博士論文

Mate Pair Library を利用した転写開始点と転写終結点の網羅的な解析

松本　京子

目次

１．序論 4

２．材料と方法 (I)

２．１．細胞培養 6

２．２．RNA 6

３．材料と方法 (II)

３．１．Mate Pair library 作成 7

３．２．ChIP seq 11

３．３．シークエンス 12

３．４．解析 12

４．結果と考察

４．１．TSS-TTS library 作成 16

４．２．選択的 TSC と TTC の評価 22

４．３．Preferred TSC-TTC ペアの同定と評価 29

４．４．転写構造とクロマチン構造の関係 37

４．５．多様な転写領域と融合遺伝子転写産物の同定における

TSC-TTC 情報の応用 44

４．６．転写産物の構造決定について TSC-TTC/Random データの応用 49

５．結語 56

６．謝辞 57

7．参考文献 58

１．序論

ゲノムでの遺伝子領域を決定し、タンパク質をコードする転写産物の正確な構造を決定するために、正確な転写開始点 transcriptional start site (TSS)と転写終結点 transcriptional termination site (TTS)を知ることは重要である。TSS についての正確な情報は、mRNA の 5'端に存在するキャップ構造を合成オリゴに置換、シークエンスタグとして用いる手法、オリゴキャップ法[1,2]や CAGE 法[3-5]が開発され、ゲノムワイドな解析が行われてきた。一方で、TTS についても、EST[6]や近年の RNA seq[7,8]に由来するデータから情報が蓄積している。

しかし、依然として多くの intergenic long non-coding RNA (lncRNA)については TSS や TTS についての情報が不完全である。また、ある一つの転写産物について、その転写領域を決定するための TSS と TTS の組み合わせは、多くの場合、明らかでない[9-12]。ENCODE[13-15]や modENCODE[16]プロジェクトでは、多種多様な生物種、細胞種で主に RNA seq を用いて転写産物のゲノムワイドなカタログ化が行われたが、lncRNA のように一般に TSS 及び TTS がゲノム上の広い範囲に分布する転写産物については RNA seq 法を主軸としたゲノムアノテーションには限界がある。これは RNA seq 法が断片化された転写産物に由来するシークエンスタグ情報を用いるために、転写領域を十分に網羅していても正確に TSS、TTS を同定することが困難であることに起因する。

実際上、転写産物の正確な構造が分からない場合は、重大な問題をはらむことになる。RNA seq のタグを利用して発現量を計算する場合、特定の転写ユニットに RNA seq タグを帰属させる必要があるが、それには転写産物の領域が明確に規定されている必要がある。また、先行研究から多くの遺伝子について選択的プロモーター、選択的転写終結点が存在することが示されているが、断片的な TSS と TTS の情報からは、それらの選択に相互相関があるかについては明らかにならない。さらに、選択的プロモーターに由

来する転写産物についての全長配列に関する情報、たとえばそれらの転写産物が十分なタンパク質コード領域を有しているのか、もしくはプロモーター関連の short RNA のようにタンパク質コード領域を含まない short RNA なのかという疑問も解決されない。これらの疑問を解決する目的には既存の RNA seq 法は不十分である。

本研究では、TSS-TTS mate pair full-length cDNA library (TSS-TTS library)を作成する方法論的開発を行った。また開発された手法を実際にヒトのトランスクリプトーム解析に応用した。本手法で構築される cDNA library は、単一分子の完全長 mRNA に由来する TSS と TTS を環状化させることにより連結し、次世代シークエンサーを使用してこれらを両末端の塩基配列として、解析を可能にするものである。類似の方法による mate-pair library の作成が Ni[17]などにより報告されている。しかし、これは、ハエでのモデルシステムを用いた系に限定した手法であり、ヒトを含む他の生物についての応用例はまだ報告されていないために、その実用性に疑念が残る。また、今回、dT アダプタープライマーの代わりにランダムプライマーで逆転写を行った TSS-Random library の作成も行った。TSS-Random library を作成することにより、TSS と cDNA 内部のエクソン構造との相関の解析を行うものであるが、それは本研究で開発された系に固有の手法である。

