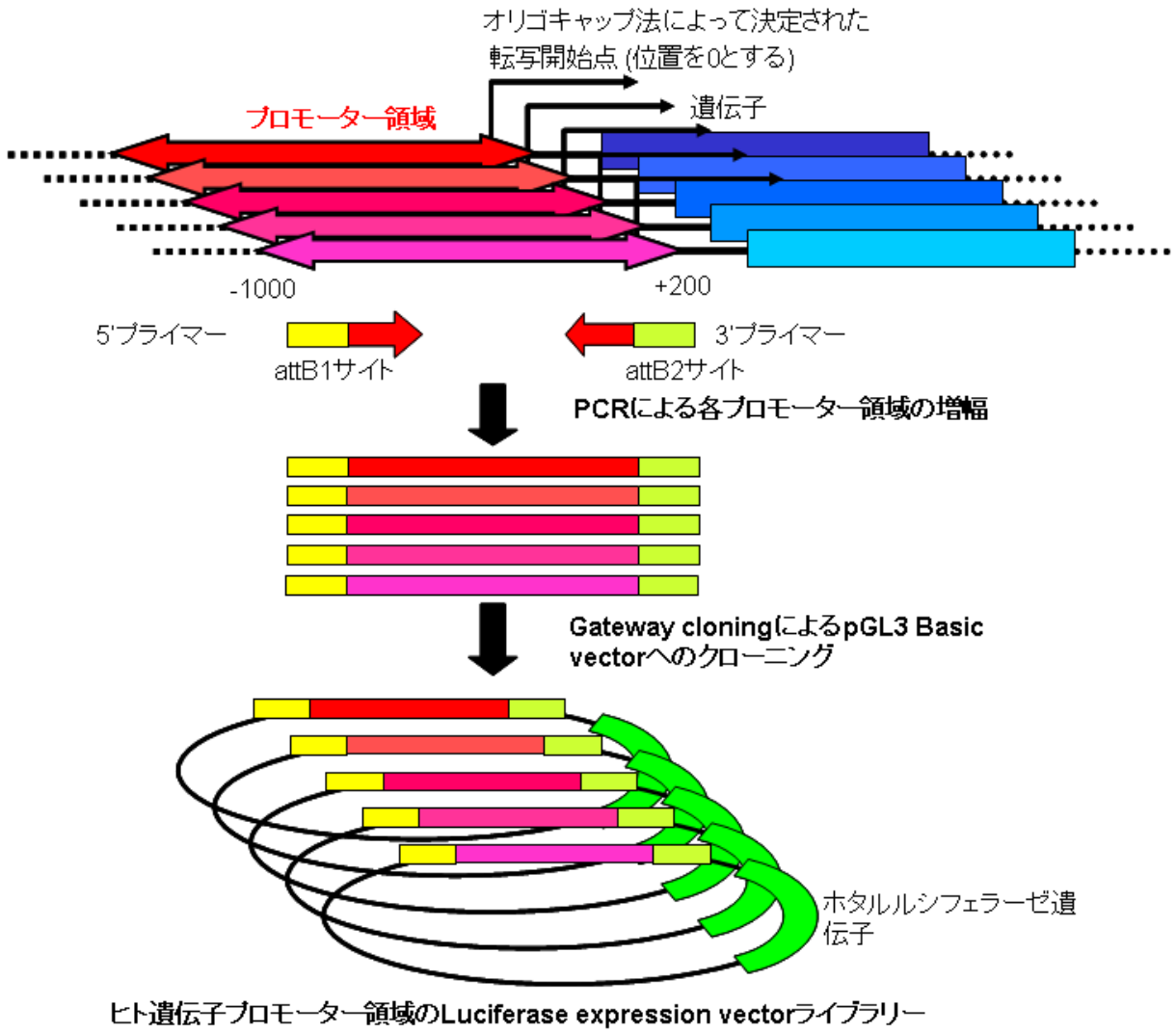


図2-1 ヒト遺伝子のプロモーター領域のクローニングの概略図

2.2.4 遺伝子のプロモーター領域のクローニング

PCRで増幅した各遺伝子のプロモーター領域を，Gateway Cloning System one tube reaction (Invitrogen)を利用し，ホタルルシフェラーゼ遺伝子を有するベクターに挿入した．Gatewayクローニングに用いたベクターは，マルチクローニングサイトの下流にホタルルシフェラーゼ遺伝子を有するpGL3 Basicベクター(Promega)のルシフェラーゼ遺伝子上流のマルチクローニングサイトの*Sma*Iサイトに，Gateway Cloning Systemに用いるGateway組み換えカセットを挿入したpGL3 Basic CaAベクターである．

反応条件は以下のとおりである．45ng/μlに調製したプロモーター領域溶液6.63μl，150ng/μl pDONR 1.33μl，BP buffer 2.67μl，BP clonase 2.67μlを加え25℃一晩反応させた．次に150ng/μl pGL3 basic CaA 2μl，0.75M NaCl 0.66μl，LR clonase 4μl加え25℃一晩反応させた．proteinase K 2μlを加え37℃,10minで反応させた．反応後のサンプルを用い，ヒートショック法によりCompetent High *E. coli* DH5a (TOYOBO)の形質転換を行った．プラスミド

DNAは1.5μ*l*，Competent High *E. coli* DH5aは15μ*l*用いた．その後，形質転換したCompetent High *E. coli* DH5aをLB amp+寒天培地（10g*/l* Bacto Yeast Extract (DIFCO)，5g*/l* Bacto Tryptone (BD)，15g*/l* Bacto Agar (BD)，10g*/l* 塩化ナトリウム，50mg*/l* アンピシリン）で，一晩37℃で培養した．続いて，アンピシリンにより選択されたコロニーを185μ*l*のLB培地（10g*/l* Bacto Yeast Extract (DIFCO)，15g*/l* Bacto Agar (BD)，10g*/l* 塩化ナトリウム，50mg*/l* アンピシリン）中に単離した．一晩37℃で培養後，45μ*l*の80%グリセロール水溶液を添加することでグリセロールストックとして保存した．

2.2.5　遺伝子のプロモーター領域の確認

各遺伝子のプロモーター領域のクローンのグリセロールストック1u*l*をテンプレートとてPCRによりプロモーター領域を増幅した． 5’-プライマー(5’-CTAGCAAAATAGGCTGTCCC-3’)と3’-プライマー(5’-GACGATAGTCATGCCCCGCG-3’)をそれぞれ3.2pmolずつ用い，30サイクル(95℃，15秒；55℃，15秒；72℃，4分)の条件で行った．PCRにはEx taq(TaKaRa)を利用した．続いて，PCR生成物に対しExoSap-IT(USB Corporation)処理を行った．その後，ExoSap-IT処理を施したサンプル2μ*l*をテンプレートとして，BigDye Terminator v3.1 Cycle Sequencing Kit (ABI)を利用し，30サイクル(95℃，10秒；50℃，5秒；60℃，2分30秒)の条件でシークエンサー用サンプルを作成した．シークエンス決定用プライマーには，5’-プライマー(5’-GCCAGAACATTTCTCTATCG-3’)，または3’-プライマー(5’-CTTTATGTTTTTGGCGTCTTCC-3')を3.2pmol用いた．5’端および3’端から増幅したシークエンサー用サンプルに対し，エタノール沈殿処理を行い，20μ*l*の水に溶解した．調製したシークエンサー用サンプルをABI PRISM 3730 DNA Analyzer (ABI)を用いて各遺伝子のプロモーター領域の配列を確認した．

2.3　遺伝子のプロモーター領域ではない，ランダムなゲノム領域のクローニング

プロモーター領域に対してのコントロールとして，プロモーター領域以外の非プロモーター領域のクローニングを行った．ヒトのゲノムDNAから緩やかなアニーリング条件のPCR法によってランダムにゲノム領域のDNA断片の増幅を行った．PCRにはポリメラーゼにEx taq(TaKaRa)を用い，テンプレートにhuman genomic DNA (Clontech)を250ng，プライマーとしてattB1配列・attB2配列を26pmolずつ用いた．プライマーがテンプレートのヒトゲノムDNAに対し容易にアニーリングできるよう，緩やかな条件(95℃，1分；40℃，1分；72℃，1分；20サイクル)でDNA断片の増幅を行った．増幅されたPCR生成物をフェノール・クロロホルム抽出，エタノール沈殿処理を行い精製した後，全量を1%アガロースゲルで電気泳動を行った．電気泳動を行ったPCR生成物をQIAquick Gel Extraction Kit(QIAGEN)を用いて約1.0kb(750bp-1250bp)のDNA断片を精

製した. 精製された約1.0kbのDNA断片溶液の吸光度をSPECTRA max PLUS 384を用いて測定し, 水で45ng/μlになるように調製した. プロモーター領域のクローニングと同様の条件(2.2.4を参照)でpGL3 Basic CaA VectorにDNA断片の挿入, クローニングを行った. さらにプラスミドDNAに挿入された各DNA断片の配列を決定し, 既知のプロモーター領域でない約1.0kbのDNA断片をランダム領域とした. クローニングを行ったランダム領域の詳細を表3-2にまとめた.

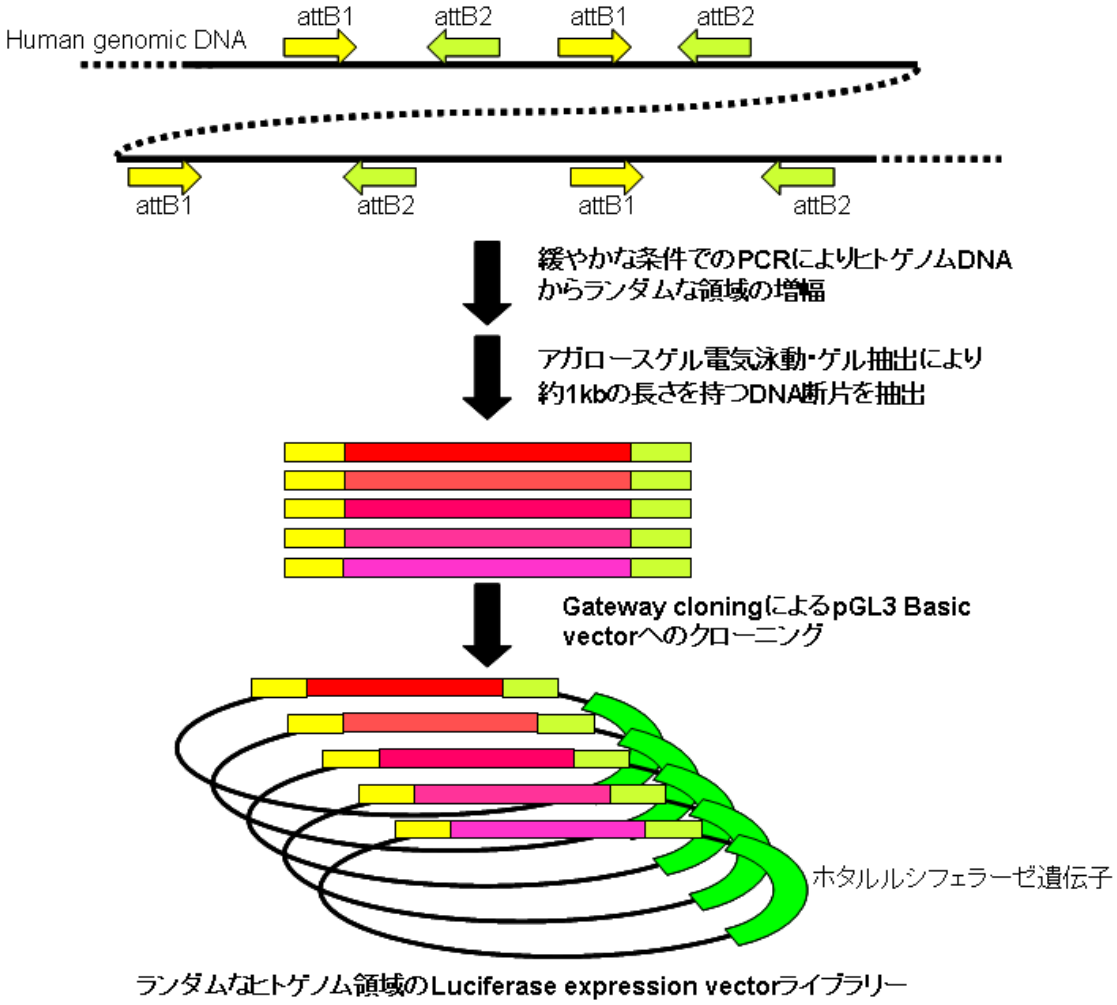


図2-2 ヒトゲノムからランダムな領域のクローニング

2.4 lncRNA のプロモーターのクローニング

2.4.1 lncRNA のプロモーター領域の同定

lncRNAおよびそのプロモーター領域の同定は以下のような手順で決定した. FLJプロジェクトにおいて取得された, ヒト完全長cDNAライブラリーから取得されたOne-pass配列より, 1)既知遺伝子・コンピュータ予測される遺伝子ではない. 2)Open Reading Frame (ORF)の長さが

100コドン以下である. 3)以前の論文に従い計算されたcoding potentialが30以下である. 4)スプライシングが確認される. 5)タンパク質のコードされないexonがコンピュータ予測される. これらの5つの条件を満たす転写産物がlncRNAと定義した. また, その配列をヒトゲノム配列にマッピングすることで768種類の転写開始点の決定およびプロモーター領域の抽出を行った. 本研究では, 768種類のlncRNAのうち, クローニングした当時の段階でアノテーションが完了していた染色体20, 21, 22番にマッピングされた88種類のlncRNAの配列情報および, そのプロモーター領域の配列情報を利用した.

2.4.2 lncRNA のプロモーター領域のクローニング

2.4.1で取得したプロモーター領域に対し, 遺伝子のプロモーター領域のクローニングと同様の方法でlncRNAのプロモーター領域のpGL3 Basic CaAベクターへのクローニングを行った. クローニングを行ったlncRNAのプロモーター領域の詳細を表3-3に示した.

2.5 ルシフェラーゼアッセイによる転写活性化能の測定

2.5.1 プラスミド DNA の精製

目的のDNA断片の挿入が確認されたクローンを, 2m/のLB培地を用い約20時間培養を行った. それぞれのクローンから, QIAwell 96 Ultra Plasmid Kit(Qiagen)を用いてプラスミドDNAを抽出した. さらに, 抽出された各プラスミドDNA溶液に対し, イソプロパノール沈殿処理による濃縮を行い, およびSPECTRA max PLUS 384で吸光度測定を行い, 水で25ng/μ/になるよう濃度調製した.

2.5.2 ルシフェラーゼレポータージーンアッセイ

ルシフェラーゼレポータージーンアッセイの概略図を図2-3に示した.

HEK293細胞5×10^3cellを100μ/のDMEM培地と共に96 WELL CULTURE CLUSTER (Coster)へ分注し, 24時間培養した. 24時間後, HEK293細胞に各DNA断片を挿入されたプラスミドDNA50ng, 補正用のウミシイタケルシフェラーゼ遺伝子を有するphRL-TK Vector (Promega)を5ng, FuGENE6 Transfection Reagent (Roche) 0.3μ/およびDMEM FCS(-)培地9.7μ/の条件でトランスフェクションを行った. 48時間後, 細胞をPhosphate buffered saline (PBS, 8g// NaCl, 0.2g// KCl, 1.44g// Na_2HPO_4, 0.24g// KH_2PO_4) 100μ/で2回洗浄し, Passive Lysis Buffer(Promega) 15μlを用いて細胞を溶解した. その後Dual-Luciferase Repoter Assay System(Promega)でプロモーター活性の測定を行った. 測定にはCentro

XS3 LB960（BERTHOLD）を用いた．測定の条件は次の通り行った．溶解したサンプル3μlを測定に用い，Luciferase assay regent II 15μlを加え2秒静置し，5秒間ホタルルシフェラーゼ活性の測定を行った．その後，Stop and Glo溶液 15μlを加え2秒静置し，5秒間ウミシイタケルシフェラーゼ活性測定を行った．

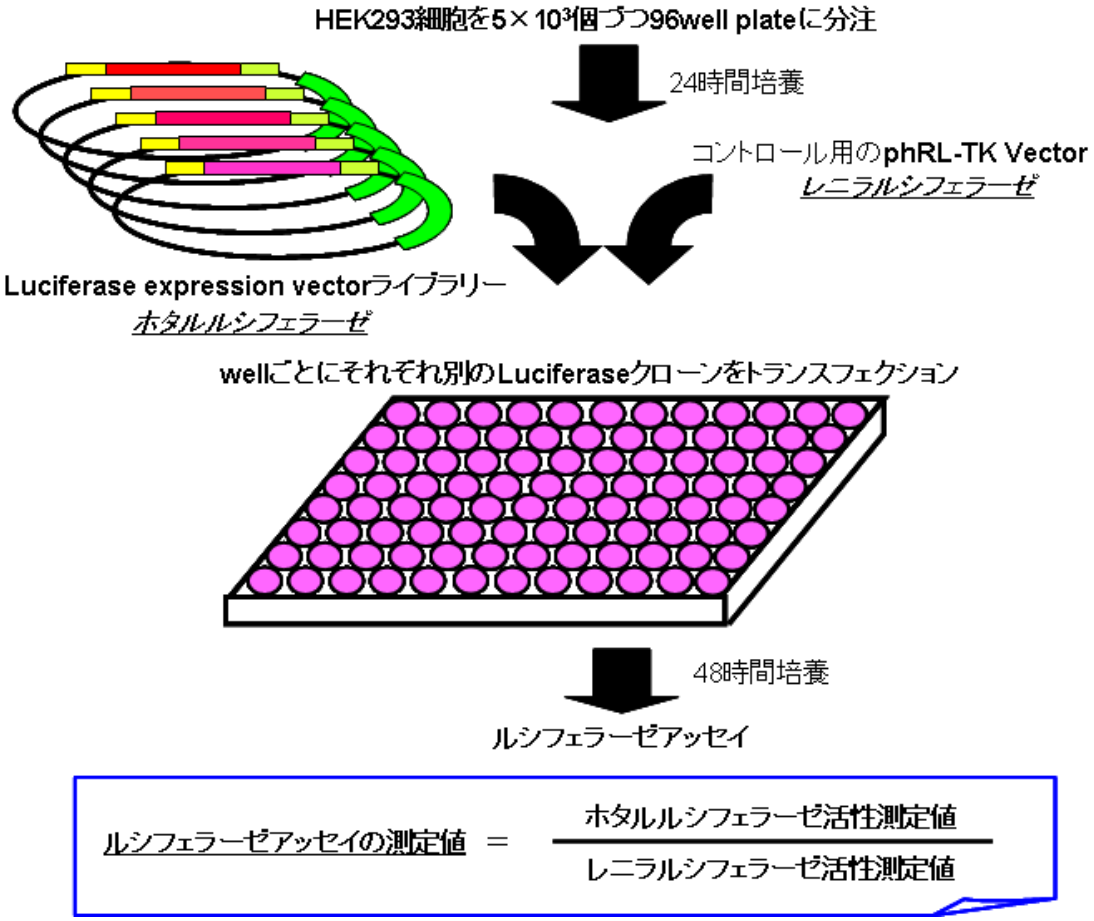


図 2-3 ルシフェラーゼレポーターアッセイ概略図

2.5.3 ルシフェラーゼアッセイの測定値

ルシフェラーゼアッセイにより計測された数値の計算方法を示す．ホタルルシフェラーゼの測定値をレニラルシフェラーゼの測定値で割った値を算出し，3回ずつ測定した値の平均値を計算した．プロモーター領域の含まれていない空ベクターであるpGL3 basic vector（Promega）の測定を実験ごとに行い，実験ごとの標準化を行った．最終的にランダム領域の平均値を1とした時の値に補正し，この値をプロモーター活性とした．転写活性化能の測定値は表3-1，表3-2，表3-3にまとめた．

2.5 転写因子結合配列(transcription factor binding site, TFBS)の探索

クローニングされた各DNA断片中に含まれる既知のTFBSのマトリックス検索を行った. マトリックス検索には転写因子結合配列検索ソフトウェアMATCH [43]を利用し, 転写因子結合配列データベースであるTRANSFAC2008.3のminFP_vertebrate_non_redundant.prfを利用した[44]. これは220種類のTFBSのマトリックスのfolse positiveを最小にするパラメーターのセットである. その後, 同名の転写因子・モチーフごとに192種類にグループ化を行った. 見つかったクローンが4以下であった25種類のTFBSについては本研究の解析には含めず, 167種類のTFBSを解析対象にした.

2.6 プロモーター活性のモデル化

2.6.1 線形和モデル

予測されたTFBSとルシフェラーゼ活性データを用いて, 重回帰分析の方法で, ルシフェラーゼ活性を説明するモデルを構築した. DNA一次配列のプロモーター活性を以下の式(1)のように定義した. 対数変換した転写活性化能は, 各転写因子結合配列の転写への寄与のスコアの総和であると仮定した線形和モデルである.

$$\log(Y) = \sum AX \tag{1}$$

それぞれの変数は
Y:ルシフェラーゼ活性データ(プロモーター活性)
A:DNA配列中に存在しているTFBSの数(またはDNAへの親和性へのスコア)
X:TFBSの転写活性化能のスコア

プロモーターの転写活性化能を目的変数, TFBSの数(またはDNAへの親和性へのスコア)を説明変数とし, 重回帰分析の手法で計算を行い, 変数の推定値を得た. 重回帰分析の計算には統計解析ソフトRのlmコマンドを用いた. 得られたプロモーターの転写活性化能の予測値と実験値との相関係数(Pearson's correlation coefficient)を計算し, モデルの精度を評価した.

2.6.2　マトリックススコアの導入

予測モデルの精度を向上させる目的として，TFBSのマトリックス検索時のスコアをDNAとの親和性のスコアとして利用した．コンセンサス配列のマトリックス検索では配列に対しスコアを与え，そのスコアの値の大きさでコンセンサス配列であるかを決定する．スコアが高いほどコンセンサス配列に近く転写因子との結合の確率も大きくなると考えられる．DNAへの親和性のスコアを線形で近似する方法については[38-39，45]を参考にした．マトリックスのスコアを親和性のスコアへ変換を以下の式(2)を用いた．

$$x' = (x - t)/(a - t) \tag{2}$$

親和性スコアはTRANSFACによって定義された閾値を0とし，マトリックススコアに従い，直線的に増加し，あるマトリックススコアで親和性のスコアが最大値1を取る，線形近似である．*x*はTRANSFACのマトリックススコア，*a*は親和性のスコアが最大をとるマトリックススコア，*t*はTRANSFACの閾値を表す．この親和性スコアをプロモーター活性モデルの*A*のTFBSの数の変わりに用いる．*a*の値を0-1の範囲で0.1の幅で変化させ，それぞれの条件で重回帰分析を行い，プロモーター活性の予測値と実測値の相関係数の値が最小の値をとる条件を最適の条件として決定した．予測値はleave-one-out 交差検定の手法で算出した．予測値と実測値の相関係数（Peason's correlation coefficient）が最小値を取る条件を最適な値とした．

2.6.3　TFBS の位置情報

予測モデルの精度向上を目的として，TFBSの出現位置を考慮した．マトリックスサーチされたTFBSの位置を最適な領域に制限することを試みた．クローニングされたDNA断片の3'端を0と基準にし．転写活性化能を測定したDNA断片の3'端を基準に100bpごとに区切り，各範囲内に存在しているTFBSを計算に用いた．各条件で重回帰分析を行い，プロモーター活性の予測値と実測値の相関係数の値が最小の条件を最適の条件として決定した．予測値はleave-one-out 交差検定の手法で算出した．予測値と実測値の相関係数（Pearons's correlation coefficient）が最小値を取る条件を最適な値とした．

2.6.4　変数選択

プロモーター予測モデルへの説明の寄与の大きい，最小のTFBSの組み合わせを選択する目的として，赤池情報量規準（Akaike's infromation criterion；AIC）[46]を用いたbackward stepwize regressionによる変数選択を行った．計算はRのstepコマンドを用いた．

2.6.5　10-分割交差検定

モデルの未知データに対する有用性及び過適合を検証する目的として10分割交差検定を行った. 全データの90%をランダムに選択し訓練用データとし重回帰分析を行った. TFBSのスコアを用いて, 残りの10%のデータを試験用データとして, プロモーター活性の予測値を算出した. 試験用データの予測値と実験値とのPearsonの相関係数を算出した. ランダムな選択, 相関係数の計算の操作を1000回繰り返した.

2.7　TFBS 変異配列

TFBSについての実験的な評価を目的として, TFBS領域ををを欠失させたプロモータークローンを作成し, ルシフェラーゼ活性の測定を行った. 欠失変異配列作成にはQuikChange II Site-Directed Mutagenesis Kit (Stratagene)を用いた. プライマーはTFBSから上流と下流15～20塩基を組み合わせた配列とその逆鎖配列のプライマーを作成した. PCRにはテンプレートとしてプロモータークローン25ng, 5pmol /μlプライマーを2μlずつ利用した, 試薬・反応条件はQuikChange II Site-Directed Mutagenesis Kitのマニュアルに従った. PCRの条件は94℃・30sec, 55℃・1min, 68℃・7minを18サイクル行った. その後, 1μl *Dpn*Iを加え37℃1hr反応させた. Competent High *E. coli* DH5a(TOYOBO)に形質転換を行った. LB amp+寒天培地に培養後, 数クローン選択して配列を確認した. 配列決定方法はプロモーターのクローニングと同様の方法である. 変異が確認できたクローンをプロモータークローンと同様の方法でプラスミドの回収・調製を行い, ルシフェラーゼ活性の測定を行った. 作成した欠失配列のクローンの情報は表3-6に示した.

2.8　IRES を用いた翻訳の 5'UTR の翻訳ノイズの評価

プロモータークローンごとの翻訳の効率の影響を評価する目的として, ルシフェラーゼ遺伝子上流にIRES配列を組み込んだベクターの作成を行った. IRESは高次構造をとりキャップ構造非依存的に翻訳をすることができ, 各クローンの翻訳効率を揃えることができると考えた. IRESにはpCITE2a vectorのIRESを用いた. 末端に制限酵素サイトを入れたプライマー（5'-ATGGCCATGGGTTATTTTCCACCATATTGC-3'）, （5'-TCTTCCATGGGCATATTATCATCGTGTTTTTCAA-3'）を用いてPCRで増幅した. QIAquick PCR purification kit（QIAGEN）で精製をおこない, *Nco*I処理を行った. pGL3 basic vectorも*Nco*IとCIAP（Calf intestine Alkaline Phosphatase）に反応をさせ, アガロースゲル電気泳動後, QIAquick gel extraction kit（QIAGEN）で精製を行った. これらのDNA断片

のライゲーションを行い，ルシフェラーゼ遺伝子上流にIRES配列を組み込んだpGL3 IRES vectorを得た．さらにgateway cassetteを上記の同じ方法で挿入し，Gateway化を行った．pGL3からpGL3 IRESへのDNA断片の乗せ換えは，pGL3ライブラリーをテンプレート，attB1,attB2配列をプライマーに用いたPCRを行い，クローニングと同様の手法により，gatewayクローニング，配列決定，プラスミド抽出を行った．

2.9 既存のプロモーター予測プログラムとの比較

本研究によって構築されたプロモーター活性予測モデルの評価を行う目的として既存のプロモーター領域予測プログラムとの比較を行った．既存のプロモーター，TSS予測モデルには，ARTS [47], Eponine [48], EP3 [49], ProSOM [50], Promoter2.0 [51], FirstEF [52]を用いた．ARTSは http://www.fml.tuebingen.mpg.de/raetsch/suppl/arts，ProSOMは http://bioinformatics.psb.ugent.be/software/details/ProSOM，Promoter 2.0は http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?promoter, FirstEFはUCSCゲノムブラウザ（http://genome.ucsc.edu/index.html）を利用した．それぞれのプログラムは任意のDNA領域をインプットとして用い，その領域に対しスコアを与える．ルシフェラーゼアッセイに用いたDNA断片のスコアを算出した．このスコアと本研究とのプロモーター活性予測モデルの比較を行った．

2.10 次世代シークエンサー（Illumina GA）を利用したデータ収集

本研究によって構築されたプロモーター活性予測値のヒトゲノムの全体像を解析する目的として，次世代型シークエンサーを用いたゲノム規模の大規模データを利用した解析を行った．本研究においてはHEK293細胞のi)TSS-Seq ii)RNAポリメラーゼIIのChIP-Seq iii)Nucleosome-Seqを利用した．実験方法を以下で述べる．

2.10.1 TSS-Seq

*in vivo*の転写活性化能情報としてTSS-Seq法により得られたHKE293細胞の約1000万のTSS（transcriptional start site）情報を利用した．TSS-Seqとは完全長cDNAの合成，オリゴキャッピング法とIllumina GAを合わせた方法である[11]．Illumina/Solexaシークエンシングに必要なアダプターをmRNAのキャップサイトに3段階の酵素反応により導入することで，TSS下流の配列を直接的に決定することを可能にしている（図2-4）．

HEK293細胞からRNeasy（Qiagen）を用いてRNAを抽出した．抽出した200μgのRNAを用いてoligo-cappingを行った．2.5U BAP（TaKaRa）を37℃ 1hr，40U TAP（Ambion）を37℃

1hr，BAP-TAP処理したRNAと1.2μgのRNAオリゴ（5′-AAUGAUACGGCGACCACCGAGAUCUACACUCUUUCCCUACACGACGCUCUUCCGAUCUGG-3′）を250U のT4 RNA ligase（TaKaRa）を用い20℃, 3hrで連結した. DNase I（TaKaRa）処理後, oligo-dTパウダー（Collaborative）を用いpoly A+ RNAを抽出した. 10 pmol のランダムヘキサマープライマー（5′-CAAGCAGAAGACGGCATACGANNNNNNC-3′）とSuper Script II（Invitrogen）を用い, 12℃ 1hr, 42℃ overnigntで1st cDNA合成を行った. テンプレートのRNAはアルカリ処理により分解した. 1/5の1st strand cDNAをPCRのテンプレートに用いた. PCRにはGene Amp PCRキットを用いた.
プライマーは5′-AATGATACGGCGACCACCGAG-3’と 5′-CAAGCAGAAGACG GCATACGA-3′を用い, 94℃ 1min, 56℃ 1min, 72℃ 2min, で15サイクルを行った. PCR産物は12%ポリアクリルアミドゲル電気泳動を行い, 150～250bpの領域を回収した. 1ngのcDNAを配列決定用のサンプルとした. Illumina GAのマニュアルに従って1’tile’ごとに15000-20000クラスターを生成し36塩基の配列の決定をした.

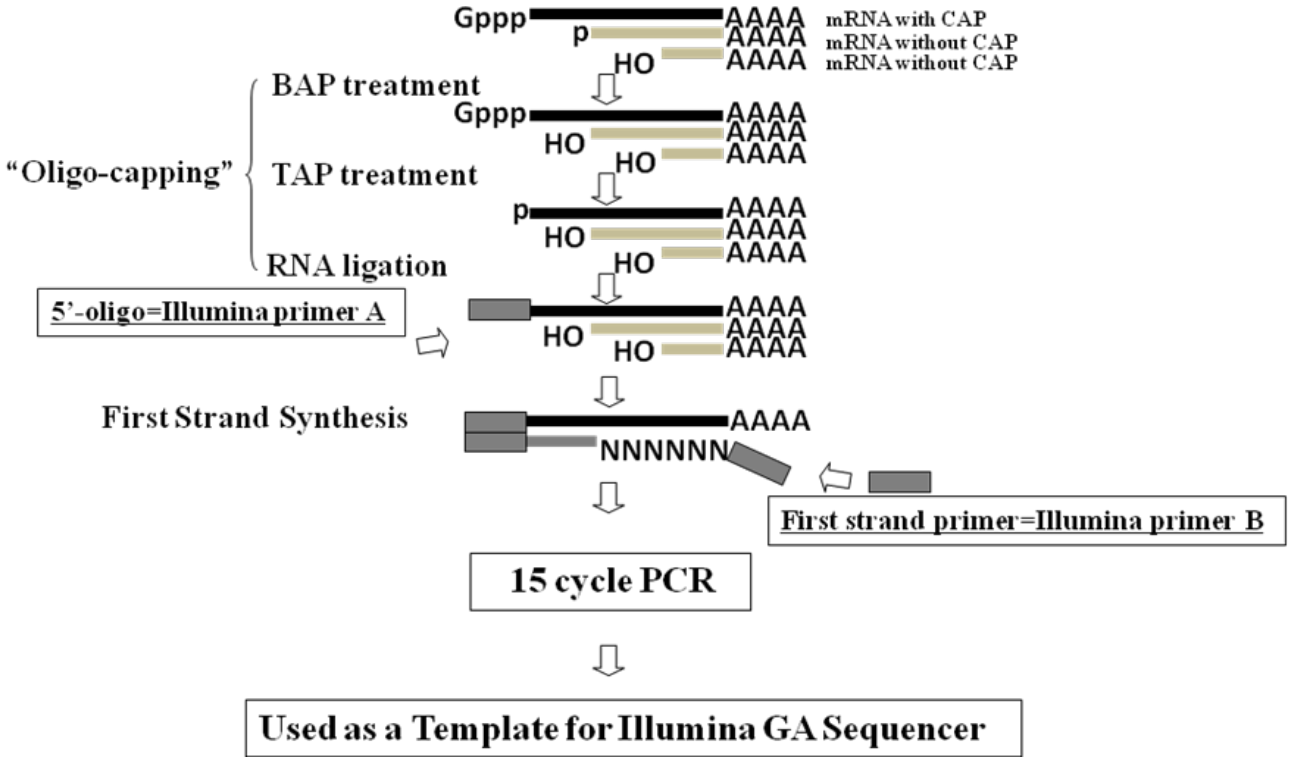


図2-4 TSS-Seq概略図

2.10.2 RNAポリメラーゼIIのChIP-Seq

RNA ポリメラーゼ II (Pol II) のChIP seqデータをPol IIのゲノム上の結合情報として利用した. ChIP-Seq法とはクロマチン免疫沈降法により得られたDNA断片をIllumina GAにより配列決定し, タグ数の情報を転写因子などの結合情報として捉え, ゲノム上における転写因子の結合箇所を調べる方法である(図2-5). 本研究ではPol IIの抗体を用いてPol IIの結合箇所を調べた.

1×10^8のHEK293細胞を終濃度1%ホルムアルデヒドにより室温で10分間クロスリンクを行った後, 終濃度150mMグリシンを添加し, 室温で5分間インキュベートして反応を停止させた. 1×PBS(Phosphate Buffered Saline)で2回洗浄して細胞をハーベストした後, 5mlの溶解バッファー1(50mM Hepes-KOH (pH 7.5), 140mM NaCl, 1mM EDTA, 10% グリセロール, 0.5% NP-40, 0.25% Triton X-100)を加え, 4℃で10分間インキュベートして細胞質画分を溶解した. 溶液を1,500rpm, 4℃で5分間遠心し, 上清を捨てた後, 5mlの溶解バッファー2(10mM Tris-HCl (pH 8.0), 200mM NaCl, 1mM EDTA, 0.5mM EGTA)でペレット(核画分)を懸濁し, 室温で10分間インキュベートして洗浄した. 懸濁液を1,500rpm, 4℃で5分間遠心し, 上清を廃棄した後, 1mlの溶解バッファー3(10mM Tris-HCl (pH 8.0), 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% デオキシコール酸ナトリウム, 0.5% N-ラウロイルサルコシン)にペレットを懸濁させた. 懸濁液をソニケーター(トミー精工)を用いて氷中で冷却しながら30秒間ソニケーション, 2分間冷却を18回繰り返した. ソニケーションした溶液に100μlの10% Triton X-100を加え, 14,000rpmで10分間遠心した. 上清のうち50μlをコントロールとして使用するために保存した(以下WCE-DNA(Whole Cell Extract DNA(全細胞抽出物))とする. 磁気ビーズに10μgのRNA polymerase II CTD repeat antibody(abcam: ab817)を結合させた. 磁気ビーズ溶液をソニケーションした溶液に加え, 4℃で一晩ローテーターを用いて攪拌した. 反応終了後, 磁気ビーズを1mlの洗浄バッファー(50mM Hepes-KOH (pH 7.5), 500mM LiCl, 1mM EDTA, 1% NP-40, 0.7% デオキシコール酸ナトリウム)で8回, 50mM NaClを含むTEバッファーで1回洗浄した. 磁気ビーズに200μlの溶出バッファー(1M Tris-HCl (pH 8.0), 0.5M EDTA (ph 8.0), 10% SDS)を加え, 65℃で15ふん加熱して磁気ビーズからDNAを溶出させた. 溶出後, 上清を新しいチューブに移し, 65℃で一晩インキュベートして脱クロスリンクを行った(以下ChIP-DNAとする). 同時にWCE-DNAに150mlの溶出バッファーを加え, 65℃で一晩インキュベートして脱クロスリンクを行った. 脱クロスリンクした溶液に200μlのTEバッファー, 8μlの10 mg/ml RNase A(フナコシ)を加えて, 37℃で2時間インキュベートしてRNAの分解を行った. その後, 4μlの20mg/ml proteinase K(タカラ)を加え, 55℃で2時間インキュベートしてタンパク質の分解を行った. その後フェノール・クロロホルム抽出とエタノール沈殿により精製した. 精製したChIP-DNA, WCE-DNAを12%ポリアクリルアミドゲル電気泳動を行い, 150-250 bpを切り出し, Illumina GAのプロトコルに従って配列決定を行った.

Illumina GAにより得られた配列情報を基にPol IIのゲノム上の結合部位の同定を行った. ゲ

ノム上にマッピングされたタグ配列の領域を，DNAのサンプルサイズを考慮して120bp延長した．ゲノムのポジションごとにタグ数を集計し，マッピングされた総タグ数に対する100万分率（ppm;parts par million）を算出する．それぞれの伸長したタグ領域についてオーバーラップしたタグ数をカウントした．タグ数情報を用いて，IPサンプルがWCEサンプルに対して5倍以上のタグ数がある領域が120bp以上続く領域を結合がある領域とした．

方法によって特定した結合部位の統計学的な有意性は，タグの分布がポアソン分布に従うものとして，以下のポアソン分布の確率密度関数(3)を用いて検定した．

$$p(x,\lambda) = 1 - \sum_{t=0}^{x} \frac{e^{-\lambda}\lambda^{t}}{t!} \qquad (3)$$

$p(x,\lambda)$は濃縮の確率表す．λはWCEサンプルから計算される120bpのウィンドウサイズでのマップされるタグの期待値を表す．tはタグ数を表す．

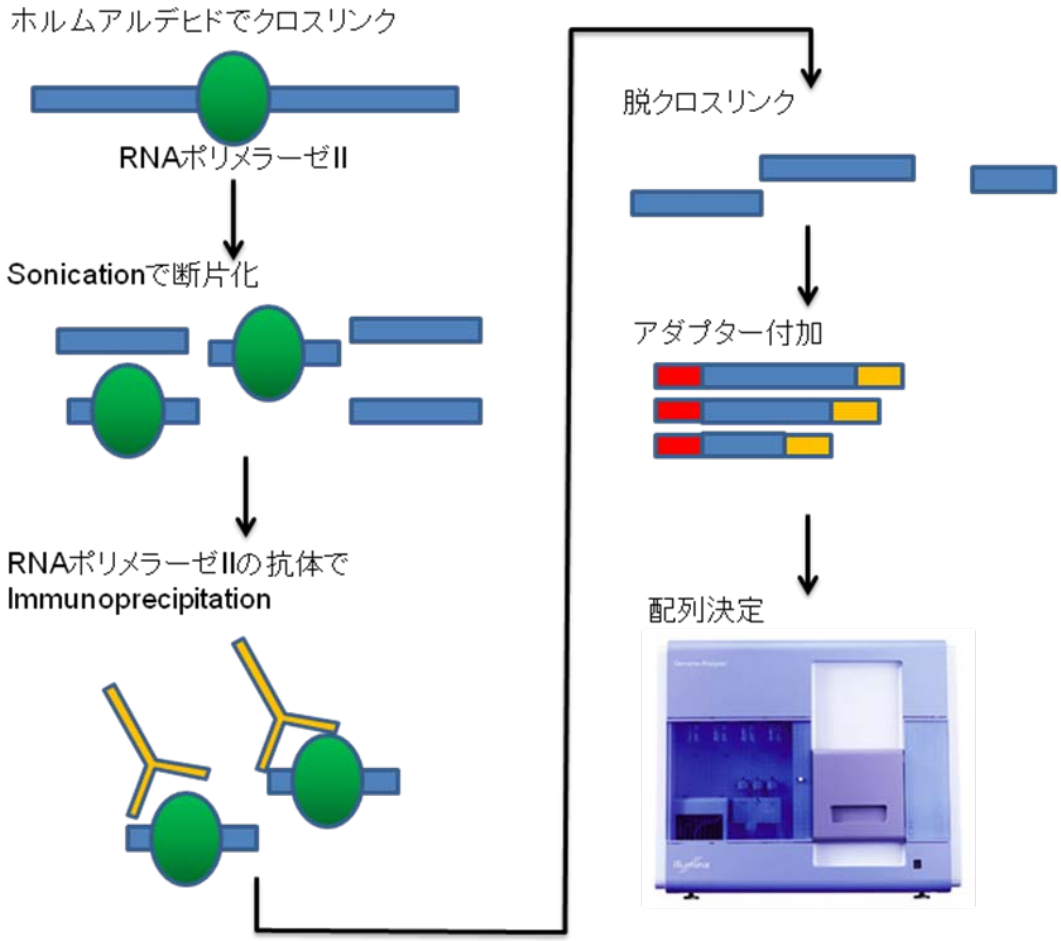


図2-5 ChIP-Seq法概略図

2.10.3 Nucleosome Seq

Nucleosome SeqによりHEK293細胞のヌクレオソーム構造の解析を行った(図2-6).

1×10^7のHEK293細胞を用いた. ChIP-IT Express Enzymatic kit(Active motif)を用いてmicrococcal nucleaseの処理によりモノヌクレオソームの精製を行った. HEK293細胞を終濃度1% ホルムアルデヒドにより室温で10分間クロスリンクを行った後, グリシンを添加して室温で5分間インキュベートしてクロスリンク反応を停止させた. 1×PBS (Phosphate Buffered Saline)で2回洗浄して細胞をハーベストした後, 冷やした溶解バッファーに細胞を懸濁させて30分間氷中でインキュベートした. その後, Dounce型ホモジェナイザーにより15-20回ホモジェナイズを行い, 核各分を分解バッファーに懸濁させ, 37℃で5分間予熱した. Enzymatic cocktail (200 U/ml)を添加し, 37℃で15分間マイクロコッカルヌクレアーゼによるリンカーDNAの分解を行った. 反応後, EDTAを加えて反応を停止させ, 遠心分離で上清を回収した. 5M NaClにより脱クロスリンクを行った後, RNaseを加えて65℃で4時間加熱してRNAを分解した. 続いてProteinase Kを加えて42℃で1.5時間反応させてタンパク質を分解した. これらのDNAサンプルをフェノール・クロロホルム抽出とエタノール沈殿により精製した. 精製したDNAサンプルをIllumina GAにより配列決定した.

ヌクレオソームのスコアを以下の式(4)に従って計算を行った. 最初に, 読まれた配列の5'端から+75bpの場所をヌクレオソームの中心iとした. ヌクレオソームセンター$c(i)$の数をゲノム上のヌクレオソームシグナルとした値$s(p_j)$に変換した.

$$s(p_j)=\log_2\left[\frac{\sum_{i=p_j-75}^{P_j+75} w(i)c(i)}{\sum_{i=p_j-75}^{Pj=75} w(i)}+1\right] \tag{4}$$

p_j = 5+10j (j=0, 1, 2, ...), $w(i)$は平均$p(j)$, 標準偏差20のガウス分布を表している.

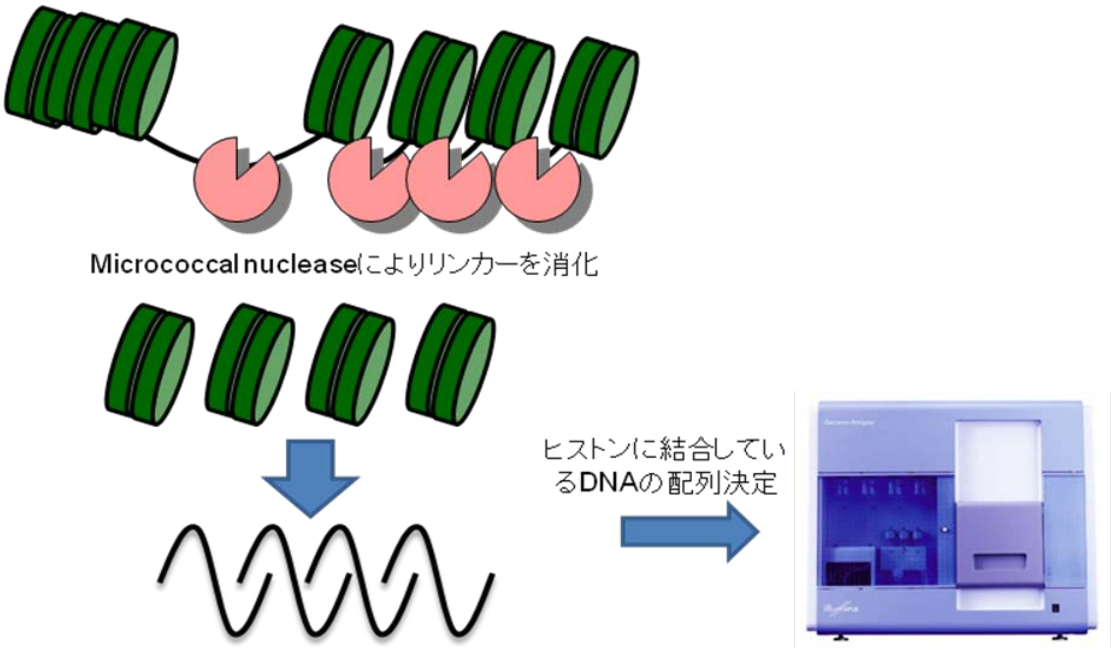


図2-6 Nucleosome-Seq法の概略

2.11 RefSeq 遺伝子 5'端周辺領域のプロモーター活性の予測値

プロモーター活性予測モデルとmRNAレベルとの比較を行った. 18,686種類のヒトRefSeq遺伝子5'端を基準として上流1,000bpから下流200bpの領域をRefSeq遺伝子のプロモーター領域とした. これらの領域の配列情報をゲノム配列から抽出し, プロモーター活性予測モデルに従ってプロモーター活性予測値を算出した. その後プロモーター活性予測値とプロモーター領域に存在したTSS-seqのタグ数との比較を行った.

2.12 RefSeq 遺伝子のプロモーター活性予測値の定性的評価

上述のプロモーター活性予測値とTSS-seqタグ数の定性的な評価を行った. HEK293細胞中で"active"なプロモーターと"silent"なプロモーターを分類能についての解析を行った. "active"なプロモーターをTSS-seqのタグ数が>5ppm以上のプロモーター領域とし, "silent"なプロモーターをTSS-seqのタグ数が0ppmのプロモーター領域とした. 定性的な評価にはPrecision(適合率)とRecall(再現率)値を用いた. 値は以下の式で測定した.

$$Precision = TP /TP + FP \quad (5)$$

$$Recall = TP / TP + FN \quad (6)$$

TP(true positive)は>5ppm, プロモーター活性予測値>1, FP(false positive)は0ppm, プロモーター活性予測値>1, FN(false negative)は>5ppm, プロモーター活性予測値<1のプロモーター数を意味している. Precision(適合率)はactiveだと予想されたものの中で実際にactive(>5ppm)出会ったプロモーターの割合, Recall(再現率)はactiveなプロモーター(>5ppm)のうちactiveである(予測値>1)と予測されたプロモーターの割合を表す. なおこの解析においては0<ppm<5のプロモーターは計算から除外した.

2.13 ヒトゲノム全体のプロモーター活性予測値の計算

本研究で構築したプロモーター活性予測値のヒトゲノム全体の分布の解析を行った. ヒトゲノム配列(hg18)(http://genome.ucsc.edu/)を1,200bpの幅で区切ったDNA領域をインプットとして, それぞれのプロモーター活性予測値を算出した.

3. 結果

3.1 ルシフェラーゼ活性の測定

本研究では, DNA一次配列情報を用いた転写活性化能の絶対値を予測するモデルを構築する目的として定量的なルシフェラーゼアッセイの情報を利用した. 合計で734種類のDNA断片のHKE293細胞内での転写活性化能の情報を得た. ルシフェラーゼライブラリーのデータセット中には, HEK293の完全長cDNAライブラリーのTSSから上流1kb, 下流200bpの領域由来のプロモーター451種類, 約1kbのランダムなゲノム領域のDNA断片248種類, lncRNA (long non-protein coding transcripts)のプロモーター領域35種類が含まれている. これらのライブラリーは先行研究で得られたものを利用した[53]. ヒトを含む哺乳類のプロモーター領域は高CG領域でCpG islandが多数含まれていることが知られているが, このプロモーターライブラリー中には, 83.8%のCpG island+プロモーターが含まれていた. また"CpG rich"プロモーター(分類方法は後述)は84.5%であった. 実験はそれぞれ3回行い, プロモーター活性を算出した. プロモーター活性の情報はゲノム領域の平均値を1とした時の値に補正した. 誤差率の平均は18.1%であった. 実験結果を図3-1, また表3-1, 表3-2, 表3-3にクローンの情報を示した.

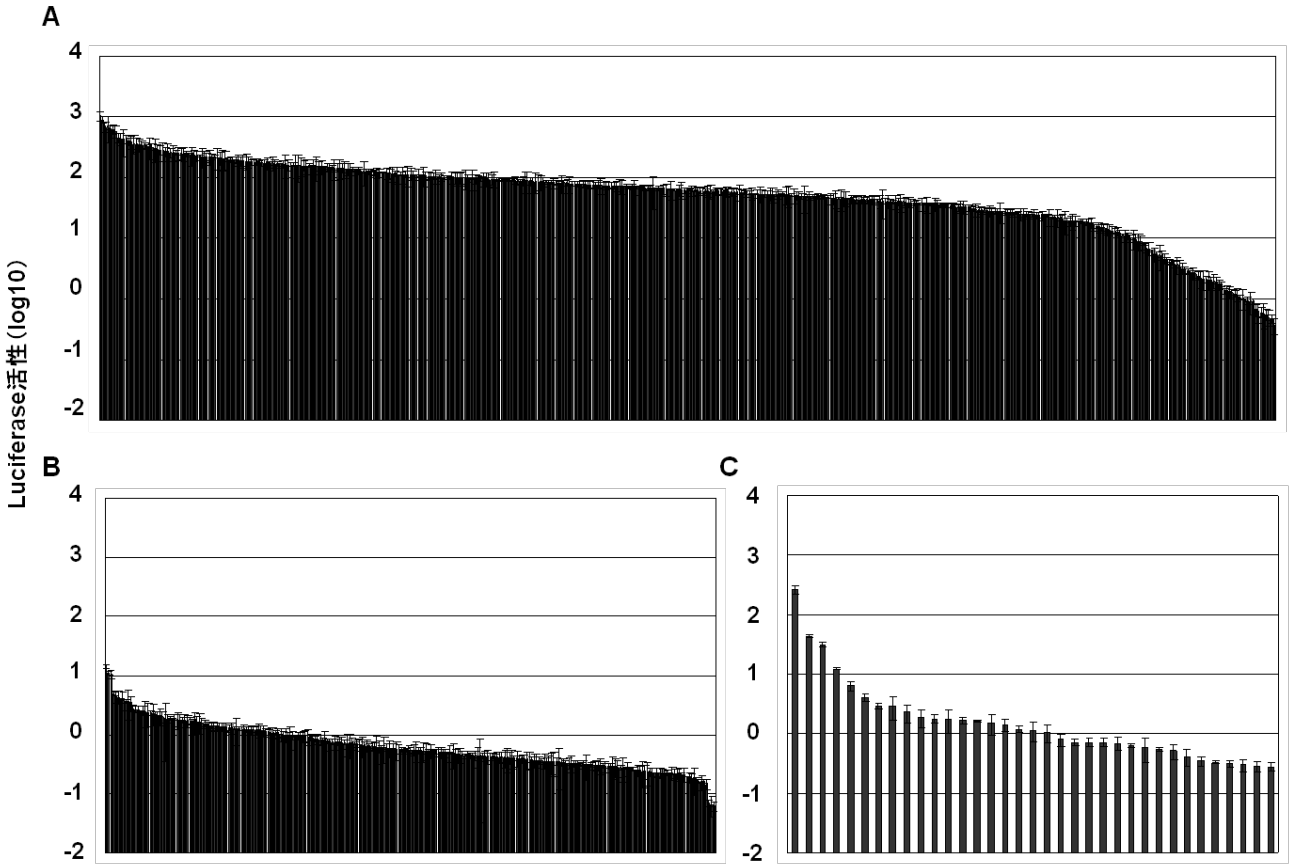


図3-1 ルシフェラーゼ活性の測定値

A プロモーター, B ランダム, C lncRNAプロモーター由来のクローンの測定値. 実験はそれぞれ3回ずつ行った. エラーバーは標準偏差を表している. 値はランダムの平均を1とした時の値に補正した. 上の図では対数表示であるため0がランダムの平均値を表す.

表3-1 プロモーターライブラリーの詳細

ID	chr	strand	start	end	length	Luc activity	sd	ID	Description
promoter1	chr7	-	97339507	97340623	1117	1042.250	177.730	NM_001673	asparagine synthetase (glutamine-hydrolyzing) (ASNS)
promoter2	chr19	-	51058109	51059189	1081	888.605	94.409	NM_004819	symplekin (SYMPK)
promoter3	chr3	+	187983070	187984198	1129	686.105	136.585	NM_001967	eukaryotic translation initiation factor 4A2 (EIF4A2)
promoter4	chr6	+	133176469	133177594	1126	683.623	317.730	NM_001016	ribosomal protein S12 (RPS12)
promoter5	chr4	-	101090329	101091443	1115	653.552	35.012	NM_002106	H2A histone family, member Z (H2AFZ)
promoter6	chr17	-	71287330	71288415	1086	619.963	104.017	NM_005324	H3 histone, family 3B (H3.3B) (H3F3B)
promoter7	chr17	-	72244834	72245941	1108	577.567	117.874	NM_003016	serine/arginine-rich splicing factor 2 (SRSF2)
promoter8	chr1	+	45821378	45822418	1041	455.366	77.681	NM_002482	nuclear autoantigenic sperm protein (histone-binding) (NASP)
promoter9	chr1	-	154257260	154258310	1051	428.217	96.524	NM_003145	signal sequence receptor, beta (translocon-associated protein beta) (SSR2)
promoter10	chr16	+	31097999	31099128	1130	428.126	203.562	NM_004960	fused in sarcoma (FUS)
promoter11	chr11	-	102467889	102469004	1116	426.237	68.096	NM_032299	DCN1, defective in cullin neddylation 1, domain containing 5 (S. cerevisiae) (DCUN1D5)
promoter12	chr22	-	35255039	35256142	1104	394.231	62.146	NM_003753	eukaryotic translation initiation factor 3, subunit D (EIF3D)
promoter13	chr6	-	27208360	27209442	1083	380.799	91.779	NM_021058	histone cluster 1, H2bj (HIST1H2BJ)
promoter14	chr19	+	62816494	62817537	1044	362.238	118.373	NM_003435	zinc finger protein 134 (ZNF134)
promoter15	chr3	+	52206230	52207308	1079	352.287	95.930	NM_000688	aminolevulinate, delta-, synthase 1 (ALAS1)
promoter16	chr8	-	145988493	145989533	1041	345.878	84.483	NM_000973	ribosomal protein L8 (RPL8)
promoter17	chr1	+	32459580	32460743	1164	345.730	48.323	NM_003757	eukaryotic translation initiation factor 3, subunit I (EIF3I)
promoter18	chr7	+	120377063	120378182	1120	337.562	27.593	NM_019071	inhibitor of growth family, member 3 (ING3)
promoter19	chr13	-	44813211	44814291	1081	324.357	16.778	NM_003295	tumor protein, translationally-controlled 1 (TPT1)
promoter20	chr20	+	47095324	47096372	1049	323.382	126.561	NM_001316	CSE1 chromosome segregation 1-like (yeast) (CSE1L)
promoter21	chr15	-	57768692	57769770	1079	315.249	38.105	NM_004330	BCL2/adenovirus E1B 19kDa interacting protein 2 (BNIP2)
promoter22	chr3	-	132704401	132705457	1057	306.432	130.339	NM_007208	mitochondrial ribosomal protein L3 (MRPL3)
promoter23	chr17	+	35696703	35697809	1107	288.610	57.164	NM_001254	cell division cycle 6 homolog (S. cerevisiae) (CDC6)
promoter24	chr11	-	65382149	65383276	1128	279.521	51.335	NM_005507	cofilin 1 (non-muscle) (CFL1)
promoter25	chr9	+	132989831	132990931	1101	277.006	97.097	NM_005085	nucleoporin 214kDa (NUP214)
promoter26	chr7	-	23538023	23539138	1116	271.857	46.715	NM_013293	transformer 2 alpha homolog (Drosophila) (TRA2A)
promoter27	chr2	-	38831819	38832912	1094	269.987	91.245	NM_001031684	serine/arginine-rich splicing factor 7 (SRSF7)
promoter28	chr20	+	42946770	42947893	1124	260.809	45.975	NM_003404	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide (YWHAB)
promoter29	chr1	-	203357649	203358694	1046	256.544	74.002	NM_005057	retinoblastoma binding protein 5 (RBBP5)
promoter30	chr1	-	43410467	43411511	1045	253.790	23.021	NM_006824	EBNA1 binding protein 2 (EBNA1BP2)
promoter31	chr16	-	68346212	68347319	1108	250.627	65.055	NM_014062	NIN1/RPN12 binding protein 1 homolog (S. cerevisiae) (NOB1)
promoter32	chr20	-	472306	473357	1052	250.490	50.901	NM_001895	casein kinase 2, alpha 1 polypeptide (CSNK2A1)
promoter33	chr1	+	152228928	152229986	1059	246.947	65.013	NM_001030	ribosomal protein S27 (RPS27)
promoter34	chr4	+	107456204	107457268	1065	245.490	28.574	NM_001142416	aminoacyl tRNA synthetase complex-interacting multifunctional protein 1 (AIMP1)
promoter35	chr11	+	66139751	66140791	1041	243.360	10.406	NM_006328	RNA binding motif protein 14 (RBM14)
promoter36	chr18	-	45272696	45273825	1130	238.050	83.750	NM_000985	ribosomal protein L17 (RPL17)
promoter37	chr20	-	5048479	5049534	1056	233.182	48.954	NM_182649	proliferating cell nuclear antigen (PCNA)
promoter38	chr10	+	103901159	103902260	1102	228.233	39.079	NM_004741	nucleolar and coiled-body phosphoprotein 1 (NOLC1)
promoter39	chr12	-	74191552	74192655	1104	226.944	2.558	NM_007043	KRR1, small subunit (SSU) processome component, homolog (yeast) (KRR1)
promoter40	chr11	+	64607284	64608398	1115	221.277	78.275	NM_006782	zinc finger protein-like 1 (ZFPL1)
promoter41	chr1	+	173234258	173235306	1049	220.743	47.838	NM_001007214	calcyclin binding protein (CACYBP)
promoter42	chr17	+	27700364	27701427	1064	220.265	32.062	NM_003457	zinc finger protein 207 (ZNF207)
promoter43	chr6	-	43134980	43136110	1131	219.762	61.228	NM_015950	mitochondrial ribosomal protein L2 (MRPL2), nuclear gene encoding mitochondrial protein
promoter44	chr19	+	12909482	12910587	1106	219.352	23.796	NM_004343	calreticulin (CALR)
promoter45	chrX	+	118591619	118592676	1058	218.303	9.009	NM_003336	ubiquitin-conjugating enzyme E2A (UBE2A)
promoter46	chr11	+	34082792	34083895	1104	214.880	83.576	NM_024662	N-acetyltransferase 10 (GCN5-related) (NAT10)
promoter47	chr17	-	62792983	62794146	1164	211.037	47.617	NM_002816	proteasome (prosome, macropain) 26S subunit, non-ATPase, 12 (PSMD12)
promoter48	chr1	+	111792635	111793792	1158	205.753	77.682	NM_001688	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit B1 (ATP5F1)
promoter49	chr20	-	2399315	2400356	1042	203.912	69.302	NM_003091	small nuclear ribonucleoprotein polypeptides B and B1 (SNRPB)
promoter50	chr14	+	21014231	21015389	1159	197.952	17.947	NM_014828	TOX high mobility group box family member 4 (TOX4)
promoter51	chr17	+	27794680	27795739	1060	197.100	21.924	NM_002815	proteasome (prosome, macropain) 26S subunit, non-ATPase, 11 (PSMD11)
promoter52	chr17	+	38178022	38179198	1177	194.943	22.152	NM_032353	vacuolar protein sorting 25 homolog (S. cerevisiae) (VPS25)
promoter53	chr3	-	49370658	49371717	1060	193.105	27.384	NM_000581	glutathione peroxidase 1 (GPX1)
promoter54	chr2	+	118287841	118288920	1080	192.989	38.646	NM_006773	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18 (DDX18)
promoter55	chr1	+	93316433	93317517	1085	188.777	33.921	NM_007358	metal response element binding transcription factor 2 (MTF2)
promoter56	chr12	-	74764607	74765682	1076	185.328	41.719	NM_004537	nucleosome assembly protein 1-like 1 (NAP1L1)
promoter57	chrX	+	77245407	77246485	1079	184.135	53.387	NM_000291	phosphoglycerate kinase 1 (PGK1)
promoter58	chr4	-	174491968	174493060	1093	181.998	37.225	NM_002129	high mobility group box 2 (HMGB2)
promoter59	chr20	+	41519002	41520121	1120	181.863	18.210	NM_006275	serine/arginine-rich splicing factor 6 (SRSF6)
promoter60	chr2	-	223228934	223229990	1057	181.696	22.298	NM_005687	phenylalanyl-tRNA synthetase, beta subunit (FARSB)
promoter61	chr7	+	107170878	107171942	1065	181.612	22.736	NM_024814	Cas-Br-M (murine) ecotropic retroviral transforming sequence-like 1 (CBLL1)
promoter62	chr22	+	36678691	36679774	1084	178.263	6.885	NM_021974	polymerase (RNA) II (DNA directed) polypeptide F (POLR2F)
promoter63	chr11	+	63509526	63510586	1061	174.438	18.496	NM_017670	OTU domain, ubiquitin aldehyde binding 1 (OTUB1)
promoter64	chr5	+	33475883	33476941	1059	173.416	19.463	NM_152295	threonyl-tRNA synthetase (TARS)
promoter65	chr17	-	45805348	45806484	1137	173.253	42.466	NM_016504	mitochondrial ribosomal protein L27 (MRPL27)
promoter66	chr21	-	34209828	34210924	1097	173.113	11.585	NM_001697	ATP synthase, H+ transporting, mitochondrial F1 complex, O subunit (ATP5O)
promoter67	chr7	-	103635561	103636656	1096	172.621	45.911	NM_002553	origin recognition complex, subunit 5 (ORC5)
promoter68	chr21	+	29317873	29318952	1080	169.906	24.923	NM_006447	ubiquitin specific peptidase 16 (USP16)
promoter69	chr6	+	26211200	26212262	1063	168.772	13.760	NM_003542	histone cluster 1, H4c (HIST1H4C)
promoter70	chr18	-	52456799	52457882	1084	167.883	2.192	NM_004786	thioredoxin-like 1 (TXNL1)
promoter71	chr6	-	83959533	83960614	1082	166.256	23.363	NM_015599	phosphoglucomutase 3 (PGM3)
promoter72	chr4	-	3503752	3504883	1132	164.927	67.918	NM_002337	low density lipoprotein receptor-related protein associated protein 1 (LRPAP1)
promoter73	chr20	+	21231049	21232121	1073	163.723	34.484	NM_012255	5'-3' exoribonuclease 2 (XRN2)
promoter74	chr1	-	37928638	37929710	1073	163.683	10.291	NM_017850	chromosome 1 open reading frame 109 (C1orf109)
promoter75	chr9	-	138414264	138415394	1131	162.891	8.463		
promoter76	chr19	-	18409823	18410885	1063	162.888	67.716	NM_016368	inositol-3-phosphate synthase 1 (ISYNA1)
promoter77	chr8	-	117837103	117838215	1113	158.505	76.097	NM_003756	eukaryotic translation initiation factor 3, subunit H (EIF3H)
promoter78	chr17	-	59932737	59933847	1111	157.369	53.009	NM_004396	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 (DDX5)
promoter79	chr6	+	142509124	142510192	1069	156.326	31.556	NM_016485	Vps20-associated 1 homolog (S. cerevisiae) (VTA1)
promoter80	chr7	+	98760518	98761611	1094	155.822	41.857	NM_006409	actin related protein 2/3 complex, subunit 1A, 41kDa (ARPC1A)
promoter81	chr2	+	27446024	27447128	1105	155.681	9.425	NM_014748	sorting nexin 17 (SNX17)
promoter82	chr3	+	140544568	140545693	1126	155.139	4.574	NM_020191	mitochondrial ribosomal protein S22 (MRPS22)
promoter83	chr8	-	66708831	66709994	1164	154.869	44.402	NM_018120	armadillo repeat containing 1 (ARMC1)
promoter84	chr10	+	1023419	1024513	1095	154.448	67.834	NM_012341	GTP binding protein 4 (GTPBP4)
promoter85	chr15	-	99652759	99653941	1183	154.252	26.529	NM_003090	small nuclear ribonucleoprotein polypeptide A' (SNRPA1)
promoter86	chr11	-	62329373	62330482	1110	151.667	39.889	NM_006362	nuclear RNA export factor 1 (NXF1)
promoter87	chr4	+	41630977	41632028	1052	146.341	9.917	NM_018126	transmembrane protein 33 (TMEM33)
promoter88	chr3	-	188006755	188007878	1124	146.316	21.377	NM_002916	replication factor C (activator 1) 4, 37kDa (RFC4)
promoter89	chr6	-	26324698	26325856	1159	146.231	24.289	NM_003518	histone cluster 1, H2bg (HIST1H2BG)
promoter90	chr7	-	127771056	127772101	1046	145.927	14.477	NM_018077	RNA binding motif protein 28 (RBM28)
promoter91	chr5	-	159778487	159779590	1104	145.062	54.526	NM_006425	SLU7 splicing factor homolog (S. cerevisiae) (SLU7)
promoter92	chr2	+	198088046	198089151	1106	144.628	20.500	NM_015387	MOB1, Mps One Binder kinase activator-like 3 (yeast) (MOBKL3)
promoter93	chr12	+	97510683	97511802	1120	143.320	52.424	NM_002635	solute carrier family 25 (mitochondrial carrier; phosphate carrier), member 3 (SLC25A3)
promoter94	chr1	+	93069211	93070309	1099	142.746	25.682	NM_000969	ribosomal protein L5 (RPL5)
promoter95	chr12	+	51685562	51686668	1107	137.641	41.541	NM_001417	eukaryotic translation initiation factor 4B (EIF4B)
promoter96	chrX	-	53466149	53467321	1173	136.617	27.945	NM_006306	structural maintenance of chromosomes 1A (SMC1A)
promoter97	chr1	-	148166138	148167271	1134	135.950	29.796	NM_181873	myotubularin related protein 11 (MTMR11)
promoter98	chr6	+	44321894	44322953	1060	135.113	12.578	NM_007355	heat shock protein 90kDa alpha (cytosolic), class B member 1 (HSP90AB1)
promoter99	chr6	-	52257359	52258468	1110	132.337	7.561	NM_002388	minichromosome maintenance complex component 3 (MCM3)
promoter100	chr6	+	7334111	7335246	1136	130.556	18.176	NM_031480	RIO kinase 1 (yeast) (RIOK1)
promoter101	chr3	+	125931063	125932122	1060	128.596	10.020	NM_000373	uridine monophosphate synthetase (UMPS)