本研究では、本手法の有用性を実践するために、14 種類のヒト由来の正常組織と 4 種類の細胞株を使用して、TSS-TTS library、TSS-Random library の作成、その解析を行った。本論文では、本手法を用いて構築された cDNA library の解析が、選択的プロモーター由来の転写産物、lncRNA、多様な転写産物が転写されるゲノム領域についての解析について有用であること、また、がん細胞における融合遺伝子の同定にも有効であることを示す。

２．材料と方法 (I)

２．１．細胞

HeLa (ATCC# CCL-2)、HEK293(ATCC# CRL-1573)、DLD1(ATCC# CCL-221)、MCF7(ATCC# HTB-22)の 4 種類の細胞株を使用した。HeLa 細胞と HEK293 細胞は、15cm シャーレに 10% Fetal bovine serum(FBS)、60mg/l カナマイシンを添加した Dulbecco's modified Eagle's medium(DMEM)培地を使用し、37℃、5% CO_2 条件下で培養した。DLD1 細胞は、15cm シャーレに 10% FBS、60mg/l カナマイシン、4.5g/l D-グルコースを添加した DMEM 培地を使用し、37℃、5% CO_2 条件下で培養した。MCF7 細胞は、15cm シャーレに 10% FBS、0.01 mg/ml human recombinant insulin、60mg/l カナマイシンを添加した phenol red free Minimum Essential Medium(MEM) 培地を使用し、37℃、5% CO_2 条件下で培養した。これらの細胞を使用して、Mate Pair library、ChIP seq library を作成した。

２．２．RNA

4 種類の細胞株から RNeasy(QIAGEN)を使用して total RNA を精製し、library 作成に使用した。14 種類のヒト組織由来の total RNA、adipose (Ambion AM7956), brain (Ambion AM7962), breast (Bio Chain R1234086-50), colon (Bio Chain R1234090-50), heart (Ambion AM7966), kidney (Ambion AM7976), liver (Ambion AM7960), lung (Ambion AM7968), lymph node(Ambion AM7894), ovary (Ambion AM6974), prostate (Ambion AM7988), skeletal muscle (Ambion AM7982), testis (Ambion AM7972), thyroid (Ambion AM6872) を購入し Mate Pair library を作成した。

３．材料と方法(II)

３．１．Mate Pair library 作成

図１．に示すスキームに基づいて、Mate Pair library を作成した。オリゴキャップ法[1,2]を用いて完全長 cDNA library、もしくは 5'端 cDNA library を作成した。それぞれ total RNA 100μg を使用した。Bacterial alkaline phosphatase(BAP)(TaKaRa) 2.5U を使用し、37℃、1 時間処理を行い、不完全長の mRNA の 5'端や、キャップ構造を持たないミトコンドリア由来の mRNA に存在するリン酸基の加水分解を行った。Tobacco Acid Pyrophosphatase(TAP)(epicentre) 40U を使用し、37℃、1 時間処理を行い、完全長 mRNA の 5'端に存在するトリリン酸結合を加水分解し、リン酸基に置換した。T4 RNA ligase(TaKaRa) 500U を使用して、20℃、3 時間処理を行い、リン酸基の残っている完全長 mRNA の 5'端のみに合成オリゴを付加した。使用した合成オリゴの配列は 5'-AGCAUCGAGUCGGCCUUGUUGGCCUACUGGCAGCAG -3' (4ng/μL; custom order, TaKaRa)で、3'端に制限酵素 EcoP15I(NEW ENGLAND BioLabs)の配列を含んでいる。DNaseI(TaKaRa) 10U を使用して、37℃、10 分間処理により、DNA の分解を行った。Oligo-dT セルロースカラム(Cosmo Bio)を使用して、polyA 配列を含む mRNA を精製した。Super ScriptII(Invitrogen) 400U を使用して、42℃、3 時間逆転写を行い、1_{st} strand cDNA を作成した。プライマーの配列は、TSS-TTS library 作成には 5'-GCGGCTGAAGACGGCCTATGTGGCC(T)17-3' (0.4 pmol/μL; custom order, Invitrogen)、TSS-Random library 作成には 3'端に EcoP15I の認識配列を含む 5'-GCGGCTGAAGACGGCCTATGTGGCCCAGCAGNNNNNNC -3' (0.4 pmol/μL; custom order, Invitrogen)を使用した。PCR を用いて DNA を増幅した。使用したプライマーは 5'-primer：5'- AGC ATC GAG TCG GCC TTG TTG -3' (0.16 pmol/μL; custom order, TaKaRa)で 10 番目、16 番目、17 番目のチミジンにビオチンが付加されている、3'-primer：5'- GCG GCT GAA GAC GGC CTA TGT -3' (0.16 pmol/μL; custom order,