表 3-1 続き

ID	chr	strand	start	end	length	Luc activity	sd	ID	Description
promoter102	chr19	-	50886908	50887962	1055	127.678	53.807	NM_004597	small nuclear ribonucleoprotein D2 polypeptide 16.5kDa (SNRPD2)
promoter103	chrX	+	16646731	16647880	1150	125.343	10.113	NM_032796	synapse associated protein 1 (SYAP1)
promoter104	chr19	-	1556282	1557411	1130	125.188	12.905	NM_006830	ubiquinol-cytochrome c reductase, complex III subunit XI (UQCR11)
promoter105	chrX	-	64671242	64672306	1065	123.349	3.560	NM_031206	LAS1-like (S. cerevisiae) (LAS1L)
promoter106	chr11	-	10786964	10788030	1067	122.070	11.936	NM_001418	eukaryotic translation initiation factor 4 gamma, 2 (EIF4G2)
promoter107	chr3	-	46012113	46013294	1182	120.662	23.261	NM_024513	FYVE and coiled-coil domain containing 1 (FYCO1)
promoter108	chr9	-	126217423	126218493	1071	120.354	19.306	NM_002799	proteasome (prosome, macropain) subunit, beta type, 7 (PSMB7)
promoter109	chr12	-	119368462	119369544	1083	119.108	4.087	NM_016399	TP53 regulated inhibitor of apoptosis 1 (TRIAP1)
promoter110	chrX	-	71413679	71414744	1066	118.179	1.261	NM_001007	ribosomal protein S4, X-linked (RPS4X)
promoter111	chr14	-	93617208	93618305	1098	118.099	25.118	NM_020414	DEAD (Asp-Glu-Ala-Asp) box polypeptide 24 (DDX24)
promoter112	chr19	+	52939629	52940770	1142	116.959	21.998	NM_015710	glioma tumor suppressor candidate region gene 2 (GLTSCR2)
promoter113	chr16	+	21871170	21872270	1101	115.951	35.304	NM_003366	ubiquinol-cytochrome c reductase core protein II (UQCRC2)
promoter114	chr11	+	124967013	124968063	1051	115.439	13.080	NM_152713	STT3, subunit of the oligosaccharyltransferase complex, homolog A (S. cerevisiae) (STT3A)
promoter115	chr2	+	85618855	85619932	1078	114.732	36.224	NM_005911	methionine adenosyltransferase II, alpha (MAT2A)
promoter116	chr11	+	47556250	47557355	1106	114.697	5.995	NM_004551	NADH dehydrogenase (ubiquinone) Fe-S protein 3, 30kDa (NADH-coenzyme Q reductase) (NDUFS3)
promoter117	chr20	+	18395078	18396177	1100	114.524	34.165	NM_006466	polymerase (RNA) III (DNA directed) polypeptide F, 39 kDa (POLR3F)
promoter118	chr6	-	100123129	100124234	1106	114.506	14.493	NM_005190	cyclin C (CCNC)
promoter119	chr1	+	155002967	155004157	1191	113.955	43.400	NM_005973	papillary renal cell carcinoma (translocation-associated) (PRCC)
promoter120	chr8	-	97316855	97317944	1090	113.746	26.697	NM_006294	ubiquinol-cytochrome c reductase binding protein (UQCRB)
promoter121	chr5	+	64099578	64100637	1060	113.047	11.699	NM_005869	CWC27 spliceosome-associated protein homolog (S. cerevisiae) (CWC27)
promoter122	chr8	-	27751108	27752148	1041	112.796	0.039	NM_018492	PDZ binding kinase (PBK)
promoter123	chr2	+	113910871	113912022	1152	112.393	24.963	NM_172003	COBW domain containing 2 (CBWD2)
promoter124	chr11	+	31486912	31488013	1102	109.430	4.922	NM_019040	elongation protein 4 homolog (S. cerevisiae) (ELP4)
promoter125	chr15	+	63947867	63948953	1087	109.313	38.180	NM_004663	RAB11A, member RAS oncogene family (RAB11A)
promoter126	chr17	+	38238015	38239100	1086	106.907	15.485	NM_005789	proteasome (prosome, macropain) activator subunit 3 (PA28 gamma; Ki) (PSME3)
promoter127	chr1	+	205991101	205992173	1073	105.769	17.670	NM_002389	CD46 molecule, complement regulatory protein (CD46)
promoter128	chr11	-	10519172	10520233	1062	105.405	11.680	NM_016422	ring finger protein 141 (RNF141)
promoter129	chr1	-	31542026	31543133	1108	104.600	13.344	NM_004814	small nuclear ribonucleoprotein 40kDa (U5) (SNRNP40)
promoter130	chr9	-	134271883	134272964	1082	104.475	25.210	NM_007344	transcription termination factor, RNA polymerase I (TTF1)
promoter131	chr12	+	56373233	56374322	1090	104.463	7.610	NM_006812	osteosarcoma amplified 9, endoplasmic reticulum lectin (OS9)
promoter132	chr5	-	86744379	86745474	1096	101.799	17.552	NM_001239	cyclin H (CCNH)
promoter133	chr2	+	219231738	219232794	1057	100.655	6.851	NM_004328	BCS1-like (S. cerevisiae) (BCS1L)
promoter134	chr10	+	62207325	62208367	1043	100.577	8.037	NM_001786	cyclin-dependent kinase 1 (CDK1)
promoter135	chr5	-	137938902	137939961	1060	100.570	3.242	NM_004134	heat shock 70kDa protein 9 (mortalin) (HSPA9)
promoter136	chr20	-	32354707	32355793	1087	100.069	21.645	NM_000687	adenosylhomocysteinase (AHCY)
promoter137	chr11	+	111449266	111450349	1084	99.008	19.362	NM_018195	chromosome 11 open reading frame 57 (C11orf57)
promoter138	chr20	+	39089951	39091010	1060	97.781	25.374	NM_003286	topoisomerase (DNA) I (TOP1)
promoter139	chr12	-	121316820	121317909	1090	97.364	20.090	NM_022916	vacuolar protein sorting 33 homolog A (S. cerevisiae) (VPS33A)
promoter140	chr19	+	5640347	5641410	1064	97.354	24.638	NM_015414	ribosomal protein L36 (RPL36)
promoter141	chr18	-	53440043	53441101	1059	97.134	35.097	NM_004539	asparaginyl-tRNA synthetase (NARS)
promoter142	chr7	-	73306503	73307679	1177	96.858	14.396	NM_002914	replication factor C (activator 1) 2, 40kDa (RFC2)
promoter143	chr17	+	38402889	38403942	1054	96.311	3.372	NM_000988	ribosomal protein L27 (RPL27)
promoter144	chr6	+	31740871	31741985	1115	96.298	35.781	NM_001320	casein kinase 2, beta polypeptide (CSNK2B)
promoter145	chr2	+	170148573	170149662	1090	96.277	16.690	NM_004792	peptidylprolyl isomerase G (cyclophilin G) (PPIG)
promoter146	chr19	+	62689941	62691105	1165	96.007	11.938	NM_024691	zinc finger protein 419 (ZNF419)
promoter147	chr17	+	44324235	44325280	1046	95.604	10.187	NM_005175	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C1 (subunit 9) (ATP5G1)
promoter148	chrX	-	129127332	129128408	1077	95.181	24.194	NM_004208	apoptosis-inducing factor, mitochondrion-associated, 1 (AIFM1)
promoter149	chr17	-	19821575	19822663	1089	94.288	24.505	NM_007202	A kinase (PRKA) anchor protein 10 (AKAP10)
promoter150	chr11	-	62170539	62171661	1123	93.825	24.356	NM_198334	glucosidase, alpha; neutral AB (GANAB)
promoter151	chr3	+	171166318	171167427	1110	93.626	17.598	NM_003262	SEC62 homolog (S. cerevisiae) (SEC62)
promoter152	chr1	+	39976190	39977281	1092	93.251	26.654	NM_006112	peptidylprolyl isomerase E (cyclophilin E) (PPIE)
promoter153	chr19	-	12905376	12906517	1142	93.153	7.440	NM_004461	phenylalanyl-tRNA synthetase, alpha subunit (FARSA)
promoter154	chr1	-	215870836	215871968	1133	93.112	8.850	NM_018040	G patch domain containing 2 (GPATCH2)
promoter155	chrX	+	54572643	54573598	956	92.483	3.647	NM_019067	guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3L)
promoter156	chr20	-	2592684	2593826	1143	91.723	5.969	NM_006899	isocitrate dehydrogenase 3 (NAD+) beta (IDH3B)
promoter157	chr8	+	67502891	67503983	1093	91.622	3.919	NM_015169	RRS1 ribosome biogenesis regulator homolog (S. cerevisiae) (RRS1)
promoter158	chr1	-	114925620	114926725	1106	91.382	3.624	NM_005872	breast carcinoma amplified sequence 2 (BCAS2)
promoter159	chr3	+	23821491	23822592	1102	90.994	5.950	NM_003341	ubiquitin-conjugating enzyme E2E 1 (UBE2E1)
promoter160	chr1	+	222367508	222368554	1047	90.341	12.548	NM_015176	F-box protein 28 (FBXO28)
promoter161	chr7	-	32496415	32497458	1044	88.934	12.711	NM_012322	LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae) (LSM5)
promoter162	chr17	-	25642981	25644112	1132	88.882	3.829	NM_000386	bleomycin hydrolase (BLMH)
promoter163	chr10	-	12278011	12279085	1075	88.475	16.890	NM_014142	nudix (nucleoside diphosphate linked moiety X)-type motif 5 (NUDT5)
promoter164	chr5	-	137576780	137577872	1093	88.402	34.280	NM_004661	cell division cycle 23 homolog (S. cerevisiae) (CDC23)
promoter165	chr4	-	4342596	4343757	1162	88.280	33.553	NM_017816	Ly1 antibody reactive homolog (mouse) (LYAR)
promoter166	chr4	-	121207336	121208388	1053	88.212	13.239	NM_002358	MAD2 mitotic arrest deficient-like 1 (yeast) (MAD2L1)
promoter167	chrX	+	70419292	70420358	1067	86.263	13.269	NM_007363	non-POU domain containing, octamer-binding (NONO)
promoter168	chr11	+	10728462	10729529	1068	85.995	30.684	NM_014633	Ctr9, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae) (CTR9)
promoter169	chr2	-	174968542	174969494	953	85.468	24.969	NM_004882	corepressor interacting with RBPJ, 1 (CIR1)
promoter170	chr17	+	46584975	46586094	1120	84.736	22.773	NM_000269	non-metastatic cells 1, protein (NM23A) expressed in (NME1)
promoter171	chr14	+	66043687	66044766	1080	84.692	14.790	NM_020806	gephyrin (GPHN)
promoter172	chr6	+	34832332	34833420	1089	84.277	2.841	NM_003093	small nuclear ribonucleoprotein polypeptide C (SNRPC)
promoter173	chr1	-	149115581	149116712	1132	83.764	4.843	NM_001668	aryl hydrocarbon receptor nuclear translocator (ARNT)
promoter174	chr19	+	16082540	16083633	1094	83.414	3.712	NM_005370	RAB8A, member RAS oncogene family (RAB8A)
promoter175	chrX	+	153643309	153644394	1086	81.766	15.902	NM_001363	dyskeratosis congenita 1, dyskerin (DKC1)
promoter176	chr12	-	107649262	107650359	1098	81.554	5.496	NM_014325	coronin, actin binding protein, 1C (CORO1C)
promoter177	chr15	-	47234898	47236043	1146	81.545	3.008	NM_004236	COP9 constitutive photomorphogenic homolog subunit 2 (Arabidopsis) (COPS2)
promoter178	chr3	+	102774789	102775879	1091	81.456	25.640	NM_020357	PEST proteolytic signal containing nuclear protein (PCNP)
promoter179	chr10	-	27189714	27190817	1104	81.452	7.760	NM_005470	abl-interactor 1 (ABI1)
promoter180	chr22	-	40414686	40415764	1079	81.435	9.532	NM_001003796	NHP2 non-histone chromosome protein 2-like 1 (S. cerevisiae) (NHP2L1)
promoter181	chr22	-	38045583	38046643	1061	79.726	24.615	NM_000967	ribosomal protein L3 (RPL3)
promoter182	chr3	+	135006361	135007518	1158	79.065	6.386	NM_021203	signal recognition particle receptor, B subunit (SRPRB)
promoter183	chr22	-	28279530	28280662	1133	78.284	5.383	NM_003678	THO complex 5 (THOC5)
promoter184	chr17	-	44847121	44848209	1089	78.052	6.262	NM_002634	prohibitin (PHB)
promoter185	chr17	+	45220130	45221224	1095	78.017	7.240	NM_007067	MYST histone acetyltransferase 2 (MYST2)
promoter186	chrX	-	74292737	74293790	1054	77.992	10.890	NM_004299	ATP-binding cassette, sub-family B (MDR/TAP), member 7 (ABCB7)
promoter187	chr16	+	30680207	30681286	1080	77.403	6.965	NM_014771	ring finger protein 40 (RNF40)
promoter188	chr3	-	52779861	52780962	1102	76.322	12.424	NM_003157	NIMA (never in mitosis gene a)-related kinase 4 (NEK4)
promoter189	chr1	+	159437801	159438946	1146	76.000	8.521	NM_001166159	NADH dehydrogenase (ubiquinone) Fe-S protein 2, 49kDa (NADH-coenzyme Q reductase) (NDUFS2)
promoter190	chr21	+	44376993	44378071	1079	75.927	11.223	NM_004649	chromosome 21 open reading frame 33 (C21orf33)
promoter191	chr11	-	89595698	89596826	1129	75.691	4.781	NM_012124	cysteine and histidine-rich domain (CHORD) containing 1 (CHORDC1)
promoter192	chr12	+	4627598	4628661	1064	75.237	11.883	NM_005002	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9, 39kDa (NDUFA9)
promoter193	chr6	-	26164555	26165626	1072	74.956	22.938	NM_005319	histone cluster 1, H1c (HIST1H1C)
promoter194	chr1	-	154941824	154942935	1112	74.142	6.674	NM_001878	cellular retinoic acid binding protein 2 (CRABP2)
promoter195	chr5	+	74015783	74016907	1125	73.961	12.969	NM_000521	hexosaminidase B (beta polypeptide) (HEXB)
promoter196	chr15	-	38118523	38119579	1057	73.780	6.637	NM_003134	signal recognition particle 14kDa (homologous Alu RNA binding protein) (SRP14)
promoter197	chr21	-	14677137	14678283	1147	73.064	8.183	NM_006948	heat shock protein 70kDa family, member 13 (HSPA13)
promoter198	chr12	+	84791248	84792335	1088	72.772	8.426	NM_006183	neurotensin (NTS)
promoter199	chr1	-	158161811	158162898	1088	72.349	14.329	NM_003564	transgelin 2 (TAGLN2)
promoter200	chrX	+	48316869	48317929	1061	71.850	18.910	NM_006743	RNA binding motif (RNP1, RRM) protein 3 (RBM3)
promoter201	chr6	-	41996707	41997795	1089	71.462	17.923	NM_004275	mediator complex subunit 20 (MED20)
promoter202	chr1	+	28434240	28435347	1108	71.298	1.606	NM_016311	ATPase inhibitory factor 1 (ATPIF1)

表 3-1 続き

ID	chr	strand	start	end	length	Luc activity	sd	ID	Description
promoter203	chr5	+	140006658	140007711	1054	70.959	7.554	NM_006083	IK cytokine, down-regulator of HLA II (IK)
promoter204	chr2	+	216681351	216682449	1099	69.655	3.724	NM_021141	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining) (XRCC5)
promoter205	chr13	+	33289319	33290384	1066	69.629	8.248	NM_002915	replication factor C (activator 1) 3, 38kDa (RFC3)
promoter206	chr2	-	198007837	198008925	1089	69.536	18.896	NM_012433	splicing factor 3b, subunit 1, 155kDa (SF3B1)
promoter207	chr14	-	19992716	19993766	1051	68.880	4.046	NM_017807	O-sialoglycoprotein endopeptidase (OSGEP)
promoter208	chr12	-	107478960	107480020	1061	68.875	3.170	NM_014706	squamous cell carcinoma antigen recognized by T cells 3 (SART3)
promoter209	chr14	-	68934621	68935703	1083	68.849	6.462	NM_004450	enhancer of rudimentary homolog (Drosophila) (ERH)
promoter210	chr20	+	61966098	61967246	1149	68.428	8.433	NM_003288	tumor protein D52-like 2 (TPD52L2)
promoter211	chr19	+	43800735	43801815	1081	67.695	5.106	NM_013234	eukaryotic translation initiation factor 3, subunit K (EIF3K)
promoter212	chr19	-	7914533	7915591	1059	67.457	7.487	NM_006351	inner mitochondrial membrane 44 homolog (yeast) (TIMM44)
promoter213	chr14	-	20921844	20923019	1176	67.016	38.076	NM_007192	suppressor of Ty 16 homolog (S. cerevisiae) (SUPT16H)
promoter214	chr1	-	114248903	114249951	1049	65.791	0.348	NM_006594	adaptor-related protein complex 4, beta 1 subunit (AP4B1)
promoter215	chr19	+	52325002	52326097	1096	65.735	5.942	NM_005500	SUMO1 activating enzyme subunit 1 (SAE1)
promoter216	chr8	-	24869920	24870990	1071	64.273	1.743	NM_006158	neurofilament, light polypeptide (NEFL)
promoter217	chr8	-	124477679	124478862	1184	64.188	10.627	NM_014109	ATPase family, AAA domain containing 2 (ATAD2)
promoter218	chr1	-	36702423	36703496	1074	63.740	14.597	NM_031280	mitochondrial ribosomal protein S15 (MRPS15)
promoter219	chr4	+	84595276	84596350	1075	63.427	1.455	NM_016067	mitochondrial ribosomal protein S18C (MRPS18C)
promoter220	chr1	-	115102021	115103080	1060	62.509	21.232	NM_007158	cold shock domain containing E1, RNA-binding (CSDE1)
promoter221	chr2	-	27739849	27740922	1074	62.493	1.928	NM_014860	suppressor of Ty 7 (S. cerevisiae)-like (SUPT7L)
promoter222	chr6	-	94185866	94186974	1109	62.372	2.549	NM_004440	EPH receptor A7 (EPHA7)
promoter223	chr19	-	57099925	57101052	1128	62.349	13.462	NM_023074	zinc finger protein 649 (ZNF649)
promoter224	chr4	+	40952735	40953788	1054	62.044	18.992	NM_004181	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase) (UCHL1)
promoter225	chr11	+	73559044	73560102	1059	61.975	6.219	NM_016147	protein phosphatase methylesterase 1 (PPME1)
promoter226	chr1	+	32888438	32889518	1081	61.408	5.972	NM_005610	retinoblastoma binding protein 4 (RBBP4)
promoter227	chr7	-	133794327	133795431	1105	61.359	2.662	NM_001628	aldo-keto reductase family 1, member B1 (aldose reductase) (AKR1B1)
promoter228	chr6	+	44462495	44463583	1089	61.233	11.598	NM_001253	CDC5 cell division cycle 5-like (S. pombe) (CDC5L)
promoter229	chr14	+	23632398	23633480	1083	61.198	4.899	NM_004563	phosphoenolpyruvate carboxykinase 2 (mitochondrial) (PCK2)
promoter230	chr10	+	14919352	14920432	1081	60.881	9.186	NM_016299	heat shock 70kDa protein 14 (HSPA14)
promoter231	chr19	+	8360220	8361346	1127	60.848	5.292	NM_004218	RAB11B, member RAS oncogene family (RAB11B)
promoter232	chr15	+	41905555	41906645	1091	60.778	9.855	NM_024908	WD repeat domain 76 (WDR76)
promoter233	chr15	-	86811429	86812584	1156	60.285	4.426	NM_022163	mitochondrial ribosomal protein L46 (MRPL46)
promoter234	chr1	-	149298574	149299643	1070	59.824	12.388	NM_020239	CDC42 small effector 1 (CDC42SE1)
promoter235	chr8	-	120937161	120938231	1071	59.811	16.935	NM_024094	defective in sister chromatid cohesion 1 homolog (S. cerevisiae) (DSCC1)
promoter236	chr11	-	47404479	47405547	1069	59.718	3.588	NM_002804	proteasome (prosome, macropain) 26S subunit, ATPase, 3 (PSMC3)
promoter237	chr11	-	18504892	18505997	1106	59.586	8.427	NM_006292	tumor susceptibility gene 101 (TSG101)
promoter238	chr3	-	150858353	150859426	1074	59.301	6.076	NM_015472	WW domain containing transcription regulator 1 (WWTR1)
promoter239	chr1	-	210654660	210655793	1134	58.891	8.760	NM_018252	transmembrane protein 206 (TMEM206)
promoter240	chr1	-	39111496	39112558	1063	58.815	2.888	NM_012333	c-myc binding protein (MYCBP)
promoter241	chr4	+	41686330	41687479	1150	58.592	11.661	NM_006345	solute carrier family 30 (zinc transporter), member 9 (SLC30A9)
promoter242	chr10	-	53129192	53130348	1157	58.425	4.985	NM_015235	cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa, tau variant (CSTF2T)
promoter243	chr11	+	65575487	65576577	1091	58.352	8.792	NM_006842	splicing factor 3b, subunit 2, 145kDa (SF3B2)
promoter244	chr1	-	6182128	6183202	1075	58.340	9.998	NM_000983	ribosomal protein L22 (RPL22)
promoter245	chr12	+	106602748	106603814	1067	58.139	5.487	NM_007062	PWP1 homolog (S. cerevisiae) (PWP1)
promoter246	chr5	+	125963569	125964663	1095	58.080	17.237	NM_032177	phosphorylated adaptor for RNA export (PHAX)
promoter247	chr12	-	55368254	55369320	1067	57.465	4.206	NM_006601	prostaglandin E synthase 3 (cytosolic) (PTGES3)
promoter248	chr11	+	62379198	62380244	1047	56.433	3.692	NM_002394	solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2 (SLC3A2)
promoter249	chr1	+	205560684	205561815	1132	56.322	9.530	NM_000574	CD55 molecule, decay accelerating factor for complement (Cromer blood group) (CD55)
promoter250	chr7	+	133981159	133982241	1083	55.439	3.241	NM_001724	2,3-bisphosphoglycerate mutase (BPGM)
promoter251	chr4	+	38859554	38860674	1121	54.961	13.183	NM_025132	WD repeat domain 19 (WDR19)
promoter252	chr2	-	38457785	38458857	1073	54.197	1.792	NM_022374	atlastin GTPase 2 (ATL2)
promoter253	chr20	+	41652047	41653220	1174	54.081	8.439	NM_016004	intraflagellar transport 52 homolog (Chlamydomonas) (IFT52)
promoter254	chr6	-	170704227	170705281	1055	53.905	14.748	NM_002793	proteasome (prosome, macropain) subunit, beta type, 1 (PSMB1)
promoter255	chr1	-	85946440	85947601	1162	53.802	7.280	NM_017953	zinc finger, HIT-type containing 6 (ZNHIT6)
promoter256	chr7	-	42938119	42939233	1115	53.653	6.818	NM_002787	proteasome (prosome, macropain) subunit, alpha type, 2 (PSMA2)
promoter257	chr20	+	54476148	54477290	1143	53.513	5.988	NM_016407	chromosome 20 open reading frame 43 (C20orf43)
promoter258	chr12	-	54497626	54498768	1143	53.449	13.546	NM_033082	SAP domain containing ribonucleoprotein (SARNP)
promoter259	chr8	+	109524072	109525199	1128	53.095	14.320	NM_014673	tetratricopeptide repeat domain 35 (TTC35)
promoter260	chr11	-	87710389	87711472	1084	53.047	3.854	NM_001814	cathepsin C (CTSC)
promoter261	chr6	-	26140140	26141268	1129	53.032	9.681	NM_003537	histone cluster 1, H3b (HIST1H3B)
promoter262	chr1	-	158579535	158580603	1069	52.957	3.035	NM_004371	coatomer protein complex, subunit alpha (COPA)
promoter263	chr12	-	81276174	81277271	1098	52.853	4.556	NM_014167	coiled-coil domain containing 59 (CCDC59)
promoter264	chr5	+	52130727	52131879	1153	52.836	16.649	NM_015946	pelota homolog (Drosophila) (PELO)
promoter265	chr19	+	12777447	12778530	1084	52.779	21.186	NM_006397	ribonuclease H2, subunit A (RNASEH2A)
promoter266	chr12	-	94953440	94954489	1050	51.242	3.076	NM_000895	leukotriene A4 hydrolase (LTA4H)
promoter267	chr12	-	47396624	47397723	1100	50.272	13.438	NM_001240	cyclin T1 (CCNT1)
promoter268	chr17	-	53439472	53440563	1092	50.003	7.892	NM_006924	serine/arginine-rich splicing factor 1 (SRSF1)
promoter269	chr3	+	138062843	138063900	1058	49.804	6.781	NM_006153	NCK adaptor protein 1 (NCK1)
promoter270	chr9	+	134894993	134896042	1050	49.723	6.124	NM_012087	general transcription factor IIIC, polypeptide 5, 63kDa (GTF3C5)
promoter271	chr11	+	124048085	124049144	1060	49.677	23.929	NM_017425	sperm autoantigenic protein 17 (SPA17)
promoter272	chr5	+	118815185	118816244	1060	49.345	3.009	NM_000414	hydroxysteroid (17-beta) dehydrogenase 4 (HSD17B4)
promoter273	chr20	+	33592256	33593301	1046	49.248	7.225	NM_015966	ERGIC and golgi 3 (ERGIC3)
promoter274	chr6	-	8047653	8048759	1107	49.181	3.280	NM_004280	eukaryotic translation elongation factor 1 epsilon 1 (EEF1E1)
promoter275	chr2	+	99319367	99320461	1095	48.658	3.415	NM_015904	eukaryotic translation initiation factor 5B (EIF5B)
promoter276	chr1	+	153981280	153982379	1100	48.540	5.741	NR_024117	misato homolog 2 pseudogene (MSTO2P)
promoter277	chr5	-	140050973	140052092	1120	48.485	2.630	NM_002109	histidyl-tRNA synthetase (HARS)
promoter278	chr17	-	70662035	70663107	1073	48.137	4.553	NM_016185	hematological and neurological expressed 1 (HN1)
promoter279	chr6	-	10946588	10947755	1168	47.879	3.477	NM_005906	male germ cell-associated kinase (MAK)
promoter280	chr6	+	26306747	26307885	1139	47.778	3.018	NM_003522	histone cluster 1, H2bf (HIST1H2BF)
promoter281	chr2	-	197372463	197373607	1145	46.076	10.028	NM_012086	general transcription factor IIIC, polypeptide 3, 102kDa (GTF3C3)
promoter282	chr15	-	64435910	64437073	1164	45.742	10.081	NM_017858	TIMELESS interacting protein (TIPIN)
promoter283	chr3	-	129325191	129326278	1088	45.547	15.231	NM_003707	RuvB-like 1 (E. coli) (RUVBL1)
promoter284	chr1	-	118273631	118274688	1058	45.419	1.597	NM_017686	ganglioside induced differentiation associated protein 2 (GDAP2)
promoter285	chr19	-	44014177	44015242	1066	45.337	17.253	NM_001398	enoyl CoA hydratase 1, peroxisomal (ECH1)
promoter286	chr6	+	10855116	10856182	1067	45.069	11.683	NM_030969	transmembrane protein 14B (TMEM14B)
promoter287	chr13	+	20611751	20612807	1057	44.533	11.891	NM_005870	Sin3A-associated protein, 18kDa (SAP18)
promoter288	chr3	-	53356466	53357597	1132	44.322	0.578	NM_018403	DCP1 decapping enzyme homolog A (S. cerevisiae) (DCP1A)
promoter289	chr17	+	20969908	20971001	1094	44.263	6.782	NM_015510	dehydrogenase/reductase (SDR family) member 7B (DHRS7B)
promoter290	chr22	+	29122036	29123175	1140	44.024	3.062	NM_012429	SEC14-like 2 (S. cerevisiae) (SEC14L2)
promoter291	chr17	-	27252755	27253798	1044	44.023	5.519	NM_018428	UTP6, small subunit (SSU) processome component, homolog (yeast) (UTP6)
promoter292	chr10	+	64562108	64563283	1176	43.945	6.894	NM_030759	nuclear receptor binding factor 2 (NRBF2)
promoter293	chr9	+	132558063	132559120	1058	43.682	6.482	NM_014285	exosome component 2 (EXOSC2)
promoter294	chr12	-	87059837	87060951	1115	43.346	3.568	NM_025114	centrosomal protein 290kDa (CEP290)
promoter295	chr1	+	22250827	22251870	1044	43.296	5.371	NM_001791	cell division cycle 42 (GTP binding protein, 25kDa) (CDC42)
promoter296	chr6	+	57289407	57290572	1166	43.264	6.180	NM_000947	primase, DNA, polypeptide 2 (58kDa) (PRIM2)
promoter297	chr6	-	33655909	33656958	1050	42.684	7.504	NM_001188	BCL2-antagonist/killer 1 (BAK1)
promoter298	chr1	-	67668543	67669603	1061	42.601	7.116	NM_015640	SERPINE1 mRNA binding protein 1 (SERBP1)
promoter299	chr3	+	10042182	10043321	1140	41.293	2.538	NM_033084	Fanconi anemia, complementation group D2 (FANCD2)
promoter300	chr2	+	170362638	170363763	1126	41.131	10.679	NM_003142	Sjogren syndrome antigen B (autoantigen La) (SSB)
promoter301	chr7	+	128165630	128166741	1112	40.882	20.223	NM_001219	calumenin (CALU)
promoter302	chr1	+	120055032	120056177	1146	40.831	3.901	NM_006623	phosphoglycerate dehydrogenase (PHGDH)
promoter303	chr8	+	125619605	125620724	1120	40.815	10.285	NM_005005	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9, 22kDa (NDUFB9)

表 3-1 続き

ID	chr	strand	start	end	length	Luc activity	sd	ID	Description
promoter304	chr1	-	45224667	45225797	1131	40.288	3.234	NM_020365	eukaryotic translation initiation factor 2B, subunit 3 gamma, 58kDa (EIF2B3)
promoter305	chr22	-	39582443	39583553	1111	40.209	1.768	NM_003932	suppression of tumorigenicity 13 (colon carcinoma) (Hsp70 interacting protein) (ST13)
promoter306	chr6	+	33046804	33047885	1082	39.966	11.078	NM_001199455	bromodomain containing 2 (BRD2)
promoter307	chr15	+	73414524	73415631	1108	39.855	5.153	NM_017828	COMM domain containing 4 (COMMD4)
promoter308	chr2	+	201755149	201756231	1083	39.671	12.278	NM_001230	caspase 10, apoptosis-related cysteine peptidase (CASP10)
promoter309	chr9	-	123172151	123173303	1153	39.507	4.915	NM_004099	stomatin (STOM)
promoter310	chr1	+	117403536	117404655	1120	39.231	7.558	NM_003594	transcription termination factor, RNA polymerase II (TTF2)
promoter311	chr10	+	112316485	112317591	1107	39.195	7.894	NM_005445	structural maintenance of chromosomes 3 (SMC3)
promoter312	chr16	+	66901448	66902626	1179	38.908	5.260	NM_019023	protein arginine methyltransferase 7 (PRMT7)
promoter313	chr11	-	111142179	111143258	1080	38.822	1.880	NM_002716	protein phosphatase 2, regulatory subunit A, beta (PPP2R1B)
promoter314	chr16	+	68015075	68016243	1169	38.301	3.245	NM_030579	cytochrome b5 type B (outer mitochondrial membrane) (CYB5B)
promoter315	chr16	+	72887280	72888345	1066	37.150	2.054	NM_002811	proteasome (prosome, macropain) 26S subunit, non-ATPase, 7 (PSMD7)
promoter316	chr15	+	86810973	86812083	1111	37.121	8.619	NM_022839	mitochondrial ribosomal protein S11 (MRPS11)
promoter317	chr15	+	40574195	40575307	1113	37.017	1.929	NM_003825	synaptosomal-associated protein, 23kDa (SNAP23)
promoter318	chr12	-	47868959	47870067	1109	36.911	10.818	NM_006009	tubulin, alpha 1a (TUBA1A)
promoter319	chr4	+	178467079	178468148	1070	36.806	1.116	NM_018248	nei endonuclease VIII-like 3 (E. coli) (NEIL3)
promoter320	chr14	+	19992197	19993311	1115	36.521	2.258	NM_001641	APEX nuclease (multifunctional DNA repair enzyme) 1 (APEX1)
promoter321	chr5	-	175748082	175749191	1110	36.405	8.004	NM_016391	NOP16 nucleolar protein homolog (yeast) (NOP16)
promoter322	chr6	+	26509555	26510613	1059	36.340	4.633	NM_007048	butyrophilin, subfamily 3, member A1 (BTN3A1)
promoter323	chr7	+	91995045	91996205	1161	35.914	3.130	NM_032120	chromosome 7 open reading frame 64 (C7orf64)
promoter324	chr21	-	29313415	29314465	1051	35.870	1.913	NM_016940	RWD domain containing 2B (RWDD2B)
promoter325	chr1	-	23635616	23636709	1094	35.081	2.985	NM_017707	ArfGAP with SH3 domain, ankyrin repeat and PH domain 3 (ASAP3)
promoter326	chr16	-	66902155	66903344	1190	35.011	2.975	NM_032178	solute carrier family 7, member 6 opposite strand (SLC7A6OS)
promoter327	chr1	+	149637738	149638782	1045	34.514	2.841	NM_002796	proteasome (prosome, macropain) subunit, beta type, 4 (PSMB4)
promoter328	chr16	+	56982913	56984037	1125	34.265	3.927	NM_022770	GINS complex subunit 3 (Psf3 homolog) (GINS3)
promoter329	chr21	+	31952925	31953967	1043	33.384	2.331	NM_000454	superoxide dismutase 1, soluble (SOD1)
promoter330	chr7	-	135063283	135064430	1148	33.068	4.663	NM_012450	solute carrier family 13 (sodium/sulfate symporters), member 4 (SLC13A4)
promoter331	chr17	-	53282214	53283343	1130	33.020	7.076	NM_016070	mitochondrial ribosomal protein S23 (MRPS23)
promoter332	chr8	+	55209379	55210512	1134	32.830	7.453	NM_014175	mitochondrial ribosomal protein L15 (MRPL15)
promoter333	chrX	-	53477890	53478983	1094	32.333	3.715	NM_004493	hydroxysteroid (17-beta) dehydrogenase 10 (HSD17B10)
promoter334	chr12	+	2855713	2856821	1109	32.001	3.489	NR_027363	chromosome 12 open reading frame 32 (C12orf32)
promoter335	chr1	+	92536278	92537392	1115	30.821	2.612	NM_024813	RNA polymerase II associated protein 2 (RPAP2)
promoter336	chr5	+	89805509	89806621	1113	30.705	2.864	NM_006467	polymerase (RNA) III (DNA directed) polypeptide G (32kD) (POLR3G)
promoter337	chr19	+	7607036	7608155	1120	29.718	7.923	NM_006949	syntaxin binding protein 2 (STXBP2)
promoter338	chr12	-	42438647	42439807	1161	29.417	2.457	NM_031292	pseudouridylate synthase 7 homolog (S. cerevisiae)-like (PUS7L)
promoter339	chr1	-	111793219	111794301	1083	29.099	4.036	NM_024102	WD repeat domain 77 (WDR77)
promoter340	chr17	-	9419874	9421034	1161	29.088	2.603	NM_004853	syntaxin 8 (STX8)
promoter341	chrX	+	118888914	118889960	1047	28.655	1.742	NM_004541	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 1, 7.5kDa (NDUFA1)
promoter342	chr22	+	30124627	30125708	1082	28.337	6.403	NM_004147	developmentally regulated GTP binding protein 1 (DRG1)
promoter343	chr15	+	89278587	89279687	1101	28.306	3.482	NM_018671	unc-45 homolog A (C. elegans) (UNC45A)
promoter344	chr22	+	39930274	39931455	1182	28.170	1.827	NM_031488	l(3)mbt-like 2 (Drosophila) (L3MBTL2)
promoter345	chr3	+	37009051	37010168	1118	27.964	2.167	NM_000249	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) (MLH1)
promoter346	chr2	+	187058335	187059443	1109	27.569	10.264	NM_018471	zinc finger CCCH-type containing 15 (ZC3H15)
promoter347	chrX	-	152712814	152713972	1159	27.299	1.272	NM_004135	isocitrate dehydrogenase 3 (NAD+) gamma (IDH3G)
promoter348	chr22	-	29317687	29318836	1150	27.274	3.074	NM_014303	pescadillo homolog 1, containing BRCT domain (zebrafish) (PES1)
promoter349	chr9	-	130124335	130125500	1166	27.097	2.680	NM_015679	TruB pseudouridine (psi) synthase homolog 2 (E. coli) (TRUB2)
promoter350	chr11	-	36267417	36268505	1089	26.626	2.512	NM_014186	COMM domain containing 9 (COMMD9)
promoter351	chr9	+	26945432	26946605	1174	26.142	4.346	NM_025103	intraflagellar transport 74 homolog (Chlamydomonas) (IFT74)
promoter352	chr12	+	105691591	105692733	1143	25.942	0.556	NM_018157	resistance to inhibitors of cholinesterase 8 homolog B (C. elegans) (RIC8B)
promoter353	chr1	-	149586211	149587327	1117	25.457	3.004	NM_000449	regulatory factor X, 5 (influences HLA class II expression) (RFX5)
promoter354	chr5	+	112223937	112225025	1089	25.358	5.441	NM_003135	signal recognition particle 19kDa (SRP19)
promoter355	chr14	+	22844936	22846062	1127	25.228	4.786	NM_004050	BCL2-like 2 (BCL2L2)
promoter356	chr2	+	136004548	136005677	1130	25.186	3.390	NM_015361	R3H domain containing 1 (R3HDM1)
promoter357	chr12	-	6830554	6831642	1089	24.742	2.841	NM_031299	cell division cycle associated 3 (CDCA3)
promoter358	chr14	-	44673291	44674366	1076	24.710	2.532	NM_002013	FK506 binding protein 3, 25kDa (FKBP3)
promoter359	chr17	+	40581309	40582390	1082	24.572	4.379	NM_006460	hexamethylene bis-acetamide inducible 1 (HEXIM1)
promoter360	chr1	+	165455826	165456868	1043	24.256	3.071	NM_002697	POU class 2 homeobox 1 (POU2F1)
promoter361	chr14	+	30097202	30098287	1086	24.096	3.100	NM_017769	G2/M-phase specific E3 ubiquitin protein ligase (G2E3)
promoter362	chr2	+	15648485	15649549	1065	24.069	4.150	NM_004939	DEAD (Asp-Glu-Ala-Asp) box polypeptide 1 (DDX1)
promoter363	chr1	-	210275334	210276494	1161	23.889	1.178	NM_015434	integrator complex subunit 7 (INTS7)
promoter364	chr6	-	33364842	33365924	1083	23.574	1.103	NM_005452	WD repeat domain 46 (WDR46)
promoter365	chr14	-	54727900	54728961	1062	22.374	6.613	NM_014750	discs, large (Drosophila) homolog-associated protein 5 (DLGAP5)
promoter366	chr3	+	25805578	25806767	1190	22.354	1.801	NM_017897	3-oxoacyl-ACP synthase, mitochondrial (OXSM)
promoter367	chr10	-	30064669	30065823	1155	22.347	1.240	NM_003174	supervillin (SVIL)
promoter368	chrX	+	114700829	114701940	1112	21.919	7.241	NM_005032	plastin 3 (PLS3)
promoter369	chr1	+	75961994	75963137	1144	21.879	2.513	NM_000016	acyl-CoA dehydrogenase, C-4 to C-12 straight chain (ACADM)
promoter370	chr3	+	69215788	69216914	1127	20.237	5.150	NM_006407	ADP-ribosylation-like factor 6 interacting protein 5 (ARL6IP5)
promoter371	chr2	+	73313988	73315106	1119	19.952	8.577	NM_006429	chaperonin containing TCP1, subunit 7 (eta) (CCT7)
promoter372	chr17	+	59257602	59258677	1076	19.681	1.335	NM_002805	proteasome (prosome, macropain) 26S subunit, ATPase, 5 (PSMC5)
promoter373	chr4	+	17186991	17188061	1071	19.496	3.923	NM_015907	leucine aminopeptidase 3 (LAP3)
promoter374	chr1	-	16885157	16886256	1100	19.440	5.777		
promoter375	chr10	+	7869220	7870283	1064	19.395	3.663	NM_005174	ATP synthase, H+ transporting, mitochondrial F1 complex, gamma polypeptide 1 (ATP5C1)
promoter376	chr17	+	73721022	73722072	1051	19.387	0.520	NM_001168	baculoviral IAP repeat containing 5 (BIRC5)
promoter377	chr15	-	88578015	88579161	1147	19.075	1.484	NM_006384	calcium and integrin binding 1 (calmyrin) (CIB1)
promoter378	chr19	-	1046166	1047339	1174	18.961	1.563	NM_002695	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa (POLR2E)
promoter379	chr4	+	140805483	140806561	1079	18.628	1.644	NM_002413	microsomal glutathione S-transferase 2 (MGST2)
promoter380	chr16	-	45280734	45281840	1107	18.284	2.488	NM_018206	vacuolar protein sorting 35 homolog (S. cerevisiae) (VPS35)
promoter381	chr1	-	28113722	28114779	1058	17.666	3.348	NM_002946	replication protein A2, 32kDa (RPA2)
promoter382	chr3	-	109423801	109424942	1142	15.832	2.584	NM_018010	intraflagellar transport 57 homolog (Chlamydomonas) (IFT57)
promoter383	chr2	-	37312071	37313207	1137	15.564	1.467	NM_005760	CCAAT/enhancer binding protein (C/EBP), zeta (CEBPZ)
promoter384	chr9	+	85784565	85785659	1095	15.294	3.084	NM_024945	RMI1, RecQ mediated genome instability 1, homolog (S. cerevisiae) (RMI1)
promoter385	chr15	-	40352737	40353845	1109	15.083	1.951	NM_015497	transmembrane protein 87A (TMEM87A)
promoter386	chr10	-	98263528	98264656	1129	14.953	2.626	NM_012465	tolloid-like 2 (TLL2)
promoter387	chr4	+	37503779	37504841	1063	14.022	2.864	NM_018290	phosphoglucomutase 2 (PGM2)
promoter388	chr5	-	132100994	132102083	1090	13.876	3.127	NM_007054	kinesin family member 3A (KIF3A)
promoter389	chrX	+	152711938	152712999	1062	13.446	1.937	NM_001204527	signal sequence receptor, delta (translocon-associated protein delta) (SSR4)
promoter390	chr7	-	132417179	132418323	1145	13.118	1.013	NM_017812	coiled-coil-helix-coiled-coil-helix domain containing 3 (CHCHD3)
promoter391	chr2	-	202774759	202775818	1060	12.859	1.932		
promoter392	chr6	+	37428800	37429950	1151	11.974	1.966	NM_003958	ring finger protein 8 (RNF8)
promoter393	chr11	+	16715586	16716698	1113	11.303	0.644	NM_014267	chromosome 11 open reading frame 58 (C11orf58)
promoter394	chr14	-	22468432	22469479	1048	11.148	2.637	NM_006109	protein arginine methyltransferase 5 (PRMT5)
promoter395	chr5	-	150118591	150119712	1122	10.699	1.274	NM_016221	dynactin 4 (p62) (DCTN4)
promoter396	chr11	-	47226769	47227865	1097	9.998	1.349	NM_001610	acid phosphatase 2, lysosomal (ACP2)
promoter397	chr1	+	111482659	111483705	1047	9.976	3.610	NM_001007794	choline/ethanolamine phosphotransferase 1 (CEPT1)
promoter398	chr10	-	69267787	69268830	1044	9.754	1.750	NM_021800	DnaJ (Hsp40) homolog, subfamily C, member 12 (DNAJC12)
promoter399	chr17	+	34610059	34611109	1051	9.052	2.764	NM_000981	ribosomal protein L19 (RPL19)
promoter400	chr11	-	34894306	34895393	1088	8.705	2.252	NM_015957	APAF1 interacting protein (APIP)
promoter401	chr17	+	32379466	32380579	1114	7.948	0.292	NM_012138	apoptosis antagonizing transcription factor (AATF)
promoter402	chr6	-	36950616	36951693	1078	7.104	1.313	NM_016059	peptidylprolyl isomerase (cyclophilin)-like 1 (PPIL1)
promoter403	chr6	-	33028726	33029805	1080	6.851	1.782	NM_006120	major histocompatibility complex, class II, DM alpha (HLA-DMA)
promoter404	chr7	-	91995619	91996693	1075	6.273	0.238	NM_000466	peroxisomal biogenesis factor 1 (PEX1)

表 3-1 続き

ID	chr	strand	start	end	length	Luc activity	sd	ID	Description
promoter405	chrX	+	46961665	46962801	1137	5.468	1.422	NM_033018	cyclin-dependent kinase 16 (CDK16)
promoter406	chr3	-	8979945	8981064	1120	5.417	0.615	NM_020165	RAD18 homolog (S. cerevisiae) (RAD18)
promoter407	chr3	-	63984563	63985681	1119	5.328	2.380	NM_014814	proteasome (prosome, macropain) 26S subunit, non-ATPase, 6 (PSMD6)
promoter408	chr8	-	120754650	120755797	1148	5.315	0.980		
promoter409	chr14	+	23491855	23492948	1094	4.547	0.901	NM_021004	dehydrogenase/reductase (SDR family) member 4 (DHRS4)
promoter410	chr1	+	109039268	109040444	1177	4.426	0.378	NM_018061	PRP38 pre-mRNA processing factor 38 (yeast) domain containing B (PRPF38B)
promoter411	chr16	+	29714787	29715830	1044	4.022	0.489		
promoter412	chr2	+	58331117	58332234	1118	3.940	0.415		
promoter413	chr6	+	125515681	125516824	1144	3.715	0.441	NM_003287	tumor protein D52-like 1 (TPD52L1)
promoter414	chr3	+	129995948	129997014	1067	3.696	1.179	NM_004637	RAB7A, member RAS oncogene family (RAB7A)
promoter415	chr10	-	43605748	43606849	1102	3.416	0.946	NR_002726	heterogeneous nuclear ribonucleoprotein A3 pseudogene 1 (HNRNPA3P1)
promoter416	chr4	-	38998548	38999653	1106	3.163	0.632	NM_002913	replication factor C (activator 1) 1, 145kDa (RFC1)
promoter417	chr12	-	6547642	6548695	1054	3.058	0.418	NM_006170	NOP2 nucleolar protein homolog (yeast) (NOP2)
promoter418	chr6	+	160101504	160102587	1084	2.845	0.362	NM_005891	acetyl-CoA acetyltransferase 2 (ACAT2)
promoter419	chr7	+	36411603	36412668	1066	2.736	0.207	NM_018685	anillin, actin binding protein (ANLN)
promoter420	chr5	+	43684023	43685142	1120	2.662	0.607	NM_012343	nicotinamide nucleotide transhydrogenase (NNT)
promoter421	chr16	+	70687426	70688480	1055	2.568	0.320	NM_014003	DEAH (Asp-Glu-Ala-His) box polypeptide 38 (DHX38)
promoter422	chr3	+	182067673	182068747	1075	2.390	0.376		
promoter423	chr15	-	56973462	56974620	1159	2.215	0.753	NM_024755	SAFB-like, transcription modulator (SLTM)
promoter424	chr5	-	157173703	157174801	1099	2.126	0.747	NM_014666	clathrin interactor 1 (CLINT1)
promoter425	chr14	-	77291099	77292198	1100	2.105	0.300	NM_012245	SNW domain containing 1 (SNW1)
promoter426	chr1	+	164126162	164127261	1100	2.074	0.695	NM_012474	uridine-cytidine kinase 2 (UCK2)
promoter427	chr8	-	48964115	48965191	1077	2.006	0.253	NM_006904	protein kinase, DNA-activated, catalytic polypeptide (PRKDC)
promoter428	chr9	-	124714035	124715111	1077	1.966	0.506	NM_006626	zinc finger and BTB domain containing 6 (ZBTB6)
promoter429	chr10	-	134966351	134967394	1044	1.868	0.474	NM_006659	tubulin, gamma complex associated protein 2 (TUBGCP2)
promoter430	chr12	+	56177286	56178329	1044	1.752	0.324	NM_004990	methionyl-tRNA synthetase (MARS)
promoter431	chr3	-	188492056	188493103	1048	1.729	0.142	NM_001879	mannan-binding lectin serine peptidase 1 (C4/C2 activating component of Ra-reactive factor) (MASP1)
promoter432	chr20	+	35840167	35841295	1129	1.356	0.125	NM_030877	catenin, beta like 1 (CTNNBL1)
promoter433	chr14	-	23725865	23726969	1105	1.353	0.349	NM_024658	importin 4 (IPO4)
promoter434	chr10	-	94257259	94258325	1067	1.307	0.303	NM_004969	insulin-degrading enzyme (IDE)
promoter435	chr3	+	128817618	128818699	1082	1.290	0.131	NM_004526	minichromosome maintenance complex component 2 (MCM2)
promoter436	chr17	+	54642427	54643517	1091	1.235	0.137	NM_018149	chromosome 17 open reading frame 71 (C17orf71)
promoter437	chr11	+	34428189	34429248	1060	1.168	0.293	NM_001752	catalase (CAT)
promoter438	chr14	-	69422514	69423553	1040	1.052	0.105		
promoter439	chr11	-	46411535	46412605	1071	1.039	0.397	NM_017749	autophagy/beclin-1 regulator 1 (AMBRA1)
promoter440	chr2	-	86127764	86128819	1056	0.938	0.044	NM_015425	polymerase (RNA) I polypeptide A, 194kDa (POLR1A)
promoter441	chr7	+	107319926	107321012	1087	0.934	0.129	NM_000108	dihydrolipoamide dehydrogenase (DLD)
promoter442	chrX	-	85227199	85228239	1041	0.922	0.371		
promoter443	chr20	+	30183492	30184617	1126	0.910	0.075	NM_014742	transmembrane 9 superfamily protein member 4 (TM9SF4)
promoter444	chr12	+	117057114	117058221	1108	0.666	0.147	NM_002567	phosphatidylethanolamine binding protein 1 (PEBP1)
promoter445	chr19	+	39865953	39867077	1125	0.619	0.225	NM_018443	zinc finger protein 302 (ZNF302)
promoter446	chr10	+	128706055	128707119	1065	0.606	0.116	NM_001380	dedicator of cytokinesis 1 (DOCK1)
promoter447	chr6	-	13804906	13806031	1126	0.584	0.261	NM_005493	RAN binding protein 9 (RANBP9)
promoter448	chr14	+	22858438	22859522	1085	0.552	0.119	NM_004643	poly(A) binding protein, nuclear 1 (PABPN1)
promoter449	chr5	-	131641066	131642109	1044	0.494	0.142		
promoter450	chr11	-	95756846	95757970	1125	0.474	0.076	NM_024725	coiled-coil domain containing 82 (CCDC82)
promoter451	chr3	+	68766601	68767681	1081	0.369	0.109		

表 3-2 ゲノムのランダム領域ライブラリーの詳細

ID	chr	strand	start	end	length	Luc activity	sd
random1	chr8	-	3817154	3818031	878	14.292	1.155
random2	chr17	-	35732194	35733153	960	10.954	0.850
random3	chr15	+	40481566	40482505	940	10.507	1.724
random4	chr12	-	76514042	76515113	1072	4.896	0.239
random5	chr3	-	139270131	139271097	967	4.403	1.101
random6	chr3	-	51700731	51701801	1071	4.345	1.132
random7	chr9	-	106196736	106197672	937	4.085	0.092
random8	chr9	+	76969348	76970411	1064	3.837	1.326
random9	chr6	+	34167376	34168285	910	3.637	0.313
random10	chr17	+	12241200	12242238	1039	3.622	1.936
random11	chr18	-	7450408	7451488	1081	3.498	0.990
random12	chr18	-	7450410	7451490	1081	2.732	0.321
random13	chr3	-	112125218	112126221	1004	2.659	0.318
random14	chr1	+	18607295	18608342	1048	2.655	0.336
random15	chr17	-	31388875	31389744	870	2.560	0.037
random16	chr2	-	17943837	17944692	856	2.529	0.404
random17	chr17	+	52548996	52550066	1071	2.456	1.177
random18	chr12	-	129352671	129353591	921	2.385	1.010
random19	chr6	+	169012702	169013861	1160	2.332	0.228
random20	chr12	+	37423728	37424667	940	2.307	0.921
random21	chr8	+	30095014	30096117	1104	2.209	0.270
random22	chr12	-	107705975	107706886	912	2.122	0.360
random23	chr21	+	31247844	31248788	945	2.108	0.700
random24	chr1	-	65595886	65596928	1043	2.004	0.441
random25	chr13	-	51439678	51440603	926	1.886	1.534
random26	chr10	+	110598940	110599889	950	1.878	0.274
random27	chr14	+	94361543	94362500	958	1.860	0.127
random28	chr9	-	5547347	5548435	1089	1.775	0.091
random29	chr1	-	65595884	65596926	1043	1.773	0.403
random30	chr3	+	58615351	58616271	921	1.769	0.546
random31	chr4	-	150404754	150405638	885	1.759	0.445
random32	chr8	+	9468178	9469320	1143	1.746	0.189
random33	chr1	+	37002124	37003108	985	1.745	0.467
random34	chr3	-	127566662	127567671	1010	1.713	0.420
random35	chr14	+	94361545	94362503	959	1.695	0.254
random36	chr2	+	201919836	201920867	1032	1.688	0.133
random37	chr6	+	47769426	47770442	1017	1.687	0.688
random38	chr8	+	9468178	9469329	1152	1.673	0.342
random39	chr3	-	46133527	46134521	995	1.647	0.613
random40	chr20	+	49104339	49105314	976	1.557	0.308
random41	chr4	-	93799524	93800538	1015	1.553	0.181
random42	chr7	-	80046261	80047311	1051	1.521	0.129
random43	chr1	-	214295359	214296376	1018	1.425	0.257
random44	chr3	-	101758789	101759818	1030	1.421	0.158
random45	chr2	+	207980539	207981522	984	1.405	0.275
random46	chr2	+	65803170	65804068	899	1.387	0.220
random47	chr11	+	62383861	62384799	939	1.361	0.240
random48	chr9	+	128855366	128856440	1075	1.360	0.223
random49	chr6	-	166817464	166818380	917	1.352	0.234
random50	chr3	+	84823480	84824478	999	1.346	0.265
random51	chr1	-	24355462	24356562	1101	1.324	0.167
random52	chr10	-	118584692	118585892	1201	1.315	0.276
random53	chr15	-	55628382	55629295	914	1.271	0.063
random54	chr2	+	143168766	143169792	1027	1.261	0.630
random55	chr17	+	22415809	22416681	873	1.258	0.106
random56	chr10	-	27280151	27281072	922	1.248	0.090
random57	chr7	-	3204150	3205193	1044	1.235	0.245
random58	chr6	+	46061149	46062061	913	1.217	0.122
random59	chr5	+	170695223	170696253	1031	1.213	0.133
random60	chr9	-	116016848	116017717	870	1.207	0.215
random61	chr3	+	195073304	195074441	1138	1.201	0.140
random62	chr14	+	61743628	61744530	903	1.195	0.204
random63	chr11	-	104245604	104246736	1133	1.192	0.327
random64	chr12	+	2493471	2494354	884	1.156	0.256
random65	chr1	-	31910475	31911652	1178	1.155	0.238
random66	chr9	-	114959966	114961131	1166	1.116	0.129
random67	chr2	+	143168759	143169794	1036	1.095	0.099
random68	chr12	-	107705973	107706884	912	1.080	0.199
random69	chr22	-	46582584	46583711	1128	1.077	0.229
random70	chr3	+	184720528	184721499	972	1.065	0.084
random71	chr4	-	37122436	37123459	1024	1.058	0.170
random72	chr20	-	58637878	58638978	1101	1.010	0.265
random73	chr1	+	214139655	214140685	1031	1.007	0.108
random74	chr12	+	10985893	10986718	826	0.995	0.266
random75	chr17	-	60638542	60639496	955	0.981	0.101
random76	chr15	-	59304807	59305978	1172	0.956	0.408
random77	chr9	+	122065343	122066319	977	0.949	0.140
random78	chr21	+	31247846	31248790	945	0.947	0.029
random79	chr14	-	21550421	21551353	933	0.944	0.110
random80	chrX	-	33112241	33113104	864	0.935	0.126
random81	chr5	+	78052495	78053486	992	0.932	0.174
random82	chr1	-	4306588	4307537	950	0.932	0.207
random83	chr13	+	106070256	106071309	1054	0.909	0.276
random84	chr21	+	32940803	32941708	906	0.901	0.115
random85	chr1	+	67870777	67871766	990	0.874	0.079
random86	chr15	-	59121301	59122227	927	0.850	0.099
random87	chr9	+	124918584	124919478	895	0.846	0.189
random88	chr11	-	35301962	35302929	968	0.820	0.184
random89	chr6	-	138941816	138942742	927	0.815	0.075
random90	chr12	-	124195020	124195831	812	0.811	0.291
random91	chr1	+	175197509	175198593	1085	0.770	0.085
random92	chrX	+	30662841	30663950	1110	0.766	0.071
random93	chr1	-	25948796	25949815	1020	0.761	0.090
random94	chr3	+	195073302	195074439	1138	0.754	0.199
random95	chr12	+	22078642	22079534	893	0.745	0.105
random96	chr12	+	37423730	37424669	940	0.739	0.190
random97	chr20	-	59630606	59631704	1099	0.729	0.171
random98	chr8	+	97578603	97579799	1197	0.726	0.070
random99	chr1	+	33778911	33779812	902	0.718	0.306
random100	chr6	+	46061147	46062059	913	0.712	0.328
random101	chr13	+	66819072	66820089	1018	0.709	0.157
random102	chr11	+	82102069	82103141	1073	0.704	0.036
random103	chr10	-	3292695	3293679	985	0.704	0.091
random104	chr1	+	155755547	155756702	1156	0.686	0.164
random105	chr9	-	123274130	123275090	961	0.664	0.264
random106	chr3	+	84823478	84824476	999	0.652	0.109
random107	chr6	-	5691876	5692697	822	0.649	0.149
random108	chr14	-	54279028	54280042	1015	0.647	0.053
random109	chr5	+	170695297	170696214	918	0.637	0.113
random110	chr13	-	41022481	41023541	1061	0.629	0.082
random111	chr3	-	50152694	50153611	918	0.626	0.310
random112	chrY	-	10254496	10255612	1117	0.623	0.145
random113	chr6	+	47769428	47770444	1017	0.612	0.034
random114	chr3	+	157723034	157723898	865	0.609	0.287
random115	chr1	-	25948798	25949817	1020	0.601	0.083
random116	chr7	-	104279619	104280619	1001	0.585	0.181
random117	chr16	-	85577013	85577983	971	0.582	0.013
random118	chr7	-	148615958	148617003	1046	0.565	0.126
random119	chr2	+	174157886	174159021	1136	0.563	0.192
random120	chr12	-	26409485	26410318	834	0.560	0.100
random121	chr4	-	78574421	78575466	1046	0.554	0.052
random122	chr5	-	144837808	144838992	1185	0.552	0.017
random123	chr22	+	22514250	22515143	894	0.552	0.041
random124	chr2	-	49159828	49160972	1145	0.551	0.093
random125	chr5	-	44529161	44530349	1189	0.549	0.056
random126	chr14	-	74060449	74061453	1005	0.545	0.117
random127	chr4	-	2481410	2482505	1096	0.544	0.209
random128	chr1	+	67870779	67871768	990	0.541	0.076
random129	chr4	-	72315071	72315990	920	0.535	0.070
random130	chr9	-	29462534	29463446	913	0.526	0.187
random131	chr6	+	34636755	34637837	1083	0.524	0.155
random132	chr5	+	146730579	146731739	1161	0.521	0.037
random133	chr2	+	208397702	208398742	1041	0.519	0.137
random134	chr21	+	31949097	31949947	851	0.517	0.030
random135	chr9	-	96613724	96614731	1008	0.514	0.042
random136	chr6	+	155739061	155740052	992	0.512	0.091
random137	chr8	-	70436425	70437443	1019	0.503	0.114
random138	chr6	-	144534901	144535834	934	0.496	0.078
random139	chr10	+	21088386	21089392	1007	0.488	0.088
random140	chr16	-	29218615	29219789	1175	0.488	0.023
random141	chr3	+	12479458	12480543	1086	0.487	0.154
random142	chr20	-	51292834	51293701	868	0.475	0.219
random143	chr7	+	145219160	145219968	809	0.467	0.106
random144	chr2	+	208397704	208398744	1041	0.457	0.123
random145	chr16	+	27801466	27802480	1015	0.456	0.130
random146	chr3	+	12479458	12480543	1086	0.455	0.077
random147	chr6	-	7144326	7145306	981	0.452	0.019
random148	chr20	-	57401406	57402338	933	0.451	0.067
random149	chr8	-	75335810	75336799	990	0.448	0.039
random150	chr2	-	5292825	5293980	1156	0.442	0.051
random151	chr6	-	67554247	67555244	998	0.441	0.081
random152	chr6	+	155739063	155740054	992	0.439	0.167
random153	chr17	+	32160350	32161454	1105	0.439	0.407
random154	chr6	+	3291380	3292517	1138	0.437	0.051
random155	chr10	-	32036084	32037150	1067	0.436	0.080
random156	chr17	+	44088517	44089540	1024	0.433	0.115
random157	chr15	+	23419318	23420072	755	0.431	0.060
random158	chr5	-	44682892	44683867	976	0.429	0.052
random159	chr1	-	59685288	59686252	965	0.420	0.043
random160	chr9	-	121995499	121996445	947	0.420	0.081
random161	chr2	-	50349147	50350313	1167	0.419	0.086
random162	chr7	-	88269596	88270505	910	0.412	0.092
random163	chr4	-	93282083	93283109	1027	0.410	0.142
random164	chr20	+	32817337	32818258	922	0.407	0.180
random165	chr13	-	64737232	64738248	1017	0.404	0.094
random166	chr1	+	155176298	155177279	982	0.399	0.144
random167	chr7	-	140848282	140849236	955	0.398	0.024
random168	chr8	-	81854834	81855932	1099	0.398	0.117
random169	chr11	-	120340797	120341978	1182	0.392	0.030
random170	chr7	+	119327066	119328117	1052	0.390	0.064
random171	chr3	-	188830545	188831528	984	0.388	0.054
random172	chrX	+	68366356	68367234	879	0.382	0.122

表 3-2 続き

ID	chr	strand	start	end	length	Luc activity	sd
random173	chr1	+	160591763	160592882	1120	0.379	0.130
random174	chr14	+	28197414	28198395	982	0.378	0.033
random175	chr9	+	130553950	130554858	909	0.369	0.097
random176	chr4	-	4847915	4849098	1184	0.366	0.098
random177	chr1	+	203266496	203267464	969	0.364	0.045
random178	chr12	+	91421042	91421891	850	0.363	0.021
random179	chr13	-	109796265	109797395	1131	0.359	0.104
random180	chr13	-	41022486	41023536	1051	0.359	0.041
random181	chr3	-	46133519	46134523	1005	0.354	0.075
random182	chr8	+	72385746	72386708	963	0.353	0.140
random183	chr17	+	60638548	60639483	936	0.350	0.157
random184	chr6	+	33782714	33783826	1113	0.349	0.043
random185	chr17	-	21267475	21268343	869	0.348	0.234
random186	chr4	-	2481409	2482503	1095	0.335	0.042
random187	chr9	-	123274128	123275088	961	0.334	0.045
random188	chr6	+	131011081	131012013	933	0.331	0.032
random189	chr16	-	66654727	66655588	862	0.327	0.066
random190	chr13	-	66819064	66820090	1027	0.327	0.102
random191	chrX	+	68366358	68367237	880	0.326	0.046
random192	chr3	-	54499114	54500228	1115	0.324	0.044
random193	chr10	-	32036151	32037085	935	0.323	0.038
random194	chr22	+	38841207	38842110	904	0.320	0.024
random195	chr1	-	14321248	14322299	1052	0.317	0.057
random196	chr12	+	15385591	15386543	953	0.310	0.143
random197	chr12	-	53637272	53638254	983	0.309	0.006
random198	chr14	-	21550402	21551351	950	0.309	0.027
random199	chr4	-	111469236	111470168	933	0.302	0.057
random200	chr3	-	28415295	28416134	840	0.300	0.065
random201	chr3	-	113196645	113197693	1049	0.299	0.047
random202	chr7	+	119880019	119881028	1010	0.297	0.091
random203	chr8	-	108937809	108938656	848	0.296	0.031
random204	chr13	+	42796888	42798035	1148	0.294	0.063
random205	chr12	+	128352603	128353605	1003	0.293	0.028
random206	chr3	-	141653255	141654179	925	0.292	0.036
random207	chr13	-	23622170	23623029	860	0.291	0.118
random208	chr13	-	109796314	109797328	1015	0.286	0.098
random209	chr18	+	5700611	5701764	1154	0.280	0.089
random210	chr11	-	104245628	104246736	1109	0.276	0.030
random211	chr3	-	113196640	113197704	1065	0.272	0.119
random212	chr3	+	56176839	56177819	981	0.269	0.012
random213	chr18	+	2232133	2233179	1047	0.269	0.076
random214	chr9	+	130553952	130554860	909	0.262	0.008
random215	chr20	+	19165120	19166168	1049	0.260	0.021
random216	chr16	+	49868775	49869758	984	0.250	0.001
random217	chr21	-	34896422	34897468	1047	0.249	0.057
random218	chr8	-	72385748	72386710	963	0.245	0.040
random219	chr20	+	11928613	11929766	1154	0.241	0.114
random220	chr18	+	23495404	23496329	926	0.238	0.068
random221	chr18	+	38398178	38399007	830	0.236	0.101
random222	chr16	-	29218617	29219791	1175	0.233	0.036
random223	chr18	+	5700612	5701766	1155	0.230	0.033
random224	chr1	-	119056770	119057854	1085	0.229	0.004
random225	chr6	+	129683902	129684753	852	0.228	0.041
random226	chr9	+	94669155	94670279	1125	0.227	0.045
random227	chr21	-	34896417	34897472	1056	0.225	0.029
random228	chr13	+	107891109	107892189	1081	0.225	0.011
random229	chr6	+	159805315	159806367	1053	0.224	0.021
random230	chr19	+	34166330	34167180	851	0.222	0.009
random231	chr15	-	91368306	91369346	1041	0.221	0.056
random232	chr15	-	36921394	36922259	866	0.221	0.039
random233	chr10	-	32642001	32643041	1041	0.217	0.042
random234	chr14	-	95802460	95803557	1098	0.216	0.054
random235	chr18	+	38398176	38399005	830	0.210	0.021
random236	chr18	-	27577770	27578954	1185	0.209	0.057
random237	chr1	+	78847119	78848093	975	0.194	0.003
random238	chr4	+	136348525	136349337	813	0.193	0.037
random239	chr3	-	29469349	29470318	970	0.183	0.035
random240	chr7	+	34382862	34383897	1036	0.181	0.090
random241	chr15	-	91368308	91369348	1041	0.171	0.016
random242	chr3	+	124996559	124997562	1004	0.161	0.014
random243	chr8	+	104433067	104434259	1193	0.160	0.047
random244	chr16	-	6428075	6429070	996	0.159	0.013
random245	chr3	-	192149273	192150312	1040	0.145	0.027
random246	chr21	+	35961522	35962601	1080	0.070	0.009
random247	chr11	+	12885159	12886286	1128	0.065	0.024
random248	chr11	-	12885154	12886310	1157	0.061	0.011

表 3-3 lncRNA プロモーターライブラリーの詳細

ID	chr	strand	start	end	length	Luc activity	sd	Gene	Description
ncRNA1	chr20	+	1253137	1254197	1061	260.987	43.206	AK055993	
ncRNA2	chr20	-	5934328	5935434	1107	43.508	2.887	AK058185	
ncRNA3	chr20	+	58146010	58147095	1086	31.334	2.342	AK309218	
ncRNA4	chr3	+	184647212	184648263	1052	12.407	0.425	AK057000	
ncRNA5	chr18	+	64532586	64533692	1107	6.405	1.236	NM_001093729	coiled-coil domain containing 102B (CCDC102B)
ncRNA6	chr20	-	577102	578104	1003	4.094	0.691	AK098418	
ncRNA7	chr22	-	27757312	27758398	1087	2.914	0.294	AK095488	
ncRNA8	chr8	-	11002470	11003565	1096	2.906	1.210	AK056119	
ncRNA9	chr21	-	46007918	46009016	1099	2.297	0.792	AK097593	
ncRNA10	chr17	-	50760029	50761082	1054	1.879	0.662	AK001871	
ncRNA11	chr9	-	108904914	108906053	1140	1.794	0.284	AK097706	
ncRNA12	chr17	-	60377506	60378587	1082	1.760	0.733	AK098507	
ncRNA13	chr20	+	43138937	43139992	1056	1.689	0.238	AK027088	
ncRNA14	chr1	-	199973073	199974252	1180	1.630	0.044	AK056048	
ncRNA15	chr21	+	45388362	45389499	1138	1.497	0.576	AK027227	
ncRNA16	chr21	-	34806248	34807371	1124	1.412	0.324	AK095442	
ncRNA17	chr22	+	39246759	39247881	1123	1.187	0.141	AK097769	
ncRNA18	chr14	-	36925289	36926435	1147	1.157	0.404	AK026823	
ncRNA19	chr1	-	166658334	166659498	1165	1.048	0.339	AK097492	
ncRNA20	chr21	+	39116901	39118029	1129	0.797	0.181	AK000535	
ncRNA21	chr22	+	44114596	44115632	1037	0.720	0.078	AK026239	
ncRNA22	chr22	+	32450030	32451172	1143	0.716	0.116	AK057493	
ncRNA23	chr20	-	47417088	47418222	1135	0.710	0.106	AK055386	
ncRNA24	chr20	-	23266667	23267757	1091	0.694	0.179	AK000935	
ncRNA25	chr1	-	146157705	146158717	1013	0.634	0.041	AK091688	
ncRNA26	chr10	+	87326539	87327619	1081	0.583	0.264	AK097655	
ncRNA27	chr15	-	32312077	32313180	1104	0.535	0.046	AK056019	
ncRNA28	chr8	+	110379017	110380184	1168	0.515	0.152	AK057158	
ncRNA29	chr22	-	39323687	39324831	1145	0.412	0.127	AK025156	
ncRNA30	chr22	+	39691180	39692243	1064	0.345	0.057	AK090764	
ncRNA31	chr9	+	127703070	127704249	1180	0.328	0.020	AK027170	
ncRNA32	chr21	+	18296821	18297958	1138	0.315	0.037	AK095296	
ncRNA33	chr20	-	4784553	4785729	1177	0.294	0.068	AK025668	
ncRNA34	chr3	+	20403792	20404886	1095	0.286	0.056	AK026452	
ncRNA35	chr3	+	144127282	144128372	1091	0.274	0.043	AK096143	

3.2 転写因子結合配列(transcription factor binding site, TFBS)の探索

ルシフェラーゼ活性を測定したDNA断片中に存在するTFBSの配列の探索を行った. TFBSの候補配列の探索には, position weight matrix (PWM)を用いたマトリックサーチを行った. マトリックスサーチにはPWMおよび閾値のデータセットとしてTRANSFAC2008.3のvertebrate_non_redundant_minFP.prfを利用した[44]. これは220種類のPWMで構成されており, false positiveを最小にするような閾値のデータセットである. 同名の転写因子, 結合サイトの場合はグループ化し, 192種類のTFBSグループを作成した. 見つかったクローン数が4つ以下の場合は解析には含めなかった. 最終的に167種類のTFBSグループを解析対象にした. クローニングされた約1kbの領域中に存在していたTFBSの種類の平均は30.2個であった.