Invitrogen)を使用した。この過程で全ての PCR 産物にビオチンを付加した。PCR 産物を 1%アガロースゲルで 100V、30 分電気泳動を行った。TSS-TTS library 作成時には、0.5~5kbp の画分を回収し、QIAquick Gel Extraction Kit(QIAGEN)を使用して DNA の精製を行った。TSS-Random library 作成時には、0.5~1kbp、1~2kbp、2~5kbp の 3 つの画分を回収し、各々精製を行った。以降の Mate Pair library 作成には、各 600ng の DNA を使用した。以降の手順は illumina の MatePair SamplePrep Kit v2 に準じる。ビオチンの付加された cDNA は T4 DNA Polymerase、T4 polynucleotide Kinase、Klenow DNA Polymerase を使用して 20℃、30 分間処理を行い、DNA 両端の突出末端の平滑化を行った。Circularization Ligase を使用して 30℃、16 時間処理を行った。環状化を行い、1 分子内での TSS と TTS もしくは cDNA 内部を結合させた。環状化を行った後、DNA exonuclease を使用して、37℃、20 分間処理を行い、存在する直鎖状 DNA を分解した。EcoP15I 20U を使用して、37℃、1 時間処理を行い、TSS-TTS library では TSS 近傍を、TSS-Random library では TSS と cDNA 内部で切断を行った。TSS-TTS library 作成時には nebulization によりさらに断片化を行った。断片化された DNA から streptavidin beads (Dynabeads M-280, Invitrogen)を使用してビオチンの付加された断片を選択的に回収した。streptavidin beads 上にビオチンの付加された DNA 断片が吸着された状態で、両端の突出末端の平滑化、3'端へのアデニン付加、アダプター付加の工程を行った。両端の突出末端の平滑化には、T4 DNA Polymerase、T4 polynucleotide Kinase、Klenow DNA Polymerase を使用して 20℃、30 分間処理を行った。3'端へのアデニン付加には、0.2mM dATP と A-Tailing Enzyme を使用して 37℃、30 分間処理を行った。アダプター付加には、Adapter Ligase を使用して 20℃、15 分間処理を行った。PCR を用いて増幅反応を行い、8%ポリアクリルアミドゲルで 120V、90 分電気泳動を行った。TSS-TTS library では 280bp 近傍、TSS-Random library では 250bp 付近の DNA 画分を回収、精製を行った。DNA 断片の両端に illumina シークエ