3.3　重回帰分析によるプロモーター活性予測モデルの構築

予測されたTFBSと計測したルシフェラーゼ活性データを用いて，DNA配列とプロモーター活性の関係式の導出を試みた．考えられる最も単純な方法である線形和モデルとしてルシフェラーゼ活性の説明を試みた．DNA断片の有するプロモーター活性をそれぞれのTFBSの持つ転写活性化能への寄与の大きさのスコアの積としたモデルと定義した（図3-2）．実際には対数変換したルシフェラーゼ活性データを各TFBSの有するプロモーター活性への寄与のスコアの線形としたモデルとして，DNA一次配列からのプロモーター活性の説明を試みた．これはBussemakerらの重回帰モデルやMEDによるモデルと類似した転写活性化能予測モデルである[31-32]．

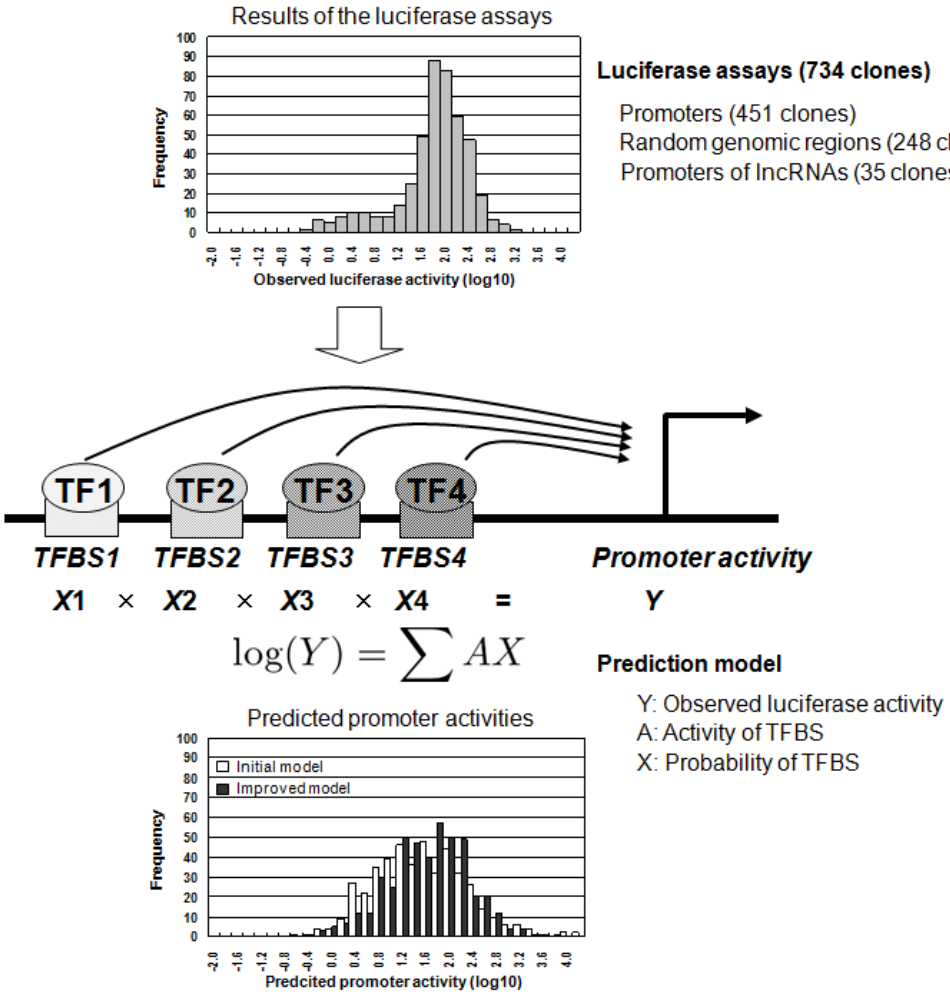


図3-2 プロモーター活性予測モデルの概略

プロモーター活性予測モデルの概略を示した．ルシフェラーゼ活性の測定値の分布（上段）予測されたプロモーター活性の分布（下段）を表わす．

プロモーター活性予測モデルのパラメーターの計算には重回帰分析の手法を用いた. 重回帰分析によって得られた各TFBSのスコアを用いて, DNA断片のプロモーター活性の予測値を算出した. 実測値と予測値の比較によりモデルの精度の検証を行った. つまり予測値が実測値の値に近ければモデルの精度が良いことを意味する. 評価には相関係数(Pearson's correlation coefficient)を利用した. 実験値と予測値との相関係数を算出した結果, *r*=0.82であった(図3-3A). またプロモーターのみではr=0.58, ランダム領域のみではr=0.25であった. 実験値の5倍以内の範囲内では約72%のクローンを予測することができた(図3-3B). TFBSの存在の有無だけを用いた線形和という単純なモデルではあるが, ある程度の精度でプロモーター活性を計算できた.

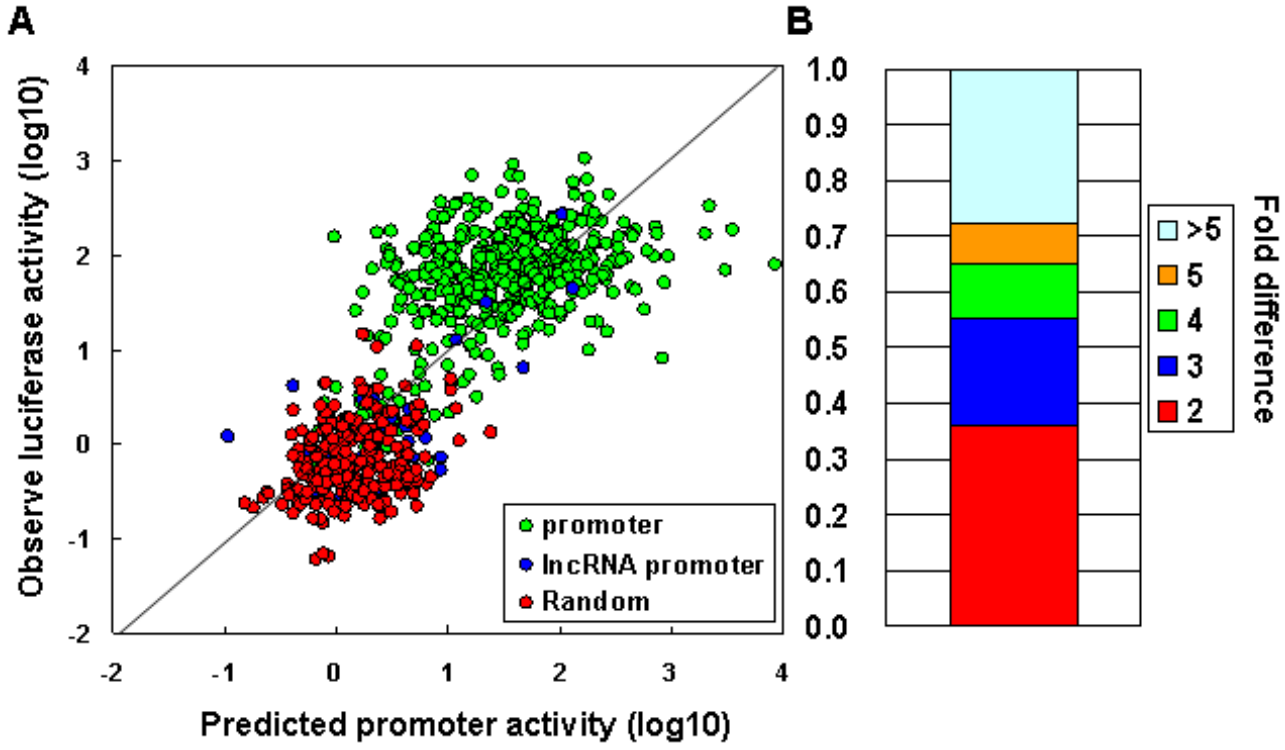


図3-3 プロモーター活性予測モデルの精度

(A)予測値と実測値との散布図, *x*軸はプロモーター活性予測値(log10), *y*軸はルシフェラーゼ活性(log10), (B)実測値からの予測値の範囲.

3.4 プロモーター活性予測モデルの改善

重回帰分析によってプロモーター活性の説明がある程度可能であると考えられた. そこで, さらに予測精度の向上が可能であるか検討した. 幾つかのプロモーター活性に関わりがあると考えられるパラメーターを導入することでプロモーター活性予測モデルの改善を試みた.

3.4.1　DNA 親和性スコアを用いたチューニング

TRANSFACのマトリックス検索時のスコアを検討した. TFBSのマトリックスサーチにはPWMがよく用いられる[43-44, 54]. PWMは位置特異的な塩基の出現確率をスコア化したもので, DNA配列を与えると配列に応じたスコアが与えられる. 通常のマトリックスサーチではある閾値よりも高い値を与えたDNA配列をTFBSとしている. PWMのスコアは高いほどコンセンサス配列に近いため, 高いスコアを与えたTFBSは転写因子との親和性・または結合する確率は高いと考えられる. そこでPWMのスコアをDNAへの親和性・確率値としたスコアとして捉えることで計算精度の向上が可能であるかどうか試みた. TFのDNAへの結合は確率的な事象として捉えられる. マトリックススコアに応じたシグモイドカーブをと捉えることができる[55-56]. マトリックスコアをDNAへの親和性へのスコアを変換する方法として, 線形で近似する方法がDasらにより報告されている[38-39]. 本研究ではこの手法を改変して利用した. TRANSFACのfalse positiveを最小にする閾値を親和性スコア0とした. マトリックススコアとともに親和性スコアも比例して直線的に上昇し, 最大値1を取る. 親和性スコアが最大値1をとるときのTRANSFACスコアの条件を計算により求めた. 重回帰分析による予測値と実験値との比較により最も計算精度の良い条件を探した. TFBSごとに条件を変えてモデル構築, プロモーター活性の予測値算出を行い, 予測値と実測値の相関係数の計算を行い, 相関係数を最大にする条件を最適な条件とした. 各TFBSにおいて最適な条件を適用した場合, 相関係数で評価したモデルの精度はr=0.84となった. 各TFBSの条件は表3-5にまとめた.

3.4.2　TFBS の位置情報を利用したチューニング

次に, TFBSのDNA上の存在位置を考慮した. TFBSとTSSの位置関係は一様では無いと考えられている. 実際に, あるTFBSについては転写開始点付近の数100bpの領域中に多く存在するTFBSが知られている[57]. すなわち転写開始点付近に機能的なTFBSがより多く存在しているといえる. また逆に機能的ではないと考えられるTFBSについては計算から除くことで予測精度の向上が可能かどうか検討した. そこでTFBSの存在位置のバイアスを考慮に入れる方法として, TFBSのTSSからの位置による閾値を導入した. 転写活性化能を測定したDNA断片の3'端を基準に100bpごとに区切り, TFBSを探索する領域を転写活性化能の測定をしたDNA断片の3'端から100bpずつ上流に広げた. 最も良い計算精度与えた領域をそのTFBSの最適な領域とした. この操作をTFBSごとに繰り返し, すべてのTFBSについて最適な条件を決定した(表3-5). 各TFBSについて最適な条件を適用することで相関係数をr=0.84からr=0.88に向上させることができた.

3.4.3　変数選択

TFBS167種類の中には予測モデルにおいてDNA断片のプロモーター活性の説明への寄与の大きさがTFBSによって異なり，モデルに必要な変数とそうではない変数が存在している．そこでAIC（Akaike's information criterion）を用いたbackward stepwise variable selectionによりプロモーター活性を説明への寄与の大きいTFBSの選択行い，説明可能な最小の変数の組み合わせを選択した．その結果，85種類のTFBSが情報量を最大にする変数の組み合わせとして選択された．選ばれたTFBSを表3-5で示す．選ばれた85種類のTFBSを用いた予測モデルの相関係数は0.86であり，より少ない変数数で同程度のプロモーター活性の説明が可能であった．またモデル構築には用いていない未知なDNA配列に対する予測に対しても改善が認められた(詳細は3.5.1で説明する)．

改善後のモデルでは実験値と予測値との相関係数がr=0.87となり，実験値の5倍以内の範囲で86%の予測可能であった（図3-4　図3-5）．またプロモータークローン内，ランダムクローン内の相関係数はそれぞれr=0.66，r=0.32であった．

予測モデルにおいて転写への強い寄与を示した転写因子の例を表3-4に示した．

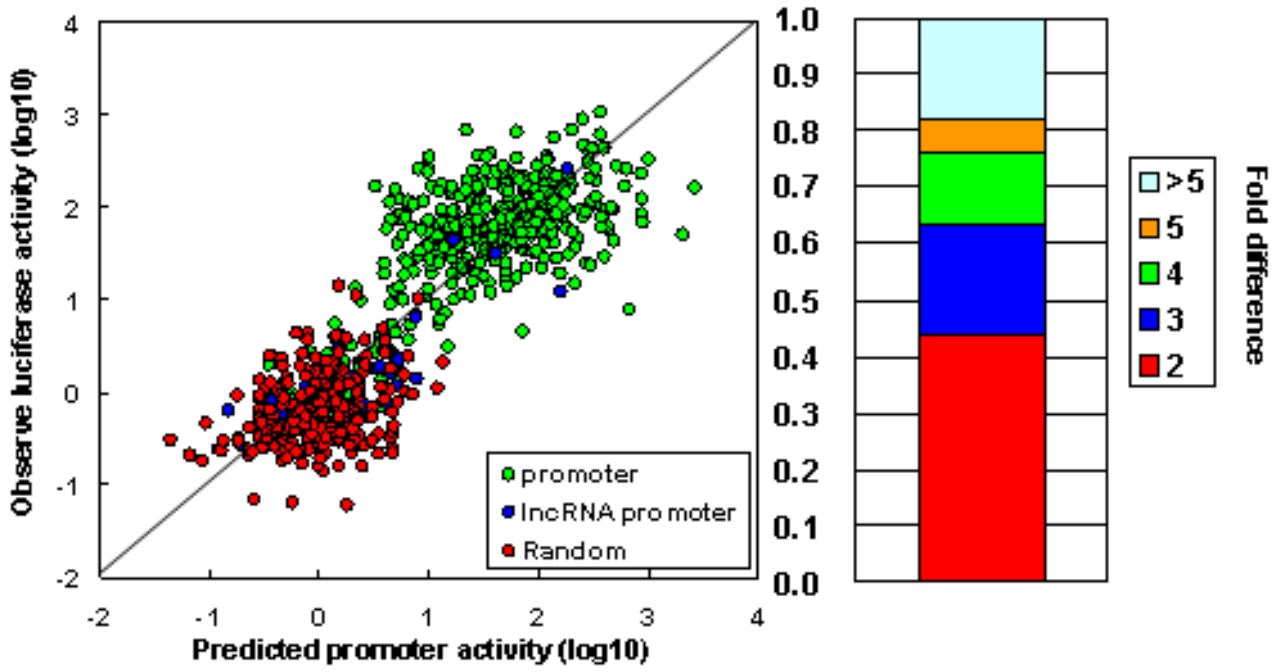


図3-4　改善後のプロモーター活性予測モデルの精度

(A)予測値と実測値との散布図，x軸はプロモーター活性予測値(log10)，y軸はルシフェラーゼ活性(log10)，(B)予測値が実測値の何倍の範囲内で予測されたクローンの割合．

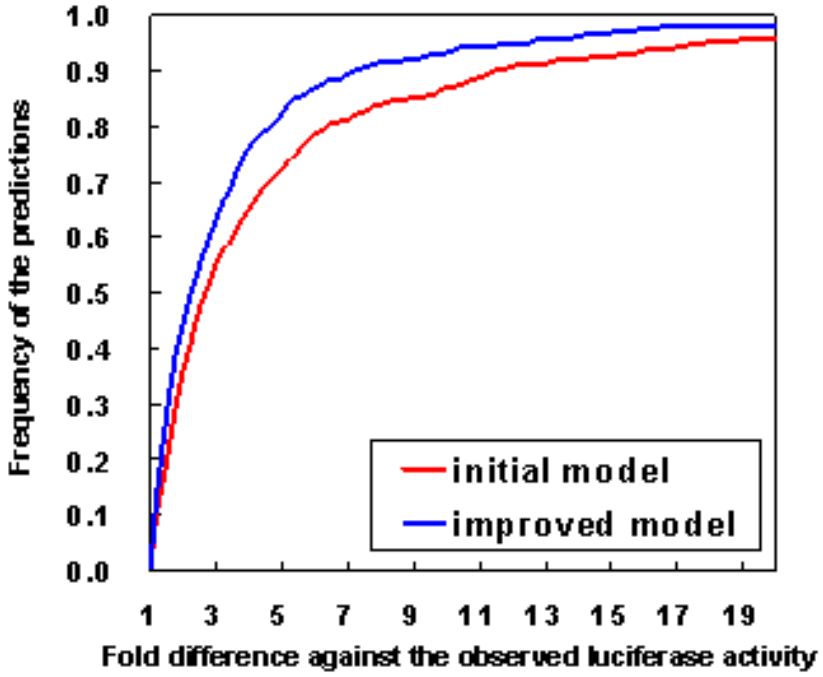


図 3-5 プロモーター活性予測モデルの精度

モデルの予測値の予測範囲の累積割合曲線. x 軸は予測された範囲, y 軸は割合を示す. 赤線, 青線はそれぞれスタートのモデル, 改善後のモデルを示す.

表3-4 プロモーター活性への寄与が大きいと考えられるTFBSの例

TF ID	TFBS ID	Assigned activity	p value
Ets1(p54)	V$CETS1P54_02,V$CETS1P54_03	0.27	<2e-16
ZF5	V$ZF5_B	0.22	1E-12
Myb	V$VMYB_02	0.17	2E-11
CREB	V$CREB_02,V$CREB_Q4_01	0.34	3E-11
Sp1	V$SP1_Q2_01	0.30	1E-10
ETF	V$ETF_Q6	0.21	1E-06

Assinged activityは重回帰分析によって得られた係数値を表す.

表3-5 プロモーター活性予測モデルに用いた85種類のTFBSと最適なパラメーター

maximum matrix score；最適なTFBSのスコアの最大値をとるTRANSFACスコア，region；最適なTFBSの位置の閾値，assigned score；重回帰分析から得られた係数値

TRANSFAC_ID	name	maximum matrix score	region	assigned score	p value
V$AHR_Q5	AHR	1	500	−0.67	1.6E−02
V$AHRARNT_01	AHR:ARNT	0	300	0.51	9.1E−02
V$AP2_Q6_01,V$AP2_Q6	AP2	0	400	0.13	4.8E−03
V$AP4_01	AP4	0	300	0.08	1.3E−01
V$BACH2_01	BACH2	0.3	1200	0.44	1.8E−01
V$BCL6_Q3	BCL6	1	500	−0.16	2.7E−03
V$CART1_01	CART1	0.3	1200	−0.14	1.7E−01
V$CDP_02	CDP	0	300	−0.26	2.2E−02
V$CDXA_02	CDXA	0	300	−0.14	4.8E−02
V$CEBPA_01	CEBPA	0.6	200	−0.58	1.4E−02
V$CEBPDELTA_Q6	CEBPD	0.7	1200	0.78	1.3E−04
V$CEBPGAMMA_Q6	CEBPG	0.7	500	−0.54	1.2E−03
V$CETS1P54_02,V$CETS1P54_03	ETS1	0	400	0.27	<2e−16
V$CHOP_01	CHOP	0.7	700	−0.30	1.8E−03
V$CKROX_Q2	CKROX	0	1000	−0.16	9.9E−02
V$CP2_02	CP2	0	1200	−0.06	1.1E−01
V$CREB_02,V$CREB_Q4_01	CREB	0	900	0.34	3.2E−11
V$DEAF1_02	DEAF1	0.5	700	0.62	7.8E−02
V$DR3_Q4	VDR, CAR, PXR	0.1	900	−0.13	1.2E−02
V$E2F_03,V$E2F_Q6_01	E2F	0.6	1200	0.23	6.6E−02
V$EBOX_Q6_01	E box	0.6	100	0.55	1.7E−01
V$EGR1_01	EGR1	0.3	300	0.29	1.5E−01
V$ER_Q6	ER	0.6	1000	0.54	5.3E−03
V$ETF_Q6	ETF	0	400	0.21	1.4E−06
V$ETS_Q6	ETS	0	300	−0.23	3.7E−02
V$FAC1_01	FAC1	1	700	−0.55	1.8E−04
V$FOXJ2_02	FOXJ2	0.1	200	−0.21	1.2E−01
V$GABP_B	GABP	0.7	200	0.27	2.4E−02
V$GATA_C	GATA	1	300	−0.53	8.6E−02
V$GEN_INI3_B	general initiator	0	200	−0.11	5.5E−02
V$GLI_Q2	GLI	0.4	1200	−0.71	3.9E−03
V$HIC1_02	HIC1	0.6	300	0.60	4.9E−02
V$HLF_01	HLF	0.1	600	0.57	5.6E−02
V$HMGIY_Q6	HMGIY	0	1000	0.22	1.2E−04
V$HNF1_Q6	HNF1	0	1000	0.23	5.3E−02
V$HNF4_Q6_01,V$HNF4ALPHA_Q6	HNF4	0.7	900	0.28	5.8E−02
V$HOX13_01	HOX1.3	0.1	1200	0.05	1.6E−01
V$HSF1_01,V$HSF1_Q6	HSF1	0.1	700	0.37	8.1E−03
V$ISRE_01	ISRE	0	1200	0.41	9.5E−02
V$KID3_01	KID3	0	300	0.06	3.6E−04
V$LEF1TCF1_Q4	LEF1TCF1	0	900	0.14	1.2E−02
V$LHX3_01	LHX3	0	1200	−0.72	3.7E−03
V$LRF_Q2	LRF	0	300	0.15	1.7E−01
V$MAZ_Q6	MAZ	0	1200	0.13	6.7E−02
V$MEF2_03	MEF2	0.3	700	0.29	6.6E−02
V$MEIS1_01	MEIS1	0	900	−0.49	4.0E−02
V$MEIS1BHOXA9_02	MEIS1B:HOXA9	0	100	−0.67	1.0E−02
V$MINI19_B	Muscle initiator sequences−19	0.2	1000	0.17	7.1E−02
V$MRF2_01	MRF2	0.3	400	−0.45	4.6E−02
V$MYOD_Q6_01	MYOD	0.7	600	−0.59	9.9E−02
V$MYOGNF1_01	myogenin:NF1	0.1	1200	−0.15	1.4E−02
V$NF1_Q6_01	NF1	0.3	100	−0.36	1.8E−01
V$NFAT_Q4_01	NFAT	0	900	−0.23	1.1E−02
V$NFKB_Q6_01	NFKB	0.7	900	0.48	8.4E−03
VNFY_01,VNFY_Q6_01	NFY	0	1200	0.18	6.0E−02
V$NKX25_02,V$NKX25_Q5	NKX2−5	0.9	300	−0.29	1.0E−03
V$OCT_Q6	OCT	1	600	0.29	2.3E−02
V$OCT1_02,V$OCT1_03,V$OCT1_Q5_01	OCT1	0.5	200	−0.36	7.7E−03
V$PAX2_01,V$PAX2_02	PAX2	0.5	400	0.09	6.6E−02
V$PAX3_B	PAX3	0	800	0.15	1.5E−04
V$PAX4_01,V$PAX4_02,V$PAX4_03	PAX4	1	1100	0.10	1.8E−04
V$PAX5_01,V$PAX5_02	BSAP	0	500	0.10	5.9E−04
V$PAX6_01,V$PAX6_Q2	PAX6	0	1200	0.08	1.4E−04
V$PAX8_01	PAX8	0	100	0.14	8.6E−02
V$PEBP_Q6	PEBP	0	800	−0.23	1.7E−01
V$POU1F1_Q6	POU1F1	0	900	−0.31	1.0E−03
V$POU6F1_01	POU6F1	0.1	900	0.35	1.0E−01
V$PPARA_01,V$PPARA_02	PPARA	0.1	700	−0.11	4.4E−04
V$RBPJK_Q4	RBPJK	0	1200	−0.27	9.4E−03
V$RFX_Q6	RFX	0.7	100	0.30	9.1E−02
V$SMAD3_Q6	SMAD3	0.4	1200	−0.18	3.7E−02
V$SOX9_B1	SOX9	0.4	300	−0.52	2.7E−03
V$SP1_Q2_01	SP1	0	700	0.30	1.2E−10
V$SRY_02	SRY	0	1200	0.12	4.1E−05
V$STAF_02	STAF	0.7	1100	0.43	1.4E−01
V$STAT_Q6	STAT	0.7	1000	−0.14	3.9E−02
V$TCF11_01	TCF11	1	200	−0.32	1.7E−01
V$TEL2_Q6	TEL2	0.1	200	0.28	6.5E−03
V$TFE_Q6	TFE	0	600	0.51	1.1E−03
V$VDR_Q3	VDR	0	100	0.28	6.0E−02
V$VMYB_02	VMYB	0	1200	0.17	1.9E−11
V$WT1_Q6	WT1	0	500	0.08	6.2E−02
V$XVENT1_01	VENTX	0.6	300	−0.26	2.9E−02
V$YY1_Q6,V$YY1_Q6_02	YY1	0.9	200	0.27	6.9E−04
V$ZF5_B	ZF5	0	300	0.22	1.2E−12

3.5　プロモーター活性予測モデルの評価

構築した予測モデルの評価を行った. 以下の3種類の方法を用いた評価を行った. (1)10分割交差検定 (2)ランダム生成した配列との比較 (3)ランダム生成したマトリックスとの比較 (4)TFBSの欠失配列のLuc活性の測定 (5)IRESを用いたベクター について評価を行った.

3.5.1　10-分割交差検定による評価

得られたプロモーター活性予測モデルに対し, モデルの未知データに対する有効性を検証する目的として10分割交差検定法による評価を行った. 10分割交差検定法はランダムに選択した90%のデータを訓練用のデータとしてモデルを構築し, 残りの10%の試験用のデータとして評価に用いる方法である. 交差検定では試験用のデータをモデルの構築には利用しないため, 未知データに対する予測精度を調べることができる. また多変数のモデルの構築の際の過学習の評価にも用いられる. 今回は, この10分割交差検定をランダムに1000回繰り返し行い, 実験値と予測値の相関係数を計算した. 167種のTFBSを用いた場合は, 相関係数の平均値はr=0.79であった. またAICにより選ばれた85種のTFBSを用いた場合はr=0.83を得た(図3-6). 全データを用いた相関係数との差も小さいため, 構築したプロモーター活性予測モデルは過剰な変数による過学習による影響も小さく, 未知のDNA断片に対しても予測に用いることができることを意味する.