ンス用のアダプターを付加した。取得されたDNA断片を次世代シークエンサーillumina Hiseq2000 を利用して塩基配列決定を行った。

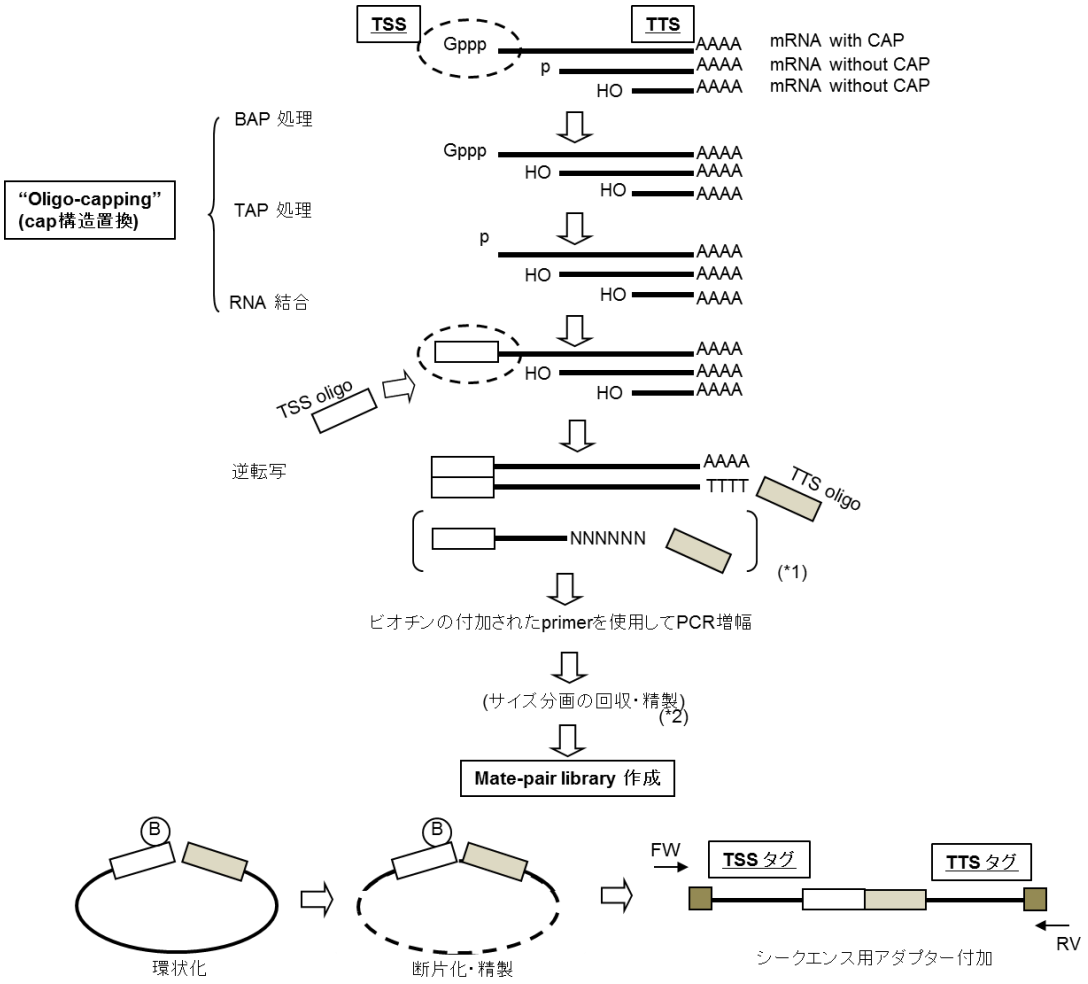

図１．Mate pair library 作成スキーム

本手法は以下の部分課程により構成される：オリゴキャップ法を利用して、完全長 mRNA の 5'端に存在するキャップ構造を、制限酵素 EcoP15I の認識配列を含む合成オリゴと置換する。dT primer もしくは random primer を用いて逆転写を行い、1st strand cDNA を作成する。3 箇所ビオチンの付加された primer を用いて DNA を増幅させる。アガロースゲルを用いて電気泳動を行い、任意の長さの cDNA を回収・精製する。環状化を行い、TSS と TTS、もしくは cDNA 内部を結合させる。断片化を行い、ビオチンの付加された結合箇所を含む断片を精製する。DNA 断片の両端に illumina シークエンス用のアダプターを付加する。

３．２．ChIP seq

4 種類の細胞株、DLD-1、HeLa、HEK293、MCF7 を使用して chromatin immunoprecipitation sequencing (ChIP seq) library 作成を行った。1×10^8 個の細胞を使用した。終濃度 1%のホルムアルデヒド溶液を加え、室温で 10 分間固定を行った。終濃度 166mM グリシン溶液を加え、室温で 5 分間放置し反応を停止させた。1×PBS で 2 回洗浄後、細胞を回収した。細胞に Lysis Buffer1 (50 mM HEPES–KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100)を 5ml 加え、4℃で 10 分間インキュベートを行い、4℃ 、1,500 rpm で 5 分間遠心を行った。上清を捨て、Lysis Buffer2(10 mM Tris–HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA)を 5ml 加え、4℃ 、1,500 rpm で 5 分間遠心を行った。上清を捨て、Lysis Buffer3 (10 mM Tris–HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine) 1ml を加え、ソニケーター(TOMY SEIKO)を使用して氷上で 30 秒、16 回超音波破砕を行った。10% Triton-X 100 を 100µl 加えて 14000rpm で 10 分間遠心を行った。上清のうち 50µl をコントロール(whole cell extract (WCE) DNA)として保存した。各抗体 10µg を磁気ビーズに吸着させ、洗浄を行った。使用した抗体は、monoclonal anti-RNA polymerase II CTD repeat YSPTSPS antibody(pol II) (Abcam, ab817), monoclonal anti- Histone H3 tri methyl Lysine 4(H3K4me3) antibody (Abcam, ab1012), monoclonal anti- Histone H3 mono methyl Lysine 4 (H3K4me1) antibody (Abcam, ab8895), monoclonal anti- Histone H3 tri methyl Lysine 27 (H3K27me3) antibody (Abcam, ab6002), polyclonal anti- Histone H3 acetyl Lysine 27 (H3K27ac) antibody (Abcam, ab4729), polyclonal anti- Histone H3 tri methyl Lysine 36 (K36me3) antibody (Abcam, ab9050), polyclonal anti-RNA polymerase II CTD repeat YSPTSPS phosphorylated serine 2 (CTD-PS2) antibody (Abcam, ab5095), polyclonal anti- CCCTC binding factor (CTCF) antibody (Millipore

07-729), polyclonal anti-Rad21 antibody (Abcam, ab992)である。磁気ビーズに上清の残りを加え、4℃で一晩ローテーターで回転させた。Wash Buffer(50 mM HEPES–KOH, pH 7.5, 500 mM LiCl, 1 mM EDTA, 1% NP-40, 0.7% Na-deoxycholate) 1ml で 8 回洗浄を行い、50mM NaCl を含む TE Buffer で 1 回洗浄を行った。200µl の elution buffer (1 M Tris–HCl, pH 8.0, 0.5 M EDTA, pH 8.0, 1% SDS)を加え、65℃、15 分間インキュベートを行い、DNA の溶出を行った(ChIP DNA)。磁気ビーズを使用して上清を回収し、65℃で一晩インキュベートを行い、脱クロスリンクを行った。保存しておいた WCE DNA に 150µl の elution buffer を加え、同様に 65℃で一晩インキュベートを行い、脱クロスリンクを行った。ChIP DNA、WCE DNA 両方に 200µl の TE buffer と 8 µl の 10 mg/ml RNase A (フナコシ)を 8µl 加え、37℃、2 時間インキュベートを行った。20mg/ml proteinase K (Takara)4µl、300mM $CaCl_2$ 7µl を加え、50 分間インキュベートを行った。フェノールクロロホルム抽出を行い、エタノール沈降を行った。サンプルは Hiseq(illumina)で塩基配列を決定した。DLD1 細胞については、ChIP seq の結果の一部を、[18]のデータを参照した。