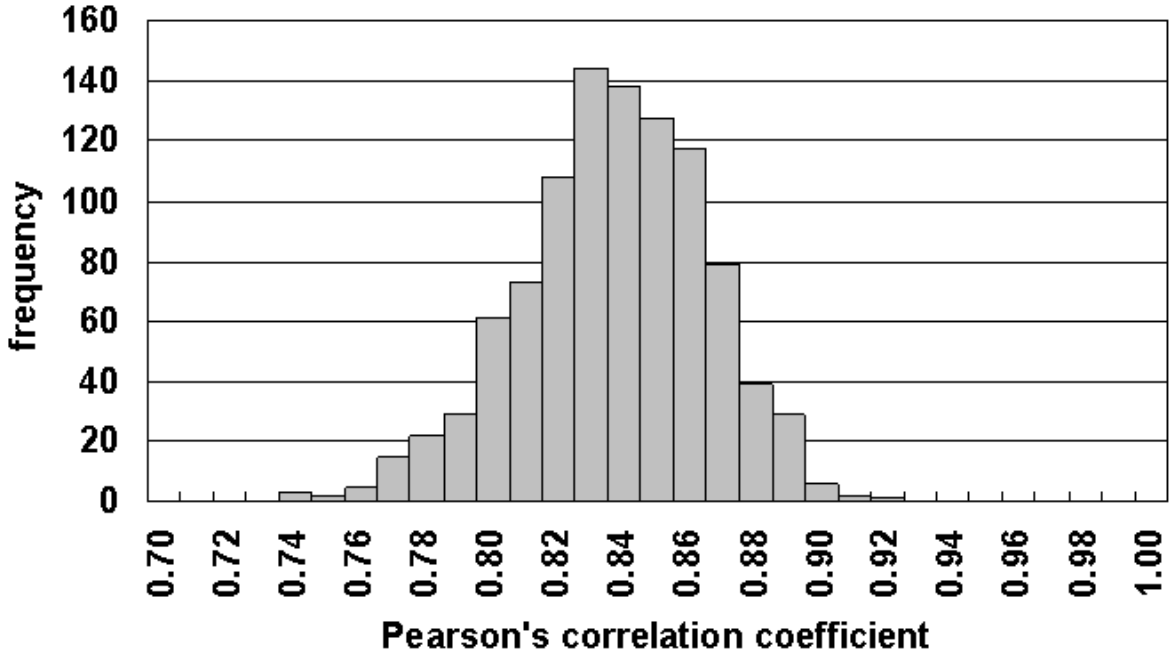


図3-6 10-分割交差検定による相関係数

ランダムに選んだ10%を試験用，残りの90%を訓練用のデータセットとし，試験用のデータの実験値と予測値の相関係数(Peason’s correlation coefficient)を計算した．この操作を1000回繰り返した．

3.5.2 TFBS 欠失させたプロモーター配列の転写活性化能の測定

重回帰分析により，TFBSを説明変数としてそれぞれの係数の値の算出を行った（表3-4）．算出された説明変数のスコアはTFBSの転写への寄与へのスコアであると期待される．そこで予測されたTFBSの係数の値の評価を行うため，実験的な評価を行った．TFBSを欠失させたプロモータークローンを作成し，ルシフェラーゼ活性の測定を行い，オリジナルのDNA断片と欠失させたDNA断片の比較を行った．実験には24種類のTFBS，61の変異DNA断片を作成し，ルシフェラーゼ活性を測定した．選択の基準はTFBSのスコアが正に大きいものを例とした．変異DNA断片のうち27(44%)プロモーターでルシフェラーゼ活性が減少した($p < 0.05$ t test)．一部の例を図3-7で示す．また全データは表Xに示した．これらの結果は，予測モデルに用いられたTFBS中の半分程度のTFBSについてはHEK293細胞中において実際に活性化に寄与していることを意味している．

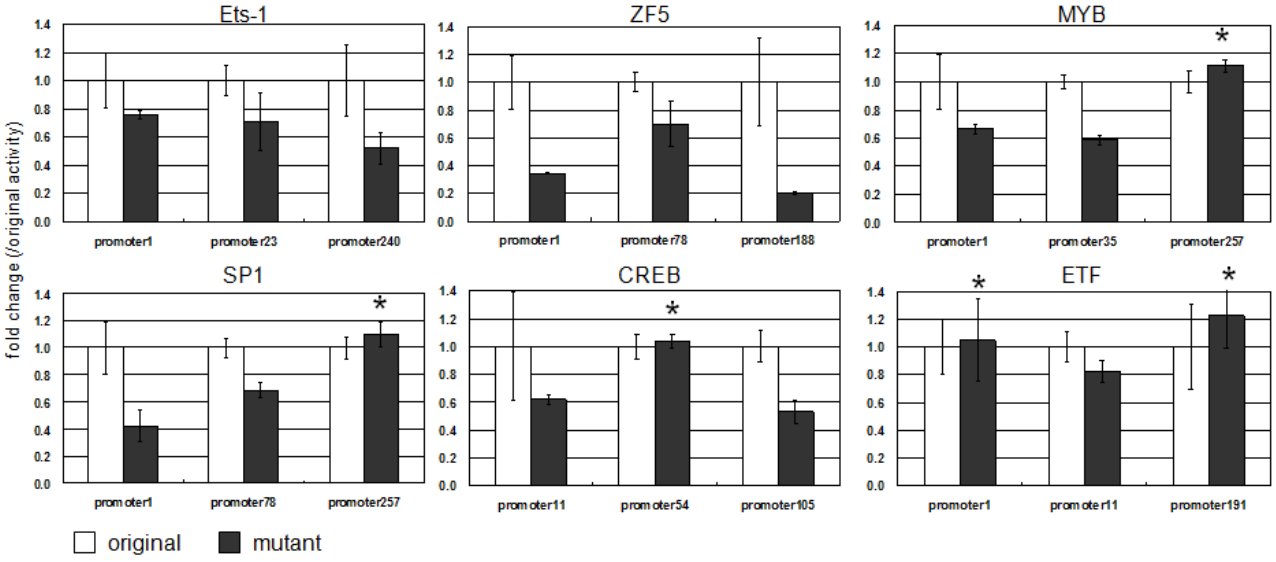


図3-7 TFBSの欠失配列の評価

TFBSを欠失させたプロモーター配列を用いて転写活性化能の測定を行った. オリジナルのプロモーター配列のルシフェラーゼ活性(白), TFBSを欠失したプロモーター配列のルシフェラーゼ活性(黒). 確認できなかったクローンを＊で表した. エラーバーは3回の実験の標準偏差を表す.

表3-6 TFBS欠失配列のプロモーター活性の測定結果

緑がオリジナルの配列と比べ活性が有意($p < 0.05$)に減少した例を示す.

clone name	TF name	TRANSFAC ID	position	fold change	s.d.	p value
promoter1	AHR:ARNT	V$AHRARNT_01	-47,-33	0.79	0.05	7.3E-02
promoter11	C/EBP delta	V$CEBPDELTA_Q6	-268,-263	0.62	0.04	1.4E-02
promoter140	C/EBP delta	V$CEBPDELTA_Q6	-406,-395	0.85	0.16	1.6E-01
promoter188	C/EBP delta	V$CEBPDELTA_Q6	-906,-895	0.36	0.12	1.5E-02
promoter39	C/EBP delta	V$CEBPDELTA_Q6	-44,-33	1.20	0.09	2.7E-02
promoter1	ETS1	V$CETS1P54_03	-26,-11	0.76	0.03	4.9E-02
promoter23	ETS1	V$CETS1P54_03	-353,-338	0.71	0.20	2.0E-04
promoter240	ETS1	V$CETS1P54_03	-317,-301	0.52	0.11	2.0E-02
promoter105	CREB	V$CREB_02	-223,-212	0.53	0.08	2.3E-03
promoter11	CREB	V$CREB_02	-487,-476	0.74	0.02	4.3E-02
promoter230	CREB	V$CREB_02	-556,-544	0.85	0.25	1.9E-01
promoter54	CREB	V$CREB_02	-164,-153	1.04	0.05	2.9E-01
promoter257	DEAF1	V$DEAF1_02	-196,-172	1.25	0.03	2.8E-04
promoter53	DEAF1	V$DEAF1_02	-101,-77	1.38	0.24	3.4E-02
promoter41	E2F	V$E2F_03	-95,-84	1.59	0.48	6.3E-02
promoter23	E2F	V$E2F_Q6_01	-148,-137	1.18	0.07	1.3E-03
promoter153	EBOX	V$EBOX_Q6_01	-126,-117	1.54	0.17	4.8E-03
promoter294	EBOX	V$EBOX_Q6_01	-63,-54	0.95	0.08	2.4E-01
promoter1	ETF	V$ETF_Q6	-298,-292	1.05	0.30	4.1E-01
promoter11	ETF	V$ETF_Q6	-83,-77	0.61	0.06	1.4E-02
promoter191	ETF	V$ETF_Q6	-277,-271	1.23	0.23	1.8E-01
promoter23	ETF	V$ETF_Q6	-414,-408	1.19	0.10	2.1E-03
promoter11	GABP	V$GABP_B	-260,-249	0.59	0.13	3.9E-02
promoter150	GABP	V$GABP_B	-195,-184	0.01	0.01	8.6E-04
promoter124	HLF	V$HLF_01	-391,-382	1.14	0.51	3.4E-01
promoter229	HLF	V$HLF_01	-397,-388	0.67	0.22	1.4E-01
promoter274	HLF	V$HLF_01	-806,-797	0.93	0.06	1.2E-01
promoter101	HNF4	V$HNF4_Q6_01	-347,-338	0.82	0.03	1.7E-02
promoter34	HNF4	V$HNF4_Q6_01	-559,-546	0.64	0.04	3.4E-02
promoter124	HNF4	V$HNF4ALPHA_Q6	-723,-711	0.98	0.45	4.8E-01
promoter64	HSF1	V$HSF1_01	-669,-663	1.29	0.07	8.9E-03
promoter11	ISRE	V$ISRE_01	-464,-450	0.77	0.06	5.6E-02
promoter353	ISRE	V$ISRE_01	-99.-85	1.22	0.23	2.0E-01
promoter373	ISRE	V$ISRE_01	-329,-315	0.75	0.16	9.4E-02
promoter134	MEF2	V$MEF2_03	-454,-433	0.77	0.07	1.5E-02
promoter164	MEF2	V$MEF2_03	-506,-485	1.06	0.17	3.4E-01
promoter227	MEF2	V$MEF2_03	-102,-81	1.04	0.11	3.2E-01
promoter23	OCT	V$OCT_Q6	-525,-515	1.02	0.11	3.5E-01
promoter57	OCT	V$OCT_Q6	-555,-545	1.03	0.15	4.2E-01
promoter13	PAX5	V$PAX5_01	-480,-453	0.35	0.03	1.6E-06
promoter1	SP1	V$SP1_Q2_01	-255,-249	0.42	0.11	5.7E-03
promoter257	SP1	V$SP1_Q2_01	-268,-254	1.10	0.09	2.3E-02
promoter78	SP1	V$SP1_Q2_01	-231,-222	0.69	0.06	1.9E-03
promoter1	SRY	V$SRY_02	-328,-317	0.73	0.08	4.5E-02
promoter35	SRY	V$SRY_02	-860,-849	1.19	0.04	9.9E-04
promoter91	SRY	V$SRY_02	-115,-104	0.93	0.06	1.7E-01
promoter108	TEL2	V$TEL2_Q6	-166,-157	0.52	0.15	4.8E-04
promoter191	TFE	V$TFE_Q6	-858,-851	1.22	0.43	2.5E-01
promoter44	TFE	V$TFE_Q6	-451,-444	3.18	0.33	1.9E-04
promoter1	MYB	V$VMYB_02	-452,-446	0.66	0.03	1.5E-01
promoter257	MYB	V$VMYB_02	-587,-579	0.59	0.04	1.4E-04
promoter35	MYB	V$VMYB_02	-585,-577	1.11	0.04	6.2E-03
promoter23	WT1	V$WT1_Q6	-336,-328	0.82	0.10	1.1E-04
promoter304	WT1	V$WT1_Q6	-337,-329	1.04	0.16	3.6E-01
promoter91	WT1	V$WT1_Q6	-283,-275	0.99	0.24	4.6E-01
promoter108	YY1	V$YY1_Q6	-107,-99	0.53	0.09	2.3E-04
promoter150	YY1	V$YY1_Q6	-113,-105	0.18	0.09	2.2E-03
promoter16	YY1	V$YY1_Q6	-97,-89	0.92	0.09	2.0E-01
promoter1	ZF5	V$ZF5_B	-245,-236	0.35	0.01	2.2E-03
promoter188	ZF5	V$ZF5_B	-181,-169	0.21	0.01	5.7E-03
promoter78	ZF5	V$ZF5_B	-277,-265	0.70	0.16	2.0E-03

3.5.3　ランダム生成したDNA断片の比較

プロモーター活性予測モデルの検証を目的として，プロモーター配列の特異性の検証を行った．Luc活性の測定に用いたプロモータークローン，ランダムクローンのDNA配列と同じGC含量を持つように，コンピュータ上でランダムDNA配列を生成し，その配列のプロモーター活性の予測値の分布を得た（図3-8A）．予測値の分布の比較を行ったところ，プロモーターの配列は同様のGC含量を持つランダムなDNA配列のプロモーター活性の予測値と比較し有意に高い予測値を得ることがわかった（p value <10^{-100};Wilcox test）．

ヒトのプロモーター配列に対する特異性の検証を行った．RefSeq遺伝子5'端領域においてHEK293細胞中で転写開始点が見られた領域のプロモーター活性の予測値の計算を行った（詳細は3.5.1を参照）．ヒト以外の下等真核生物（酵母，線虫，ショウジョウバエ）のプロモーター配列のプロモーター活性予測値の算出をした．それぞれのプロモーター活性予測値の分布を比較した結果，ヒト以外の真核生物のプロモーターと比べヒトプロモーター配列が有意に高いプロモーター活性予測値を得ることがわかった（酵母，線虫，ショウジョウバエにおいてp value <10^{-100};Wilcox test）．したがって，プロモーター活性予測モデルはヒトのプロモーター配列に特異的に高いスコアを与えるモデルであった．

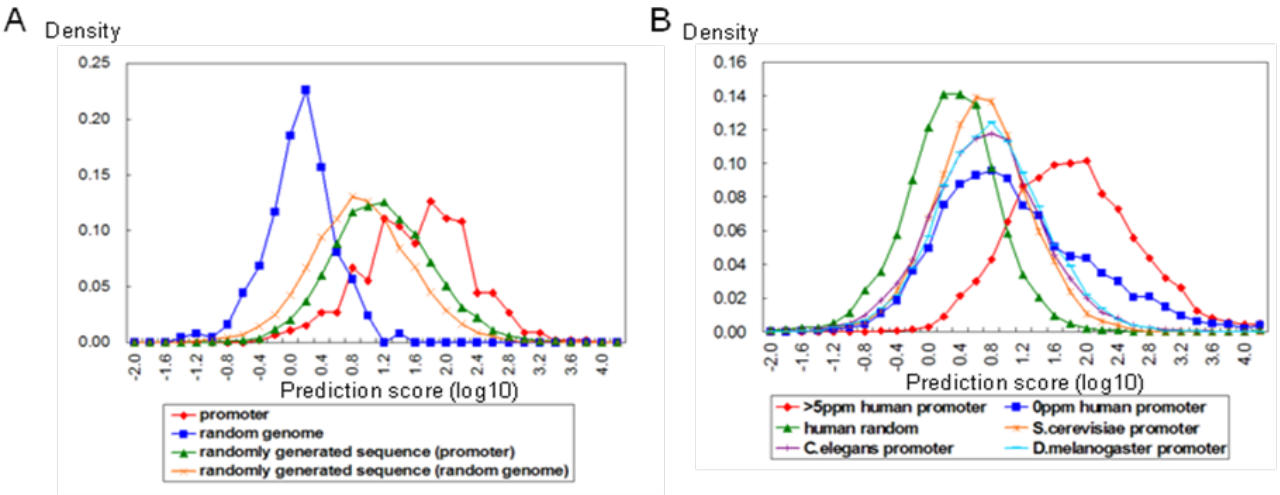


図3-8　プロモーター予測モデルのDNA配列特異性の評価

(A)　ルシフェラーゼアッセイに用いたDNA断片と同じGC含量を持つランダム生成DNA断片のプロモーター活性予測値の分布．ルシフェラーゼアッセイに用いたプロモーター領域の予測値（赤），ルシフェラーゼアッセイに用いたランダムなゲノム領域の予測値（青），プロモータークローンと同様のGC含量を持つランダムDNA配列の予測値（緑），ランダムなゲノム領域と同様のGC含量を持つランダムDNA配列の予測値（黄）の分布を示す．

(B)　各生物由来のプロモーター配列のプルモーター活性の予測値の分布．ヒトRefSeq遺伝子

5’領域（>5ppm TSS）の予測値（赤），RefSeq遺伝子5’端領域（0 ppm TSS）（青），ランダム抽出したヒトゲノム領域（緑），*S.cerevisiae*のプロモーター領域（黄），*C.elegans*のプロモーター領域（紫），*D.melanogaster*のプロモーター領域（淡青）の予測値の分布を示す．ゲノム配列，遺伝子情報はUCSC genome browserを利用した．

3.5.4　ランダム生成したマトリックスを用いた評価

TFBSの検索に用いたTRANSFACのPWMに対して塩基配列の特異性についての検証を行った．方法はTRANSFACのPWMの塩基の位置をランダムにシャッフルしたランダムPWMの作成を行った．ランダムPWMを用いて配列のサーチを行い，プロモーター活性予測モデルと同様に重回帰分析を行い，係数の値の算出を行った．ランダムPWMを用いた時の係数値の分布とオリジナルPWMの係数値を比較した．その結果TFBSごとに配列特異性があるもの配列の特異性が弱いものがあった．配列特異性のある例については表3-7の緑で示した．プロモーター活性への説明の寄与の大きかった例では，Ets1，ZF5，MYB，CREBなどのTFBSがオリジナルのPWMにおいて有意に高い係数値を与えため，PWMの配列特異性があるといえる．ETF，SP1についてはオリジナルで高い計数値を与えたが，ランダムのPWMの分布との有意差が見られなかった．これらの例については高GCでcomplexityが低いマトリックスであり，ランダムマトリックスであっても高GCな領域で頻繁に見つかること，またランダムPWMのバリエーションが少ないため，ランダムとオリジナルで差がなかったと考えられる．

表3-7 TRANSFACのPWMの塩基の位置をランダムにシャッフルしたPWMを利用した時の重回帰分析の係数値

original matrix:オリジナルのPWMを用いた係数値, ave. random matrix:ランダムシャッフルしたPWMを用いた係数値の平均, s.d.: ランダムシャッフルしたPWMを用いた係数値の標準偏差, *P*:ランダムシャッフルした係数値が正規分布に従うと仮定したときのオリジナルの係数値をとる確率. P<0.05となるTFを緑色で示した.

TF name	TRANSFAC_ID	original matrix	ave. random matrix	s.d.	*P*
AHR	V$AHR_Q5	-0.67	0.11	0.47	4.8E-02
AHR:ARNT	V$AHRARNT_01	0.51	0.12	0.32	1.1E-01
AP2	V$AP2_Q6_01,V$AP2_Q6	0.13	0.08	0.05	1.2E-01
AP4	V$AP4_01	0.08	0.10	0.11	4.2E-01
BACH2	V$BACH2_01	0.44	0.12	0.55	2.8E-01
BCL6	V$BCL6_Q3	-0.16	0.01	0.13	8.6E-02
CART1	V$CART1_01	-0.14	-0.02	0.08	7.3E-02
CDP	V$CDP_02	-0.26	-0.09	0.12	8.4E-02
CDXA	V$CDXA_02	-0.14	-0.15	0.14	4.8E-01
CEBP/A	V$CEBPA_01	-0.58	0.00	1.29	3.3E-01
CEBP/D	V$CEBPDELTA_Q6	0.78	-0.02	0.28	2.4E-03
CEBP/G	V$CEBPGAMMA_Q6	-0.54	-0.17	0.31	1.1E-01
ETS1	V$CETS1P54_02,V$CETS1P54_03	0.27	0.08	0.06	7.3E-04
CHOP	V$CHOP_01	-0.30	0.00	0.18	4.9E-02
KROX	V$CKROX_Q2	-0.16	0.13	0.16	3.4E-02
CP2	V$CP2_02	-0.06	0.07	0.07	2.4E-02
CREB	V$CREB_02,V$CREB_Q4_01	0.34	0.05	0.07	4.3E-05
DEAF1	V$DEAF1_02	0.62	0.02	0.90	2.5E-01
DR3	V$DR3_Q4	-0.13	0.04	0.07	5.7E-03
E2F	V$E2F_03,V$E2F_Q6_01	0.23	0.08	0.09	5.7E-02
EBOX	V$EBOX_Q6_01	0.55	0.13	0.64	2.5E-01
EGR1	V$EGR1_01	0.29	0.11	0.27	2.5E-01
ER	V$ER_Q6	0.54	0.08	0.37	1.1E-01
ETF	V$ETF_Q6	0.21	0.07	0.10	8.3E-02
ETS	V$ETS_Q6	-0.23	0.05	0.27	1.5E-01
FAC1	V$FAC1_01	-0.55	-0.03	0.15	3.2E-04
FOXJ2	V$FOXJ2_02	-0.21	-0.21	0.24	5.0E-01
GABP	V$GABP_B	0.27	0.25	0.45	4.8E-01
GATA	V$GATA_C	-0.53	-0.11	0.49	2.0E-01
Initiator	V$GEN_INI3_B	-0.11	0.02	0.10	1.1E-01
GLI	V$GLI_Q2	-0.71	0.11	0.52	5.6E-02
HFH1	V$HFH1_01	0.60	0.20	0.55	2.3E-01
HLF	V$HLF_01	0.57	-0.01	0.26	1.4E-02
HMGIY	V$HMGIY_Q6	0.22	-0.02	0.09	3.1E-03
HNF1	V$HNF1_Q6	0.23	-0.02	0.11	1.1E-02
HNF4	V$HNF4_Q6_01,V$HNF4ALPHA_Q6	0.28	0.06	0.23	1.7E-01
HOX13	V$HOX13_01	0.05	-0.01	0.03	4.4E-02
HSF1	V$HSF1_01,V$HSF1_Q6	0.37	0.03	0.25	8.4E-02
ISRE	V$ISRE_01	0.41	0.11	0.50	2.7E-01
KID3	V$KID3_01	0.06	0.05	0.02	3.7E-01
LEF1:TCF1	V$LEF1TCF1_Q4	0.14	0.02	0.09	9.0E-02
LHX3	V$LHX3_01	-0.72	0.04	0.25	1.1E-03
LRF	V$LRF_Q2	0.15	0.08	0.11	2.7E-01

TF name	TRANSFAC_ID	original matrix	ave. random matrix	s.d.	*P*
MAZ	V$MAZ_Q6	0.13	0.16	0.11	4.0E-01
MEF2	V$MEF2_03	0.29	-0.01	0.32	1.7E-01
MEIS1	V$MEIS1_01	-0.49	0.11	0.43	8.2E-02
MEIS1B:HOXA9	V$MEIS1BHOXA9_02	-0.67	0.18	0.51	4.8E-02
Initiator	V$MINI19_B	0.17	0.11	0.13	3.2E-01
MRF2	V$MRF2_01	-0.45	-0.19	0.49	2.9E-01
MYOD	V$MYOD_Q6_01	-0.59	0.13	0.91	2.1E-01
MYOGNF1	V$MYOGNF1_01	-0.15	0.05	0.08	7.8E-03
NF1	V$NF1_Q6_01	-0.36	0.08	1.22	3.6E-01
NFAT	V$NFAT_Q4_01	-0.23	-0.02	0.19	1.4E-01
NFKB	V$NFKB_Q6_01	0.48	0.13	0.62	2.9E-01
NFY	VNFY_01,VNFY_Q6_01	0.18	0.07	0.36	3.8E-01
NKX25	V$NKX25_02,V$NKX25_Q5	-0.29	-0.10	0.11	3.6E-02
OCT	V$OCT_Q6	0.29	-0.01	1.76	4.3E-01
OCT1	V$OCT1_02,V$OCT1_03,V$OCT1_Q5_01	-0.36	-0.13	0.21	1.4E-01
PAX2	V$PAX2_01,V$PAX2_02	0.09	0.01	0.09	1.8E-01
PAX3	V$PAX3_B	0.15	0.05	0.04	1.2E-02
PAX4	V$PAX4_01,V$PAX4_02,V$PAX4_03	0.10	0.09	0.03	3.9E-01
PAX5	V$PAX5_01,V$PAX5_02	0.10	0.08	0.04	2.9E-01
PAX6	V$PAX6_01,V$PAX6_Q2	0.08	0.01	0.02	6.0E-04
PAX8	V$PAX8_01	0.14	0.01	0.40	3.7E-01
PEBP	V$PEBP_Q6	-0.23	0.01	0.21	1.3E-01
POU1F1	V$POU1F1_Q6	-0.31	-0.04	0.15	3.2E-02
POU6F1	V$POU6F1_01	0.35	0.03	0.52	2.7E-01
PPARA	V$PPARA_01,V$PPARA_02	-0.11	0.02	0.05	4.7E-03
RBPJK	V$RBPJK_Q4	-0.27	0.06	0.20	4.5E-02
RFX	V$RFX_Q6	0.30	0.06	0.73	3.8E-01
SMAD3	V$SMAD3_Q6	-0.18	0.08	0.15	4.8E-02
SOX9	V$SOX9_B1	-0.52	-0.21	1.63	4.2E-01
SP1	V$SP1_Q2_01	0.30	0.25	0.14	3.8E-01
SRY	V$SRY_02	0.12	0.03	0.03	2.1E-03
STAF	V$STAF_02	0.43	0.47	2.54	4.9E-01
STAT	V$STAT_Q6	-0.14	0.03	0.19	1.8E-01
TCF11	V$TCF11_01	-0.32	0.03	0.59	2.8E-01
TEL2	V$TEL2_Q6	0.28	0.02	0.42	2.7E-01
TFE	V$TFE_Q6	0.51	0.08	0.36	1.1E-01
VDR	V$VDR_Q3	0.28	0.06	0.38	2.9E-01
MYB	V$VMYB_02	0.17	0.06	0.05	1.2E-02
WT1	V$WT1_Q6	0.08	0.10	0.06	4.0E-01
XVENT1	V$XVENT1_01	-0.26	-0.05	0.19	1.3E-01
YY1	V$YY1_Q6,V$YY1_Q6_02	0.27	0.06	0.53	3.5E-01
ZF5	V$ZF5_B	0.22	0.13	0.05	4.7E-02

3.5.5 IRESを用いた翻訳への影響の評価

プロモータークローンごとの翻訳効率の影響について評価を行った. 実験にはプロモーター配列のDNA断片をルシフェラーゼ遺伝子上流に組み込んだベクターを用いている. 発現する遺伝子は各クローンとも同一のルシフェラーゼ遺伝子であるため, ルシフェラーゼ活性の違いはDNA配列の特徴として捉えることができる. 実際には, クローニングした領域には転写開始点以降の領域も含まれているため, 発現するルシフェラーゼ遺伝子にはそれぞれのプロモーター領域由来の5'UTRが含まれていると考えられる. 5'UTRが下流のルシフェラーゼ遺伝子の翻訳の効率に影響を与える可能性があるため[58], 各クローンの5'UTRの翻訳に及ぼす影響の評価を行った.

その目的にルシフェラーゼ遺伝子上流にinternal ribosome entry site (IRES)配列を組み込んだベクターを用いた. IRESは高次構造をとりキャップ構造非依存的に翻訳をすることができるため[59], 各クローン間の翻訳の効率を揃えることができると期待した. DNA断片をルシフェラ

ーゼ遺伝子上流にIRES配列を組み込んだベクターへ乗せ替えを行い，704種類のDNA断片について，同様の方法でLuc活性の測定を行った．通常のプロモータークローンとIRESベクタークローンの比較を行った結果，相関係数は0.93であった(図3-9)．この結果から5'UTR配列の違いに由来する翻訳のバイアスは，今回のモデルにおいては無視できる程度に小さいと結論した．

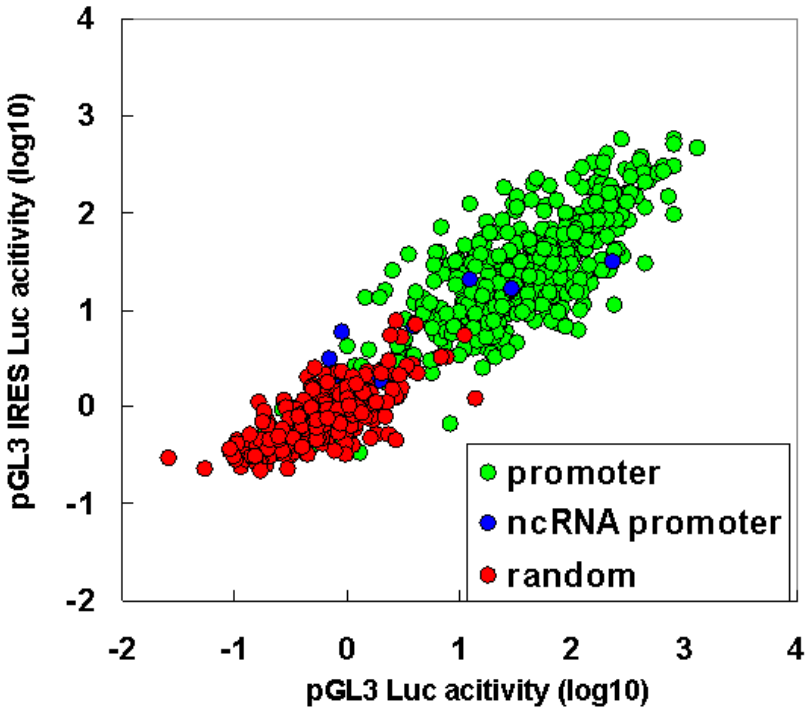


図3-9　IRESを用いたベクター系による翻訳の影響の評価
通常のルシフェラーゼベクター（*x*軸）ルシフェラーゼ遺伝子5'端にIRES配列を組み込んだベクター（*y*軸）

3.6　既存のアプローチとの比較

本研究による手法と既存のアプローチとの比較を行った．(1)既存のプロモーター予測プログラム，(2)Landolineらの体系的ルシフェラーゼアッセイ情報[60]との比較を行った．

3.6.1　既存のプロモーター予測との比較

先行研究にはプロモーター活性の"絶対量"を予測するモデルはほとんどないが，類似するものとしてプロモーター"領域"を予測プログラムがある．そこで本研究のプロモーター"活性"予測モデルと既存のプロモーター"領域"もしくは転写開始点予測プログラムを用いた比較を行った．既存のプロモーター領域の予測プログラムにはARTS [47]，Eponine [48]，EP3 [49]，

ProSOM [50], Promoter2.0 [51], FirstEF [52]を用いた. 既存の予測プログラムはDNA配列に対し, プロモーターらしさについてのスコアや確率値などを与える. 与えられたスコアとルシフェラーゼ活性の相関係数(Pearson's correlation coefficient)を算出した(表3-8). その結果, 既存の予測モデルはプロモーター活性の予測に用いることが出来るが, 本研究のプロモーター活性予測モデルの相関係数が高い傾向にあった.

表3-8 既存のプロモーター領域予測プログラムとの比較

	本研究	ARTS	EP3	Eponine	ProSOM	Promoter2.0	firstEF
all clone	0.83	0.79	0.40	0.37	0.60	0.11	0.75
promoter clone	0.60	0.53	0.26	0.21	0.35	0.017	0.43

数値はルシフェラーゼ活性値と予測スコアとの相関係数(Peason's correlation coeffieicent). 本研究の相関係数についてはleave-one-out cross validationによって得られた数値.

3.6.2 他の体系的ルシフェラーゼデータとの比較

2010 年に Landolin らによって 4,565 種のプロモーター配列と 8 種類の細胞を用いた体系的ルシフェラーゼアッセイデータを用いた解析について報告されている[60]. 論文では, "ユビキタス"な発現を示すプロモーターは正規化された CpG スコア(= CpG の数 × 塩基長 / (C の数 × G の数))で予測できると述べている(相関係数 r=0.75). また高 CG プロモーター(正規化 CpG スコア>0.5)と低 CG プロモーター(正規化 CpG スコア<0.5)で分類した時, 高 CpG プロモーターの方が低 CpG プロモーターと比べると相関係数が低い傾向にあった(高 CpG プロモーター, 低 CpG プロモーターの相関係数はそれぞれ r=0.22, r=0.5 と報告されている). そこで本研究のプロモーター活性予測モデルにおいて同様の比較を行った. 本研究による予測モデルでは, 予測値と実測値の相関係数がクローン全体では r=0.86, プロモータークローンのみでは r=0.66 であった. またプロモータークローン内で high-CG プロモーター(正規化 CpG スコア>0.5)と low-CG プロモーター(正規化 CpG スコア<0.5)に分類したとき, それぞれの相関係数は r=0.34, r=0.77 であった. また Landoline らの Luc データを用いて, 本研究と同様の手法を用いてプロモーター活性予測モデルの構築を試みた. 表 3-9 に示すようにどの細胞においても構築可能であり, 相関係数が約 r=0.6 程度であり, 本研究で用いた HEK293 細胞と同程度の予測精度を達成した.

表3-9 Landoline et al,Genome Res, (2010) のデータを用いたプロモーター活性予測モデル

細胞種ごとの4,565プロモーターのプロモーター活性情報とDNA配列情報を用いて予測モデルを本研究の手法を用いて構築した.

cell type	ht1080	g402	t98g	hct116	hela	hepg2	ags	u87mg
correlation coefficient (*r*)	0.60	0.55	0.67	0.64	0.68	0.64	0.67	0.63

3.7 *in vivo* での予測モデルの評価

ルシフェラーゼアッセイを用いたプロモーター活性予測モデルについてはある程度可能になったと考えている. 次に, ヒトゲノム配列を用いてプロモーター活性の予測値と細胞内で起こっている転写との関連について解析を行った. 細胞内の転写の指標にTSS-seq法によるタグ数の情報を用いた. TSS-seqはオリゴキャップ法によって作成された完全長cDNAの5'端のDNA断片を"次世代型"シークエンサーと呼ばれるIllumina GAIIで配列決定を行う実験手法である. HEK293細胞中のmRNA由来のTSS-seq情報を利用した解析を行った. タグ情報の集計結果については図3-10, 表3-10にまとめた.

A

TSS tag statistics	
#TSS tags in RefSeq region	8,880,335
#TSS tags in integenic region	736,616
Total	9,734,314

Positions of the TSS tags	
# TSS tag clusters (total)	15,932
#TSS tag clusters in RefSeq regions (>5ppm)	5,605
# Represented RefSeq genes (>5ppm)	5,249
# intergenic TSS tag clusters (total)	27,959
#Intergenic TSS tag clusters (>5ppm)	833

Position of the TSS tags in RefSeq regions		
TSS tags in Refseq regions (total)		8,880,335
	upstream	2,687,663
	first exon	5,112,453
	second or later exon	707,253
	intron	372,966
	%full	0.92

B

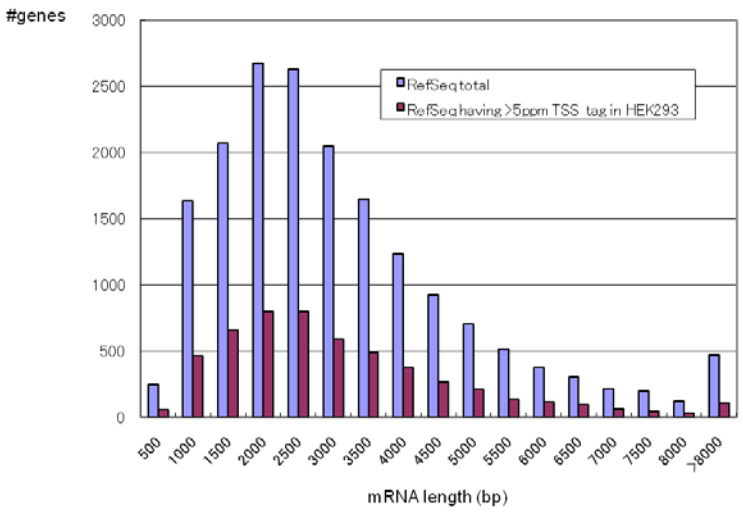


図3-10 TSS集計データ

(A) HEK293のTSSライブラリーの評価. RefSeq領域を中心に, 第一エクソン, 上流領域(50-kb), イントロン領域にマップされたTSSの割合(%full)を計算した(0.92)(右パネル). (B) TSSの位置から推定された転写産物の長さの分布.

表3-10 プロモーター活性予測モデルの評価に用いたIllumina GAの配列情報の集計結果

*in vivo*のプロモーター活性予測値の評価を目的として転写に関わる情報としてTSS-Seq, Nucleosome-Seq, ChIP-Seq（RNAポリメラーゼII）を用いた.

TSS Seq	**#total reads**	**9,734,314**
	expected accuracy to detect correct TSSs	0.9
	#total TSS Clusters of >5ppm	6,641
	#total TSS Clusters	135,579
Nucleosome Seq	**#total paired-end reads**	**15,071,279**
	median insert size	163 bp
ChIP Seq (pol II)	**#total reads**	**15,864,405**
	#WCE reads	5,774,736
	#IP reads	10,089,669
	#peak detected	43,214
	#peak in RefSeq region	37,696 (87%)
	#total of TSS Clusters of >5ppm in HEK293	6,641
	#peak overlapping >5ppm TSS Clusters in HEK293	5,499 (83%)
	#total of TSS Clusters of <5ppm in HEK293	86,704
	#peak overlapping TSS Clusters of <5ppm in HEK293	12,410 (14%)

3.7.1 TSS seq データを利用した *in vivo* での予測モデルの評価

次に本研究によって予測されたプロモーター活性予測値とヒトゲノム中の分布とmRNAレベルでの発現レベルの比較解析を行った. この解析にはHEK293細胞由来のTSS-seq法による転写開始点情報を用い, 1400万の36bpのタグ数を転写開始点及び転写量情報として用いた. 18,686種類のRefSeq遺伝子の5'の上流1kbの領域を用い, そこにマップされたTSSタグのカウントとその領域のプロモーター活性の予測値を比較した. 予測値とTSSのタグカウントによる実測値との相関は低く, 転写産物レベルの絶対量の予測は困難であった（図3-11）.

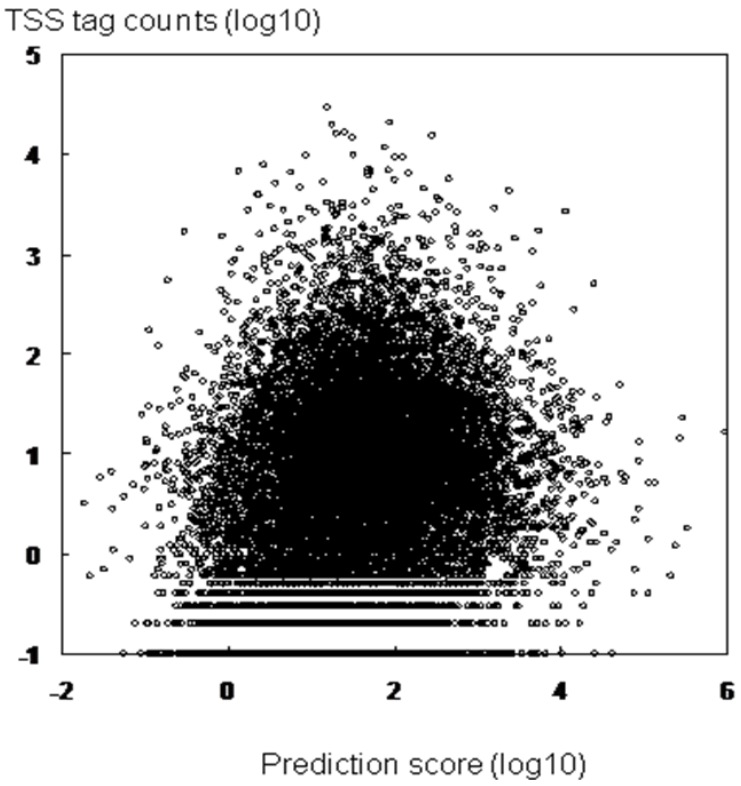


図3-11　RefSeq5'端領域のプロモーター活性予測値とTSS-seqによるタグ数との相関
RefSeq5'端領域の配列のプロモーター活性予測値(*x*軸)とその領域にマップされたTSSのタグ数($\log_{10}$)(*y*軸)

次に定性的な予測精度について評価を行った. HEK293細胞において, 転写活性があるプロモーター配列と無いプロモーター配列を判別することが可能であるかどうか検証した. 18,686種類のRefSeq遺伝子をHEK293細胞で発現確認されているものと発現が確認できないものの2群に分類した. 発現情報にはTSSのタグカウント情報を用いた. 発現量の閾値はTSS-tagのタグ数>5ppm (parts per million)とした. 粗い推定ではあるが, これは細胞中に総体で100万コピーのmRNAがあるとしたときに, 5コピーの転写産物があるということを示している[自分のREF論文28を入れる]. >5ppmのTSSがあるRefSeq5'端領域を転写活性化能のある"active"なプロモーター領域とし, 0ppmの領域を転写が起こらない"silent"なプロモーター領域とした. 0<TSS<5ppmのRefSeq5'端の5,622領域は解析には含めなかった. それぞれのプロモーター群のプロモーター活性予測値の分布を調べたところ, これらの群の分布には差があることがわかった(p<1e-100; Wilcoxon rank test)(図3-12). したがって定性的な分類は可能であると言える.

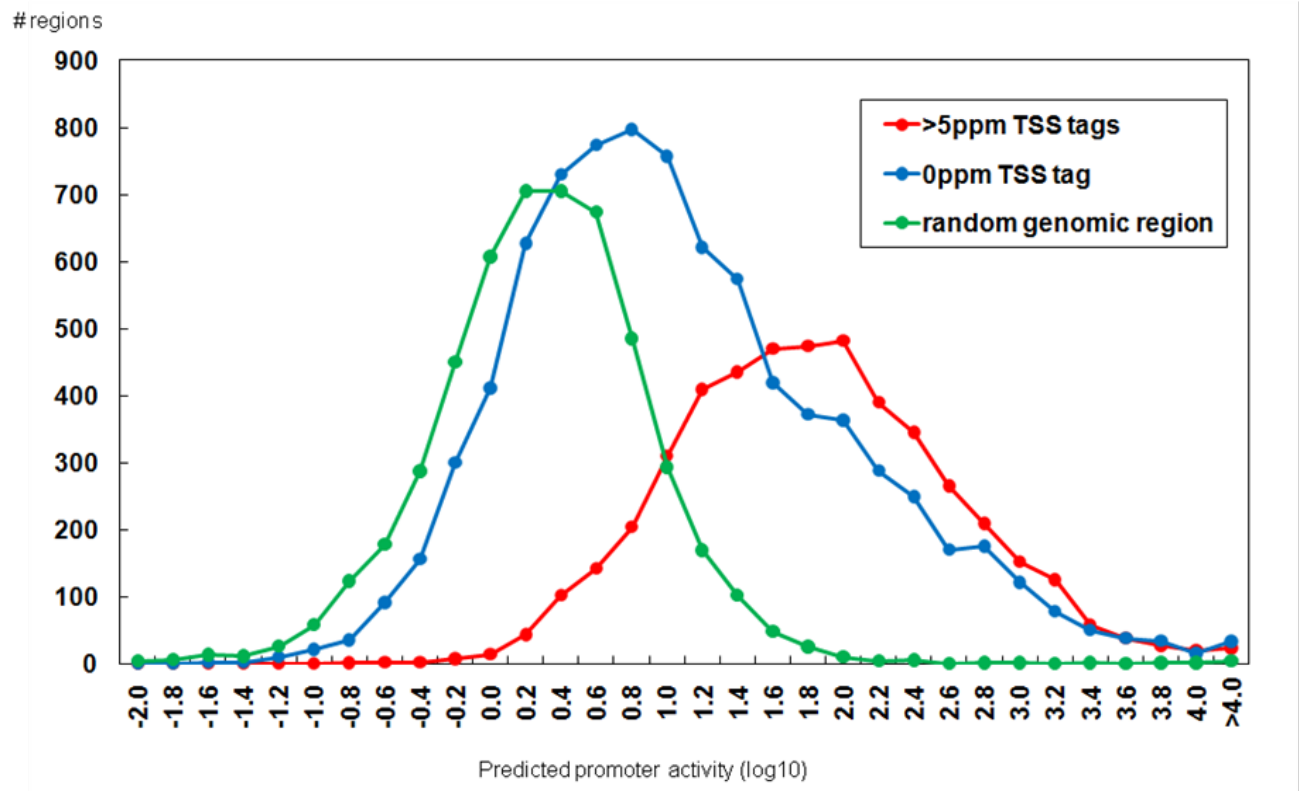


図3-12　RefSeq遺伝子5'領域のプロモーター活性予測値の分布

(赤)>5ppm以上のRefSeq5'端領域，(青)0ppmのRefSeq5'端領域，(緑)ヒトゲノム配列からランダム抽出された1.2kbの領域のプロモーター活性予測値の分布を示す．

全18,686RefSeq遺伝子中，4,749 (25%)はHEK293細胞中でTSSタグが>5ppmであった．このうちの3,922 (83%)については予測スコアが>1であった．またPrecision（適合率）とRecall（再現率）値での評価を行った．Precision（適合率）とRecall（再現率）値はそれぞれ，予測スコアが>1の領域中で>5ppmのTSSがあるRefSeq5'端領域の割合，>5ppmのTSSがある領域中で予測スコアが>1のRefSeq5'端領域の割合を示す．この閾値ではTSSタグの予測のPrecision，Recall値は(0.52，0.83)であった．またTSSの閾値を>1ppmの条件ではPrecision，Recall値は(0.63，0.83)であった．また予測スコア，ppmについて別の閾値を用いた条件の結果は表3-11にまとめた．本研究においてはPrecision，Recall値がどちらも高い値をとる，プロモーター活性予測値1を基準にした．定性的にプロモーターを分類することが可能であったが，Precisionの値が低い値を取る傾向にあったため，予測スコア>1のうちにTSSが見られないプロモーターのグループが多数あることを意味している．

表3-11 TSSタグ数とプロモーター活性予測値それぞれの閾値ごとのPrecisionとRecall値

評価には中間のタグ数のRefSeq遺伝子については計算には用いなかった. TSSの閾値がそれぞれppm >5, >2.5, >1のとき0<TSS<5ppmの5,622領域, 0<TSS<2.5ppmの4,352領域, 0<TSS<1ppmの2,850領域については中間領域として計算から除外してある.

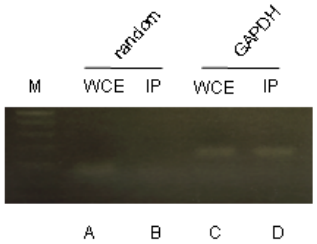
	Prediction score >2	Prediction score >1	Prediction score >0
TSS tag ppm >5			
Precision	0.57	0.52	0.39
Recall	0.35	0.83	0.99
TSS tag ppm >2.5			
Precision	0.63	0.58	0.45
Recall	0.35	0.83	0.99
TSS tag ppm >1			
Precision	0.69	0.63	0.51
Recall	0.36	0.83	0.99
TSS tag ppm >0			
Precision	0.76	0.70	0.58
Recall	0.37	0.82	0.99

3.7.2 RNA ポリメラーゼ II の ChIP-Seq を用いた評価

TSSが観察されなかったRefSeq5'端が8,315領域あり, その中の3,600(43%)領域では予測スコア1以上の高いプロモーター活性予測値を得た. その理由を調べる目的に, 転写に関わっている情報(ChIP-Seq(Pol II), Nucleosome-Seq)との比較を行った.

RNAポリメラーゼII (Pol II)のChIP-Seqの情報を利用して, RefSeq5'端周辺領域のPol IIの結合位置を調べた. Pol IIのChIP-Seqのタグの集計結果やPol IIの結合の判定条件について図3-13にまとめた. その結果, 図3-14に示すように, 高いプロモーター活性予測スコアを与え, Pol IIの結合, TSSのタグが観察された遺伝子例(A)だけではなく, TSSのタグは観察されなかったが, 高いプロモーター活性予測値を与えた遺伝子で, Pol IIの結合が確認される遺伝子例(B)が見つかった. またPol IIの結合しているRefSeq5'端領域の割合を図3-15に示した. TSSがないRefSeq5'端領域において, プロモーター活性予測値のスコアと比例して, Pol IIの結合している割合の増加が見られた.

A



Enrich rate = (A/B)/(C/D)= 408

B

	#total peak	#peak in RefSeq region	#peak outside of RefSeq region	estimated false detection rate	
fold (IP/WCE in 120 bp bin)= 1	2,058,235	1,240,231	818,004	0.26	
2	215,099	158,564	56,535	0.08	
3	71,609	60,241	11,368	0.02	
4	51,508	44,445	7,063	0.004	
5	43,214	37,696	5,518	0.0006	*parameters used in this study
6	38,332	33,694	4,638	8E-05	
7	34,867	30,862	4,005	1E-05	
8	32,431	28,858	3,573	1E-06	
9	30,288	27,096	3,192	1E-07	
10	28,480	25,575	2,905	1E-08	

C

	#total peak	#peak in RefSeq region	#peak outside of RefSeq region	
window size= 60 bp	60,756	52,431	8,325	
120 bp	43,214	37,696	5,518	*parameters used in this study
180 bp	29,773	26,460	3,313	

図3-13 RNAポリメラーゼIIのChIP-Seqの評価

(A)Real-time PCRによる既知のPol II結合サイトの確認. ランダムゲノム領域には5'-CGTGTCCCCCATATCAGAAC-3'と5'-TCAGCCTCAGTCTCCCTTGT-3'のプライマーを用い, GAPDH(TSSから-95bp ～ 49bp)には5'-CGTAGCTCAGGCCTCAAGAC-3'と5'-GTCGAACAGGAGGAGCAGAG-3'のプライマーを用いた. "Enriched rate"の計算にはreal-time PCRのCt値を利用した.

(B), (C)異なる閾値条件でのPol IIの結合サイトの集計.

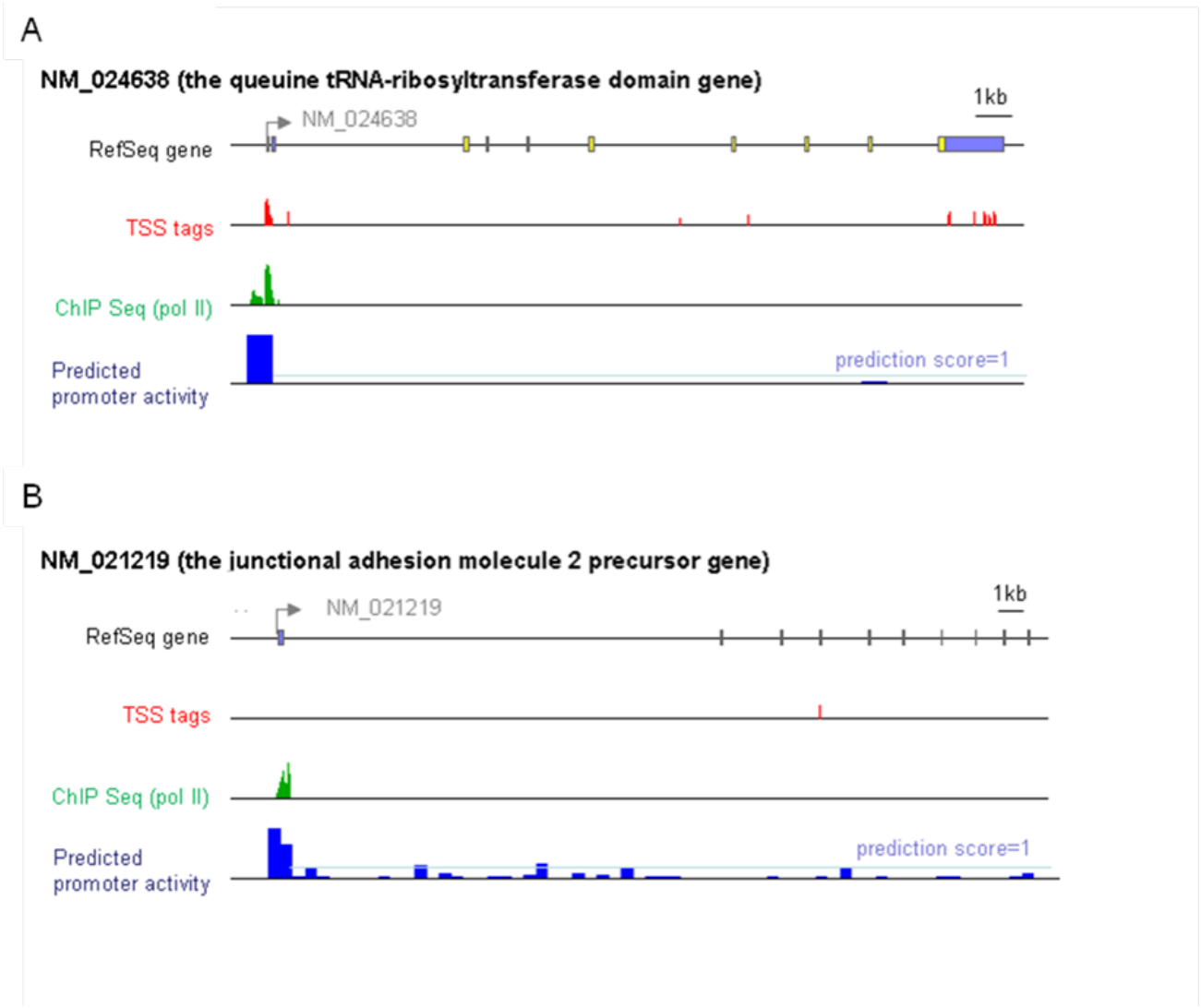


図3-14 プロモーター活性予測値，TSS，PolIIIの結合についての遺伝子例．

上から，TSS-Seqのタグ数，Pol IIのChIP-Seq，プロモーター活性予測スコアの分布を表す．淡青の先はプロモーター活性予測スコアが1を示す．(A)高いプロモーター活性予測スコア，TSSのタグ，Pol IIの結合が確認された遺伝子例（NM_024638，QTRTD1）．(B)TSSのタグはないが，高いプロモーター活性，Pol IIの結合が確認された遺伝子例（NM_021219，JAM2）．

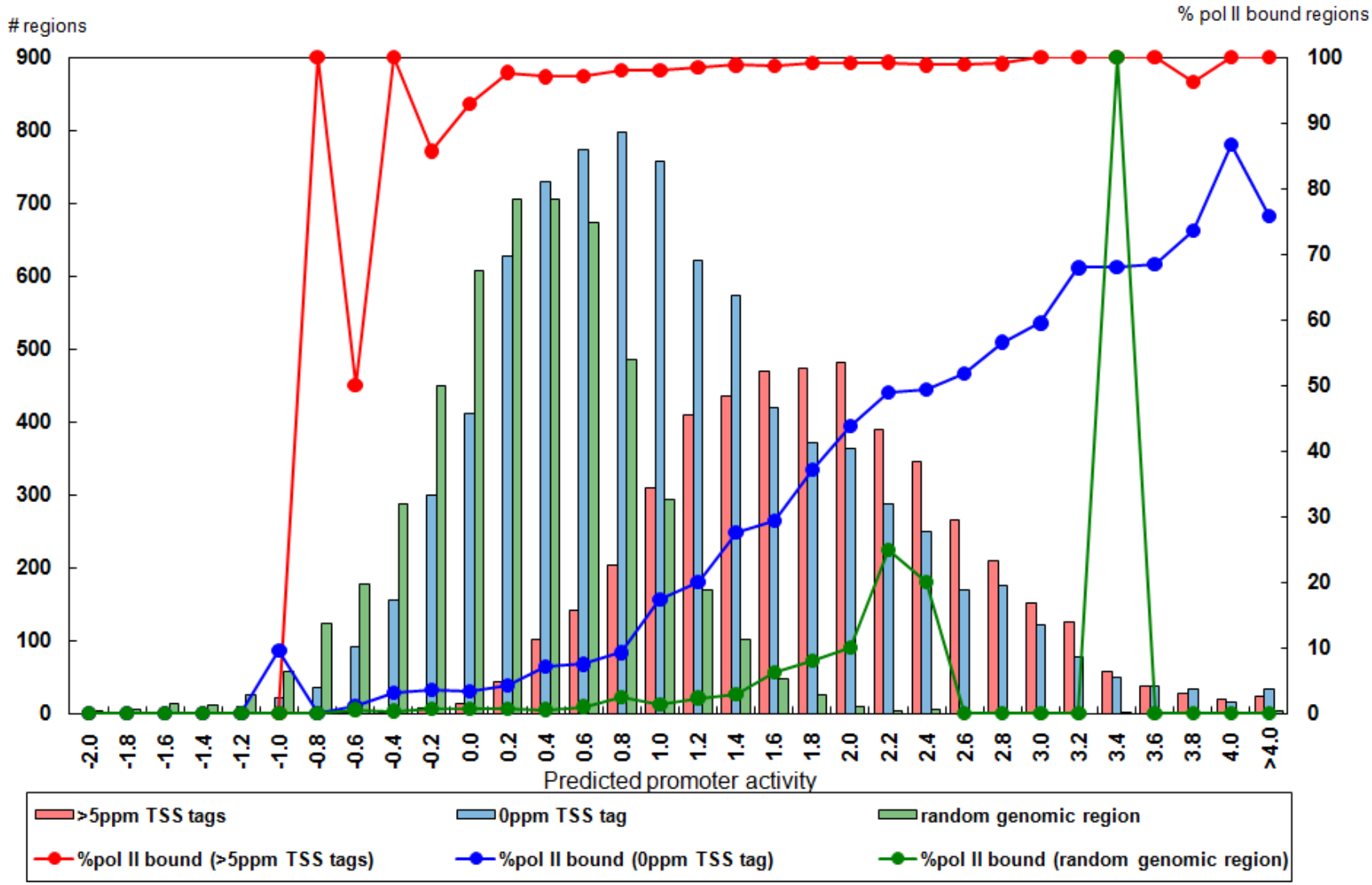


図3-15 TSS seqとPol II seqを用いたin vivoにおける予測モデルの評価

RefSeq遺伝子の5'上流領域のプロモーター活性の予測値スコア(x軸)の分布のヒストグラム(頻度; y軸左側), 棒グラフ赤(>5ppm), 青(0ppm), 緑(ランダム領域). 折れ線グラフはChIP Seq(pol II)の結合が確認された領域の割合 (y軸右側). 赤(>5 ppm), 青(0ppm), 緑(ランダム領域)

3.7.3　Nucleosome-Seq を用いた評価

HEK293細胞のヌクレオソーム構造を調べた．ヌクレオソーム構造の解析には，HEK293細胞のNucleosome-Seqデータを利用した(図3-16)．Nucleosome-Seqのタグの集計結果は表3-10にまとめた．RefSeq遺伝子周辺のヌクレオソームの位置情報の解析を行ったところ，予測スコア>1，TSSタグカウント>5ppmのRefSeq領域については，転写開始点近傍でヌクレオソーム占有率のスコアが低い傾向にあった．これはクロマチンが開いた構造をとっていることを示している（図3-17A）．さらに予測スコア>1，TSSタグカウント0の領域についても同様に開いたクロマチン構造をとっていた（図3-17B）．このような領域は潜在的に高いプロモーター活性を持ち，開いたクロマチン構造を形成し，Pol IIのリクルートの効率を制御しうる能力を持つ領域であるが，転写伸長を起こす何らかの因子が不足している例を示唆している．またプロモーター予測モデルが，転写産物の有無に関わらず，潜在的に転写活性化能を持つ領域を予測できたと考えている．

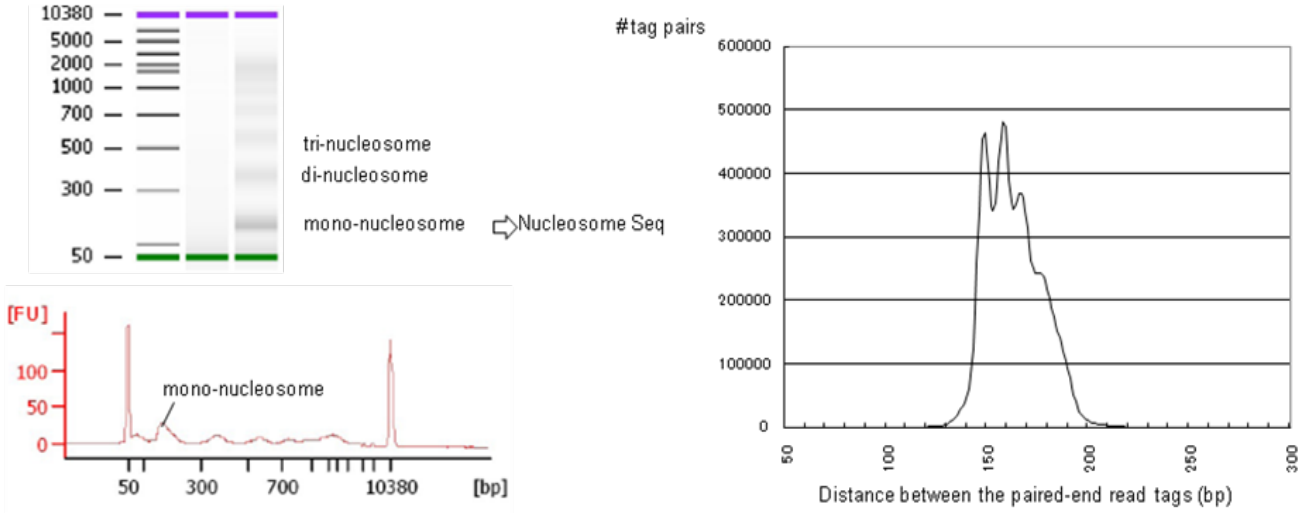


図3-16　Nucleosome-Seqの評価

（左）Nucleosome-Seqに用いたMicrococcal nuclease処理後のDNAをBIOANLYZER (Agilent)で定量を行った．モノヌクレソームに当たるDNA断片を回収しIllumina GAによる配列決定を行った．（右）Paired-endタグから算出したDNA断片の長さ．

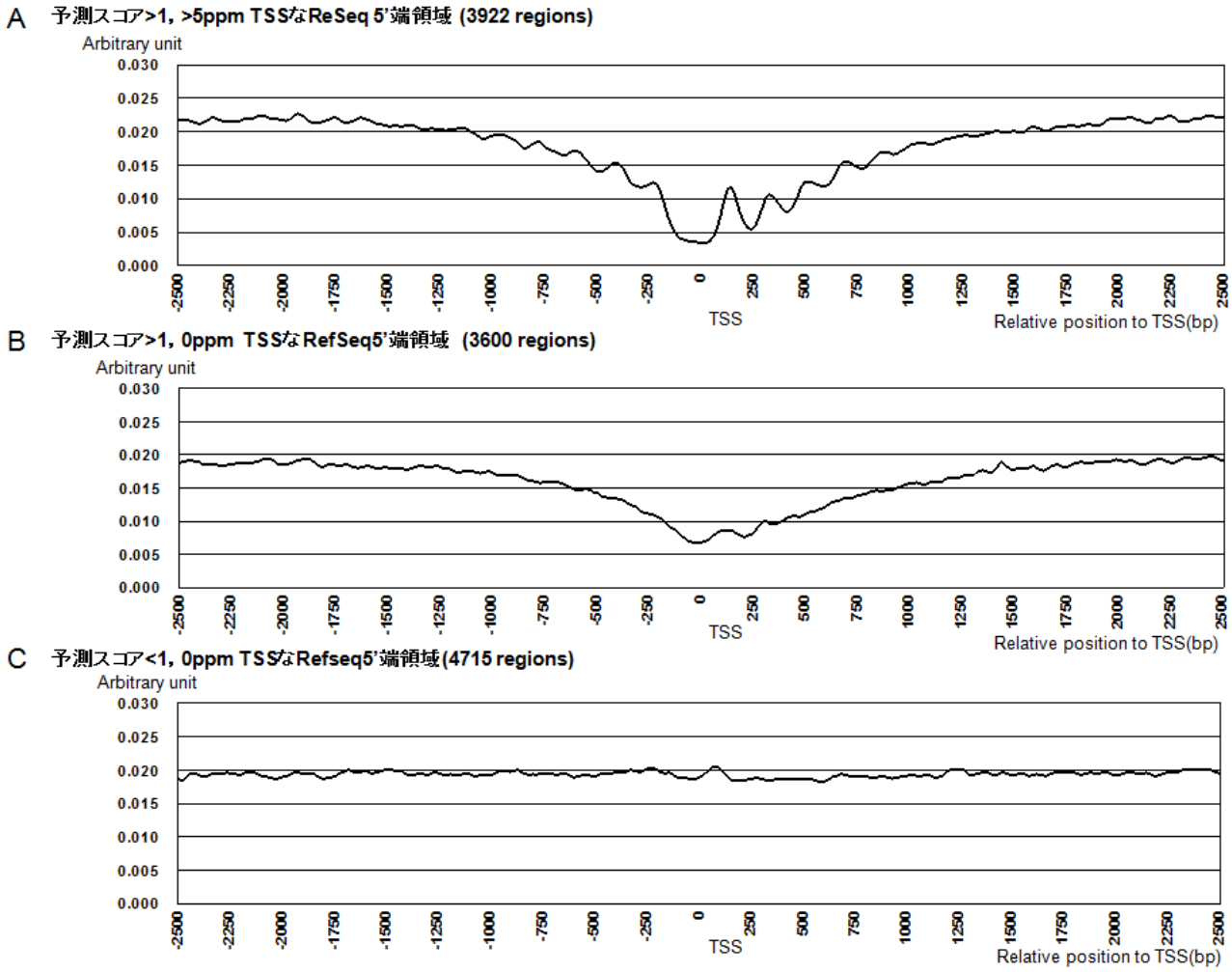


図3-17　RefSeq遺伝子5'端のヌクレオソーム構造

RefSeq遺伝子の5'端のヌクレオソーム構造を示す. (A)予測スコア>1, TSSタグ>5ppmのReSeq遺伝子. (B)予測スコア>1, TSSタグ0ppmのRefSeq遺伝子. (C)予測スコア<1, TSSタグ0ppmの遺伝子. 転写開始点を基準(0)として, ゲノム領域(*x*軸)のヌクレオソーム占有率(*y*軸)の計算した. 計算方法については2.10.3を参照.

3.8　ヒトゲノムの予測プロモーター活性のランドスケープ

ヒトゲノム全体のプロモーター活性予測値の計算を行った. ヒトゲノム配列を1.2kbの幅で分け, それぞれの領域でプロモーター活性予測値の算出を行った. 得られたプロモーター活性予測値のヒトゲノム上での分布とTSS-seq, Pol IIのChIP-Seq, Nucleosome-Seqとの比較を行った. RefSeq遺伝子領域以外のゲノム領域を1.2kbの幅で分け, 合計2,650,838領域のプロモーター活性予測値の算出とTSS-Seq, Pol IIのChIP-Seq, Nucleosome-Seqとの比較を行った. 解析した185,018領域についてプロモーター活性の予測スコア>1を与えた. TSS-Seqのタグ数を調べたところ, このうち147領域について>5ppmのTSSが存在していた. このうち97(66%)についてはPol IIの結合がChIP-Seqにより確認された. したがって高いプロモーター活性予測値を

持っている領域であるため，転写制御を受けた転写産物であることが示唆される．詳細を表3-12にまとめた．遺伝子間領域についての例として，プロモーター活性予測値，TSS-seq，ChIP-Seqのピークを図3-18に示した．例に示したのはChr5:7872000-103720000の領域で，NM_024010とNM_012073の間の遺伝子間領域を表している．これらRefSeq遺伝子の例では高いプロモーター活性予測値があり，Pol IIが結合し，転写開始点が得られている典型的と考えられる例である．これらの遺伝子間領域中には幾つかlncRNAとか考えられるcDNAが得られている領域があった．こらのlncRNAの上流領域に高いプロモーター活性予測値を得る領域があり，Pol IIの結合，TSSが確認できる例と高いプロモーター活性予測値とPol IIの結合はあるがTSSが確認できない例などRefSeq遺伝子と同様の例を見つけることが可能であった．これらのlncRNAはノイズレベルの転写産物ではなく，転写制御を受けたものであることを示唆している．

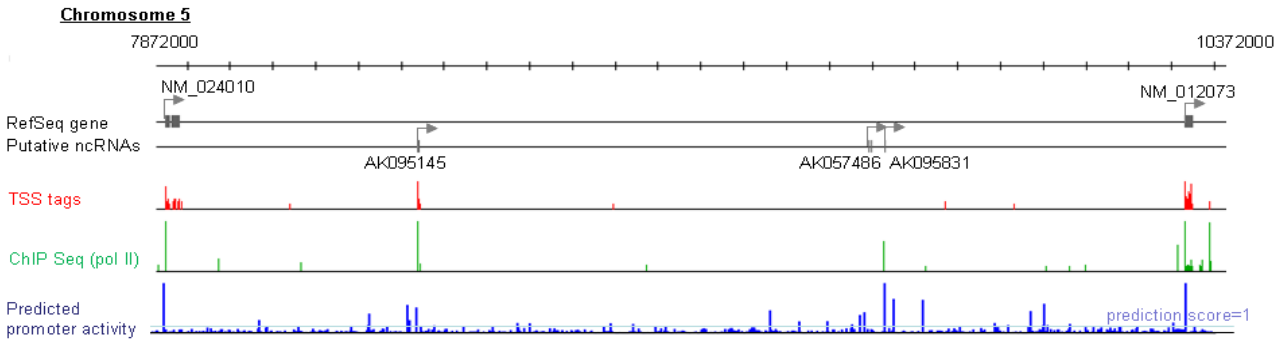


図3-18　ヒトゲノムのプロモーター活性予測値のランドスケープ

ヒトゲノム配列を用いてプロモーター活性予測値の算出を行った．Chr5:7872000-103720000の領域を例として示す．上から，TSS-Seqのタグ数，Pol IIのChIP-Seq，プロモーター活性予測スコアの分布を表す．淡青の先はプロモーター活性予測スコアが1を示す．

表3-12　潜在的に高いプロモーター活性予測値を有していた領域のTSSタグ，Pol IIの結合についての統計．

	RefSeq regions	intergenic regions
#genomic regions having prediction scores of > 1	12,089	185,018
#genomic regions having >5ppm TSS tags	4,749	766
#genomic regions having prediction scores of > 1 and TSS tags of >5ppm	3,922	147
#genomic regions having prediction scores of >1, TSS tags of >5ppm and pol II binding signals	3,884	97
#genomic regions having prediction scores of >1 and no TSS tag	3,600	182,140
#genomic regions having prediction scores of >1, no TSS tag and pol II binding signals	1,401	3,077
#genomic regions having prediction scores of 0-1	5,444	1,499,210

次に遺伝子間領域において高いプロモーター活性予測値を得た領域についてのヌクレオソーム構造について調べた. 高いプロモーター活性予測値, >5ppmのTSSを得た147領域についてヌクレオソーム構造を調べたところ, 開いたクロマチン構造を取り(図3-19 A). プロモーター活性予測値>1, 0ppmの182,140領域についてのヌクレオソーム構造についても調べたところ開いたクロマチン構造を取る傾向にあった(図3-19 B). これらの結果は遺伝子間領域の大部分の領域についても, 潜在的なプロモーター活性を持つ可能性があると考えられる. また転写産物レベルにおいては別の要因によって抑制され, 実際の転写産物が見つからないものと考えられる. また遺伝子間領域のプロモーター活性もしくはcDNAクローニング時の実験的なエラーやランダムに起こる散発的な転写産物では無いことを示唆する.

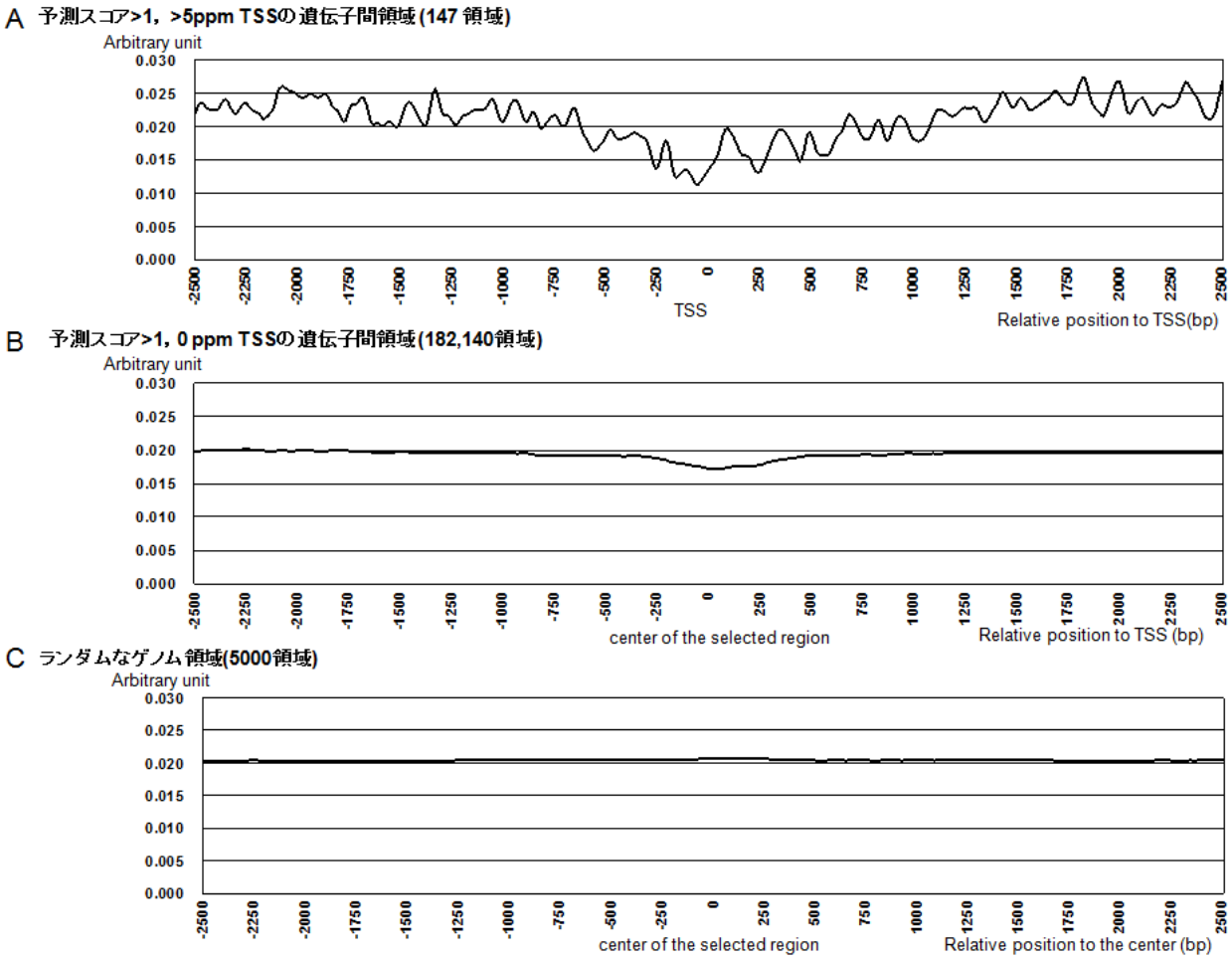


図3-19 遺伝子間領域のヌクレオソーム構造

遺伝子間領域ののヌクレオソーム構造を示す. (A)予測スコア>1, TSSタグ>5ppmのReSeq遺伝子. (B)予測スコア>1, TSSタグ0ppmのRefSeq遺伝子. (C)予測スコア<1, TSSタグ0ppmの遺伝子. 転写開始点を基準(0)として, ゲノム領域(x軸)のヌクレオソーム占有率(y軸)の計算した.

4. 考察

4.1 プロモーター活性予測モデルについて

本研究では，HEK293細胞における体系的なルシフェラーゼ活性情報を利用して，ヒトのDNA一次配列のプロモーター活性予測モデルの構築を行った．予測モデルの精度は実験値の5倍以内の範囲で86%のDNA断片の活性を予測が可能であり，ルシフェラーゼによる実験値と予測値の相関係数（Pearson's correlation coefficient）はr=0.87（プロモーターのみ，ランダムのみの場合それぞれr=0.66，r=0.32）であった（図3-4）．プロモーター活性の説明は，各々のTFBSのスコアの和という比較的単純な予測モデルではあるが，高い予測精度であり，先行研究と比較しても高い精度で予測が出来たと考えている．この要因としては，先行研究の殆どはマイクロアレイを用いた遺伝子発現情報を用いていたのに対し，本研究ではプロモーター活性情報に定量的なルシフェラーゼアッセイ情報を基にしており，DNA配列と内在しているプロモーター活性の間のみの関係の情報を用いたためであると考えられる．また今回は線形和モデルを基にしたシンプルな予測モデルであるため，各種パラメーターを入れ込むなど改良を加えやすいため，さらに予測精度の高いモデルを構築することができると期待される．プロモーター活性予測モデルに加えることで予測モデルの精度の向上が可能であると考えられるパラメーターとしては，他生物との進化的保存度，新規のDNAモチーフの探索などがあげられる．また転写開始点付近のCGコンテント，CpGスコアやDNA構造の湾曲能（bendability）・塩基対の融解温度（melting temprature）などDNAの物理的なパラメーターもプロモーター領域予測に用いている例[50, 61]もある．また今回のプロモーター活性予測モデルでは考慮しなかったが転写因子間の相互作用についても考慮すべきかもしれない．

本研究ではプロモーター活性予測モデルは線形和モデルとしパラメーターの推定に重回帰分析の手法を用いた．単純な手法ではあるが，利点としては変数として用いたTFBSのスコアをプロモーター活性として利用できることである．Support vector machine (SVM)やベイジアンネットワークなどの手法も予測モデルに用いられることもある．しかしこれらのモデルはしばしば変数間や説明変数との関連性が複雑になるため，TFBSとプロモーター活性の直接の関連付けが難しいことある．本研究のプロモーター活性予測モデルでは167種類のTFBSを変数として用い，AICによる変数選択で85種類のTFBSがプロモーター活性の説明に寄与が大きい変数として選択された（表3-5）．これら変数として用いたTFBSのスコアをプロモーター活性と用いることができる利点がある．ヒト遺伝子においてDNA結合ドメインを持つ遺伝子は約2600個存在しているといわれいている[62]．一部が細胞中で発現しているとしても，モデル構築に用いた変数の数よりは多い数の転写因子が発現していると考えられる．そのためプロモーター活性予測モデルの予

測精度向上のためには結合箇所が未知の転写因子の結合配列の情報が必要になってくる可能性がある. 逆に, 85種類程度でも予測モデルが構築できていることから, 転写因子間で結合配列が縮重しており, 結果的にTFBS側の配列のcomplexityはそれほど大きくないのかもしれない. complexityを下げている要因としてはCpGサイトが挙げられれ, 実際にCpGスコアとルシフェラーゼ活性との間にある程度相関が見られた.

TFBS欠失配列の実験結果から約半数のTFBSの例について活性の減少を見出すことができ(図3-7, 表3-6), DNA配列改変を行うことでプロモーター活性に寄与する領域の検出が可能であるという点もプラスミドを用いたレポーターアッセイの利点であるといえる. 変異DNAとプロモーター活性の変化を体系的に解析する手法を確立することができれば, 塩基単位の変化とプロモーター活性の関係性を明らかにすることが可能になるかもしれない. さらに1塩基レベルの違いでプロモーター活性予測の説明が可能であれば, プロモーター領域に存在しているSNPである”Regulatory SNP”中で特にプロモーター活性の変化に関わっているSNPの候補の探索が可能になるかもしれない.

Landolin et al Genome Res (2010)[60]の8細胞種におけるプロモーター活性情報を用いてプロモーター活性予測モデルを構築した結果, それぞれの細胞ごとに相関係数約0.6の精度のモデル構築が可能であった(表3-9). これは本研究において用いたHEK293細胞のプロモーター活性予測モデルと同程度の精度(r=0.66)であった. したがってHEK293細胞以外の細胞環境下においても同様の手法でプロモーター活性予測モデルの構築が可能であるといえる.

細胞ごとのプロモーター活性予測モデルでは細胞ごとにTFBSのスコアが与えられるためそれぞれの細胞に応じたプロモーター活性予測が可能であるといえる. したがってTFBSのスコアを適当な数値に改変する事で異なる組織におけるプロモーター活性予測が可能になると考えられる. この時必要になると考えられる情報としては, 転写因子の発現量の情報である. 例えば本研究で用いたTSS-Seqが有用であると考えられる. 組織ごとに得られたmRNAからのTSS-Seqの転写因子の発現頻度情報を利用する. HEK293細胞で構築した予測モデルに対して発現量の高い転写因子のTFBSには高いスコアを与え, 発現が見られない転写因子のTFBSのスコアを0にするというようなTFBSのスコアに変更を加えることで細胞・組織ごとのプロモーター活性予測モデルへの変換することができるかもしれない.

さらにプロモーター活性予測モデルの逆問題である任意のプロモーター活性をもつようなDNA断片のデザインが可能になるかもしれない. 任意の細胞環境下において任意のプロモーター活性をもつようなDNA次配列のデザインが可能であれば産業的にも有用であると考えられる.

4.2　プロモーター活性予測モデルとヒトゲノム配列について

構築したプロモーター活性予測モデルとHEK293細胞における*in vivo*の転写活性化能の指標としたTSS数の比較を行った. プロモーター活性予測モデルによる予測値とTSSの指標にした転写活性化能との相関は見られなかった. 図1-1で示したように, mRNAの合成は様々な制御を経ているため, 細胞内でのmRNAのコピー数を決めるためにはそれぞれの制御段階のモデル化が必要であると考えられる. RNAの発現量の絶対値の推定を試みる場合は, DNAのメチル化, ヒストン修飾やmRNAの合成速度・分解速度などの情報を取り入れた数学モデルが必要であると考えられる. しかしながら, 定量的な相関は見られなかったものの, HEK293細胞において転写開始点が存在している"active"な遺伝子のプロモーターと転写開始点が見られなかった"silent"な遺伝子のプロモーターとの間でプロモーター活性予測値の定性的な分類を行うことが可能であった(図3-12). これは転写が行われているプロモーターと行われていないプロモーターの分類が可能であることを意味している. TSSを指標としたプロモーターの定性的な分類は可能であったが, TSSが観測されないものの高いプロモーター活性予測値を有する領域も存在していた. 次世代シークエンサーを用いたRNAポリメラーゼII(Pol II)のChIP-Seq, ヌクレオソーム構造を調べるNucleosome-Seqの情報を用いた解析を行ったところ, TSSが見られないプロモーターであっても, 高いプロモーター活性予測値を得たプロモーターにおいてはPol IIの結合が起こり, 転写が行われている遺伝子に特徴的である開いたクロマチン構造をとる傾向にあった(図3-14, 図3-15, 図3-17). このような高いプロモーター活性予測値を得た領域は転写産物が速やかに分解される遺伝子か, クロマチン構造を開き, Pol IIのリクルートが行われるものの転写伸長に関する因子にかけているため転写産物が見られない例であると考えられる. 近年, 転写開始数十塩基辺りで転写が一時停止を引き起こすpromoter proximal pausing[63-66]という現象が知られているが, 転写が行われていないが高いプロモーター活性予測値を得た領域のうちの幾つかについてもこのような転写停止の例であると考えられ, 刺激や分化などにおいて, 素早い転写応答を行う必要がある遺伝子例である可能性がある.

プロモーター活性予測モデルを用いてヒトゲノム配列全体のプロモーター活性予測値の分布の解析を行った. 遺伝子間領域に存在しているlncRNAと考えられる完全長cDNAクローンとのオーバーラップを調べた結果, 完全長cDNA5'端領域付近に高いプロモーター活性予測値を得る領域があった. したがってこのような完全長cDNAの例についてはノイズレベルの転写産物ではなく高いプロモーター活性をもつ領域から制御を受けた転写産物であることを示唆する. したがって, 何らかの転写制御されてきたRNAであるというlncRNAの意義づけが可能になるのではないかと期待している. 遺伝子間領域のプロモーター活性予測値の分布を解析した結果, 転写の有無に関わらず高いプロモーター活性予測値を与える領域が多数見られた. このような高いプロモーター活性予測値を得た領域付近を詳細に解析することで新規の遺伝子の発見が可能になるかもしれない. 遺伝子間領域中にはに転写は見られないものの高いプロモーター活性予

測値を得る領域が182,140あった. これらの領域のヌクレオソーム構造を調べた結果, 開いたクロマチン構造をとる傾向にあり(図3-19), ヒトゲノム中において下流の転写産物は観察されないながらも, 潜在的に高いプロモーター活性を有する領域が多数存在することが示唆された. 遺伝子間領域の潜在的にプロモーター活性を持つ領域が生物学的意義を持っていいるのかどうかについては今のところ明らかではない. 潜在的なプロモーター活性を持つ領域が, 進化上で新規な遺伝子のためのプロモーター領域を提供するような役割を担っているのかもしれない. 細胞中のヒトゲノム全体にわたる転写の正確な予測のためには, その他にも様々な情報が必要になると考えられ, 転写予測のためのさらなる改良点には, 様々な生物学的パラメーターを考慮に入れることも必要である. 例えば, エンハンサーなどの遠位DNAエレメント[67], DNAのメチル化の影響[68-69], ヒストンのメチル化やアセチル化[19], ゲノムDNAの3次元の立体構造[70-73]等のゲノムやクロマチンの構造やmRNAの合成・分解速度[74]が挙げられる. これらの情報を用いた統合的なアプローチを通じて細胞内の転写の全体像が見えてくると思われる.

5. 総括

5.1. プロモーター活性予測モデルについて

本研究では，DNA一次配列からの内在的なプロモーター活性予測モデルの構築を行った．ヒト培養細胞のHEK293細胞内のプロモーター活性情報として約700種類のDNA断片のルシフェラーゼ活性情報を基にし，プロモーター活性をTFBSの転写への寄与のスコアの和とした線形和モデルとして捉え，重回帰分析の手法でプロモーター活性モデルの構築を試みた．プロモーター活性の実験値と予測値の相関係数（Pearson's correlation coefficient）がr=0.87，プロモータークローン内，ランダムクローン内の相関係数はそれぞれr=0.66，r=0.32であり予測精度を達成することができた．さらにプロモーター活性に関与していると考えられるパラメーターを導入することでプロモーター活性予測モデルの精度向上が可能であると考えられる．また，10分割交差検定の結果，相関係数の平均が0.83となり未知なDNA配列へも有用なモデルを構築できたと考えている．以上のことから定量的ルシフェラーゼアッセイを基準にしたプロモーター活性予測モデルできたと考えている．

*in vivo*な転写活性の指標としてTSS-Seqの情報を用いてプロモーター活性予測モデルの評価を行った．RefSeq5'端領域のプロモーター活性予測値とTSSのタグ数との相関は見られなかったが，RefSeq5'端領域をTSSが観察された領域と観察されない領域で分類を行ったとき，それぞれの群でプロモーター活性の予測値の分布に有意な差が見られた(図3-12)．さらにRNAポリメラーゼIIのChIP-SeqとNucleosome-Seqの情報との比較を行った結果，TSSが観察されないRefSeq5'端領域で高いプロモーター活性予測値を得た領域もおいて，RNAポリメラーゼIIの結合している割合が増加し（図3-14），また開いたクロマチン構造をとる傾向にあり（図3-17），転写の行われている領域と同様の特徴が見られた．このような例はプロモーター活性予測モデルが転写の有無によらず，このような高いプロモーター活性予測値を得た領域は転写産物が速やかに分解される遺伝子か，クロマチン構造を開き，Pol IIのリクルートが行われるものの転写伸長に関する因子にかけているため転写産物が見られない例であると考えられる．また遺伝子間領域においても転写産物が観察されないものの高いプロモーター活性予測値を得た領域についてはそうでない領域と比べ，Pol IIの結合，開いたクロマチン構造をとる傾向にあった．したがって，プロモーター活性予測モデルは潜在的にプロモーター活性を持つ領域にも高いスコアを与える傾向にあるといえる．また潜在的にプロモーター活性を持っていると領域がゲノム中に多数存在していることが示唆された．

5.2. 今後の展望

本研究のプロモーター活性予測モデルではプロモーター活性を既知のTFBSを説明変数とした線形和モデルで記述していており，比較的単純なモデルであるといえる．ある程度の予測精度を達成することができたが，さらに予測精度を向上させることが可能であると考えている．例えば，他生物との進化的保存度，新規のDNAモチーフの探索などがあげられる．また転写開始点付近のDNA構造の湾曲能（bendability）・塩基対の融解温度（melting temprature）などDNAの物理的なパラメーターやCGコンテントなど予測モデルに応用が可能であると考えられる．またTF-TF間の相互作用も転写活性の説明に重要であると考えられる．

Landolinらの8種類の培養細胞における体系的なルシフェラーゼアッセイの情報[60]を用いてプロモーター活性予測モデルの構築を行った結果，それぞれの細胞環境下において相関係数で約r=0.6程度を得た（表3-9）．つまり本研究で用いたHEK293以外の細胞においても同様の手法を用いてプロモーター活性予測モデルが構築可能であった．各細胞間におけるプロモーター活性予測モデルは説明変数であるTFBSのスコアの値が細胞ごとに最適化されているため，各細胞のプロモーター活性の説明が可能になっているといえる．そのため今後は本研究の予測モデルを基準にして，TFBSのスコアを変更する行列変換のようなことを行うことで細胞ごとのプロモーター活性予測が可能になると考えられ，数百種類のクローンを用いたルシフェラーゼアッセイを行わなくても細胞ごとのプロモーター活性予測が可能になるかも知れない．そのためには転写因子のmRNAなどの発現量情報が必要になってくると考えられ，3-7で用いたTSS-Seqのようなデジタルな遺伝子発現量情報が有用であると考えられる．

DNA一次配列情報からのプロモーター活性の予測モデルがある程度構築可能であった．次はその逆問題である，ある細胞中またはあらゆる細胞において任意のプロモーター活性を持つようなDNA断片のデザインが可能であるかどうか，というような問題が出てくる．もし自由にプロモーター配列のデザインができるのであれば，産業的にも非常に有用であるため，期待される課題であると思われる．

ヒトゲノム配列のプロモーター活性予測値の解析を行い，RefSeq遺伝子5’端領域にマップされたTSS-Seqのタグ数との相関を調べたところ，相関はほとんど見られなかった．ゲノム配列からmRNAの発現量予測のためには，プロモーター配列以外の変数を加えたモデル化も必要になってくる．そのための改良には，プロモーター配列以外の転写に関わると考えられる情報を考慮に入れることも必要であり，例えばエンハンサーなどの遠位DNAエレメント[67]，DNAのメチル化の影響[68-69]，ヒストンのメチル化やアセチル化[19]，ゲノムDNAの3次元構造[70-73]等が挙げられる．さらにmRNAの合成・分解速度[74]のパラメーターを組み込んだ数理モデルの構築によって細胞内でのmRNAのコピー数予測も可能になると考えられる．これらの情報は，ゲノム規模での網羅的な解析手法がその根幹となるが，次世代型シークエンサーの登場によって転写の全体像の統合的な理解が進み，ゲノム規模での転写制御モデルの現実味を帯びてきているといえる．本研究において構築したプロモーター活性予測モデルが，ゲノム配列からトランスクリ

プトームを繋ぐ，転写制御機構の包括的な理解へ向けての足がかりになると期待している．

6. 参考文献

1. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-151.
2. Collins, F.S., et al., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.
3. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
4. Suzuki, Y. and S. Sugano, *Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method.* Methods Mol Biol, 2003. **221**: p. 73-91.
5. Imanishi, T., et al., *Integrative annotation of 21,037 human genes validated by full-length cDNA clones.* PLoS Biol, 2004. **2**(6): p. e162.
6. Suzuki, Y., et al., *Identification and characterization of the potential promoter regions of 1031 kinds of human genes.* Genome Res, 2001. **11**(5): p. 677-84.
7. Suzuki, Y., et al., *Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites.* EMBO Rep, 2001. **2**(5): p. 388-93.
8. Yamashita, R., et al., *DBTSS provides a tissue specific dynamic view of Transcription Start Sites.* Nucleic Acids Res, 2009.
9. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-1563.
10. Kimura, K., et al., *Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes.* Genome Res, 2006. **16**(1): p. 55-65.
11. Tsuchihara, K., et al., *Massive transcriptional start site analysis of human genes in hypoxia cells.* Nucleic Acids Res, 2009. **37**(7): p. 2249-63.
12. Sanger, F., et al., *Nucleotide-Sequence of Bacteriophage Phichi174 DNA.* Nature, 1977. **265**(5596): p. 687-695.
13. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-380.
14. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-59.
15. Metzker, M.L., *Sequencing technologies - the next generation.* Nature Reviews Genetics, 2010. **11**(1): p. 31-46.
16. Wilhelm, B.T. and J.R. Landry, *RNA-Seq-quantitative measurement of expression*

through massively parallel RNA-sequencing. Methods, 2009. **48**(3): p. 249-257.

17. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature Reviews Genetics, 2009. **10**(1): p. 57-63.
18. Sultan, M., et al., *A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.* Science, 2008. **321**(5891): p. 956-60.
19. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-837.
20. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nature Reviews Genetics, 2009. **10**(10): p. 669-680.
21. Jiang, C.Z. and B.F. Pugh, *Nucleosome positioning and gene regulation: advances through genomics.* Nature Reviews Genetics, 2009. **10**(3): p. 161-172.
22. Lister, R., et al., *Highly integrated single-base resolution maps of the epigenome in Arabidopsis.* Cell, 2008. **133**(3): p. 523-536.
23. Shendure, J. and H.L. Ji, *Next-generation DNA sequencing.* Nature Biotechnology, 2008. **26**(10): p. 1135-1145.
24. Wheeler, D.A., et al., *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-U5.
25. Cox-Foster, D.L., et al., *A metagenomic survey of microbes in honey bee colony collapse disorder.* Science, 2007. **318**(5848): p. 283-287.
26. Wilhelm, B.T., et al., *Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution.* Nature, 2008. **453**(7199): p. 1239-U39.
27. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature Methods, 2008. **5**(7): p. 621-628.
28. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-1502.
29. Schones, D.E., et al., *Dynamic regulation of nucleosome positioning in the human genome.* Cell, 2008. **132**(5): p. 887-898.
30. Morin, R.D., et al., *Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.* Genome Research, 2008. **18**(4): p. 610-621.
31. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
32. Nguyen, D.H. and P. D'Haeseleer, *Deciphering principles of transcription regulation in eukaryotic genomes.* Mol Syst Biol, 2006. **2**: p. 2006 0012.
33. Zhang, Z. and J. Zhang, *Accuracy and application of the motif expression decomposition method in dissecting transcriptional regulation.* Nucleic Acids Res,

2008. **36**(10): p. 3185-93.

34. Beer, M.A. and S. Tavazoie, *Predicting gene expression from sequence.* Cell, 2004. **117**(2): p. 185-98.
35. Gertz, J. and B.A. Cohen, *Environment-specific combinatorial cis-regulation in synthetic promoters.* Mol Syst Biol, 2009. **5**: p. 244.
36. Gertz, J., E.D. Siggia, and B.A. Cohen, *Analysis of combinatorial cis-regulation in synthetic and genomic promoters.* Nature, 2009. **457**(7226): p. 215-8.
37. Suzuki, H., et al., *The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line.* Nature Genetics, 2009. **41**(5): p. 553-562.
38. Das, D., Z. Nahle, and M.Q. Zhang, *Adaptively inferring human transcriptional subnetworks.* Mol Syst Biol, 2006. **2**: p. 2006 0029.
39. Das, D., N. Banerjee, and M.Q. Zhang, *Interacting models of cooperative gene regulation.* Proc Natl Acad Sci U S A, 2004. **101**(46): p. 16234-9.
40. Khodursky, A.B. and J.A. Bernstein, *Life after transcription - revisiting the fate of messenger RNA.* Trends in Genetics, 2003. **19**(3): p. 113-115.
41. Kent, W.J., *BLAT - The BLAST-like alignment tool.* Genome Research, 2002. **12**(4): p. 656-664.
42. Florea, L., et al., *A computer program for aligning a cDNA sequence with a genomic DNA sequence.* Genome Research, 1998. **8**(9): p. 967-974.
43. Kel, A.E., et al., *MATCH (TM): a tool for searching transcription factor binding sites in DNA sequences.* Nucleic Acids Research, 2003. **31**(13): p. 3576-3579.
44. Wingender, E., *The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.* Briefings in Bioinformatics, 2008. **9**(4): p. 326-332.
45. Das, D., M. Pellegrini, and J.W. Gray, *A primer on regression methods for decoding cis-regulatory logic.* PLoS Comput Biol, 2009. **5**(1): p. e1000269.
46. Akaike, H., *Factor-Analysis and Aic.* Psychometrika, 1987. **52**(3): p. 317-332.
47. Sonnenburg, S., A. Zien, and G. Ratsch, *ARTS: accurate recognition of transcription starts in human.* Bioinformatics, 2006. **22**(14): p. e472-80.
48. Down, T.A. and T.J. Hubbard, *Computational detection and location of transcription start sites in mammalian genomic DNA.* Genome Res, 2002. **12**(3): p. 458-61.
49. Abeel, T., et al., *Generic eukaryotic core promoter prediction using structural features of DNA.* Genome Res, 2008. **18**(2): p. 310-23.
50. Abeel, T., et al., *ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles.* Bioinformatics, 2008. **24**(13): p. i24-31.

51. Knudsen, S., *Promoter2.0: for the recognition of PolII promoter sequences.* Bioinformatics, 1999. **15**(5): p. 356-61.
52. Davuluri, R.V., I. Grosse, and M.Q. Zhang, *Computational identification of promoters and first exons in the human genome.* Nat Genet, 2001. **29**(4): p. 412-7.
53. Sakakibara, Y., et al., *Intrinsic promoter activities of primary DNA sequences in the human genome.* DNA Res, 2007. **14**(2): p. 71-7.
54. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Research, 2008. **36**: p. D102-D106.
55. Carey, M., *The enhanceosome and transcriptional synergy.* Cell, 1998. **92**(1): p. 5-8.
56. Veitia, R.A., *A sigmoidal transcriptional response: cooperativity, synergy and dosage effects.* Biological Reviews, 2003. **78**(1): p. 149-170.
57. Koudritsky, M. and E. Domany, *Positional distribution of human transcription factor binding sites.* Nucleic Acids Research, 2008. **36**(21): p. 6795-6805.
58. Gray, N.K. and M. Wickens, *Control of translation initiation in animals.* Annual Review of Cell and Developmental Biology, 1998. **14**: p. 399-458.
59. Bochkov, Y.A. and A.C. Palmenberg, *Translational efficiency of EMCV IRES in bicistronic vectors is dependent upon IRES sequence and gene location.* Biotechniques, 2006. **41**(3): p. 283-+.
60. Landolin, J.M., et al., *Sequence features that drive human promoter function and tissue specificity.* Genome Research, 2010. **20**(7): p. 890-898.
61. Florquin, K., et al., *Large-scale structural analysis of the core promoter in mammalian and plant genomes.* Nucleic Acids Research, 2005. **33**(13): p. 4255-4264.
62. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Current Opinion in Structural Biology, 2004. **14**(3): p. 283-291.
63. Guenther, M.G., et al., *A chromatin landmark and transcription initiation at most promoters in human cells.* Cell, 2007. **130**(1): p. 77-88.
64. Muse, G.W., et al., *RNA polymerase is poised for activation across the genome.* Nature Genetics, 2007. **39**(12): p. 1507-1511.
65. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters.* Science, 2008. **322**(5909): p. 1845-1848.
66. Gilmour, D.S., *Promoter proximal pausing on genes in metazoans.* Chromosoma, 2009. **118**(1): p. 1-10.
67. Visel, A., E.M. Rubin, and L.A. Pennacchio, *Genomic views of distant-acting enhancers.* Nature, 2009. **461**(7261): p. 199-205.

68. Deng, J., et al., *Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming.* Nature Biotechnology, 2009. **27**(4): p. 353-360.
69. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nature Biotechnology, 2009. **27**(4): p. 361-368.
70. Schoenfelder, S., et al., *Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells.* Nature Genetics, 2010. **42**(1): p. 53-U71.
71. Li, G., et al., *ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.* Genome Biology, 2010. **11**(2): p. R22.
72. Lieberman-Aiden, E., et al., *Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome.* Science, 2009. **326**(5950): p. 289-293.
73. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).* Nature Genetics, 2006. **38**(11): p. 1348-1354.
74. Friedman, N., et al., *Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells.* Nature Biotechnology, 2011. **29**(5): p. 436-U237.

7. 論文目録

原著論文

Predicting promoter activities of primary human DNA sequences.

Takuma Irie, Sung-Joon Park, Riu Yamashita, Masahide Seki, Tetsushi Yada, Sumio Sugano, Kenta Nakai, Yutaka Suzuki

Nucleic Acids Research. 2011 Jun 1;39(11)

8. 謝辞

本研究の遂行にあたり，修士課程の頃から長年の指導をして頂きました，菅野純夫教授と鈴木穣准教授に御礼申し上げます．常日頃からの温かいご指導のおかげで何とか研究を遂行することができました．心から感謝いたします．

ご多忙の中，本論文の審査をお引き受け下さり，貴重なご意見を下さいました，東京大学大学院新領域創成科学研究科　上田卓也教授，佐藤均准教授，東京大学大学院医科学研究所　中井謙太教授，東京大学大学院理学系研究科　伊藤隆司教授にお礼申し上げます．

ゲノム制御医科学分野　渡邊学助教には研究室のセミナーなどで数多くのご助言を頂き，大変参考になりました．感謝いたします．

京都大学大学院情報学研究科　矢田哲士准教授，東京大学医科学研究所　山下理宇特任助教，朴聖博士にはプロモーターのモデリングなどをはじめとした数学的・情報学的な解析について多大なご助言を頂きました．心から感謝申し上げます．

本研究は研究室のメンバーの助力が無ければ完成しませんでした．
菅野・鈴木研究室卒業生の榊原雄太様，鈴木佐和子様に整備して頂いたプロモータークローンを用いました．金井昭教博士（現　広島大学原爆放射線医科学研究所特任助教）にはRNAポリメラーゼIIのChIP-Seqについて，荒内貴子様にはNucleosome-Seqについてのご助力をいただきました．

菅野研究室の実験補助員の今村聖実様，西澤理絵様にはクローンの配列決定などを手伝って頂きました．心よりお礼申し上げます．メディカルゲノム博士課程の関真秀様にはルシフェラーゼアッセイの条件検討などを手伝って頂きました．

株式会社ダイナコム　若栗浩幸博士，関森悦子様，堀内映実様には情報学的解析でご助力頂きました．

メディカルゲノム専攻同期として共に研究室生活を送った田崎真哉博士，谷本幸介博士（現　東京医科歯科大学　難治疾患研究所　ゲノム解析室　助教），Nuankanya Sathirapongsasuti博士に深く感謝いたします．皆様のこれからのご活躍を心よりお祈り申し上げます．

本研究を行うにあたりお世話になりましたゲノム制御医科学分野の皆様に心から感謝いたします. 充実した研究室生活を送ることができました.

生活費など, 金銭面で援助して頂きました独立行政法人日本学生支援機構に感謝いたします. 私をリサーチアシスタントとして採用して下さいました, COEプログラム「言語から読み解くゲノムと生命システム」とグローバルCOEプログラム「ゲノム情報ビッグバンから読み解く生命圏」に感謝いたします.

最後に, 大学院進学に快く賛成し, 学生生活を支えてくれた両親に感謝いたします.

平成23年8月