３．３．シークエンス

塩基配列の決定は、illumina Hiseq2000 を利用して行った。サンプル調整は付属のプロトコールに従って行った。Mate Pair library については、1 つの library につき、両端から 101bp ずつ、1,000 万タグ以上の塩基配列を取得した。ChIP library については、1 つの library につき、片側から 36bp、平均 3400 万タグ以上の塩基配列を取得した。

３．４．解析

すべての TSS-TTS library、TSS-Random library において、TSS を決定する際に、cDNA の 5'端に隣接する塩基配列 CTGCTGCC を標識配列として使用した。TSS-TTS library において、TTS を決定する際には、TSS が確認されたタグのペアのうち、ゲノム上にマッピングされた Fw の座標が Rv の座標よりも大きい、という条件に当てはまるものを選択した。TSS-Random libray の 3'端を決定する際に、cDNA の 3'端内部断片に隣接する塩基配列 CTGCTGGG を標識配列として使用した。同定された TSS タグと TTS タグをそれぞれ独立に、500bp の bin でクラスタリングを行った。同定したクラスターをヒトゲノム参照配列 RefSeq 遺伝子(UCSC Genome Browser; hg19; http://genome.ucsc.edu/; NM ： タンパク質コード遺伝子、NR ： ノンコーディング遺伝子)にマッピングを行った。TSC の場合は代表転写産物の 5'端上流 50kb 以内、TTC の場合は代表転写産物の 3'端下流 50kb 以内にマッピングされた場合、その転写産物に属するとみなした。RefSeq 転写産物の内部エクソンにマッピングされた TSC、TTC は今回の解析には使用しなかった。TSC と TTC の発現量はシークエンスタグ数を元にして計算を行った。part par million tags (ppm)であらわされる発現量を解析に使用した。[19]の計算方法を参考にした。

組織特異的な発現パターンを解析するために、Z-score を以下の計算式を用いて計算した。

$$z = (x-\mu)/\sigma ;$$

x：タグ数(ppm)の log2、μ：xの平均、σ：xの標準偏差をそれぞれ示す。

TSC や TTC の近傍に位置するシス因子を解析するために、TRANSFAC (version 2011.1; http://www.gene-regulation.com/pub/databases.html)を利用した。TATA box (V$TATA_01, V$TATA_C)を探索するために、minFP のカットオフ値を使用した。CpG island を探索するために、UCSC Genome Browser の情報を利用した。polyA 付加シグ

ナルの探索のために、AATAAA に完全一致する配列を探索した。統計解析は R (http://www.r-project.org/)を使用した。

“preferred” TSC-TTC の相関について検証するために、まず 5ppm 以上、10 タグ以上発現している TSC-TTC ペアを抽出した。ランダム抽出からポワソン分布を仮定して、統計的に有為な偏り、$p<0.05$、があるペアを抽出した。また、ある遺伝子で発現しているすべてのペアのうち、それぞれのクラスターで最も発現頻度の高い TSC-TTC ペアを抽出した。

融合遺伝子を同定するために、4 種類のがん細胞株において、5ppm 以上発現しており、同じ染色体上で 3Mb 以上離れた座標に位置する 2 つの遺伝子、もしくは異なる染色体上に位置する 2 つの遺伝子上に存在する TSC-TTC のペアを抽出した。このさい、マッピングクォリティスコアが 37 であり、RefSeq 遺伝子の転写領域内に TSC、TTC 両方が存在し、既存の RefSeq 遺伝子内での発現量の 5%以上発現している TSC-TTC ペアのみを抽出した。

図 2. に示すスキームに基づいて、取得したシークエンスタグをゲノム上にマッピングすることによりアセンブルを行った。TSC-TTC　library から同定できた転写領域の内部に、TSC を共有する TSS-Random library のペアとなる 3’端タグのマッピングを行い、その結果を用いてアセンブルを行った。マッピングソフトには TopHat v2.0.8b を使用した。既存の RefSeq モデル遺伝子が存在する場合は、エクソン上に 1 タグ以上マップされれば転写領域として同定した。1 つのタグが分割されてマッピングされた場合、その間の領域をイントロンとして同定した。

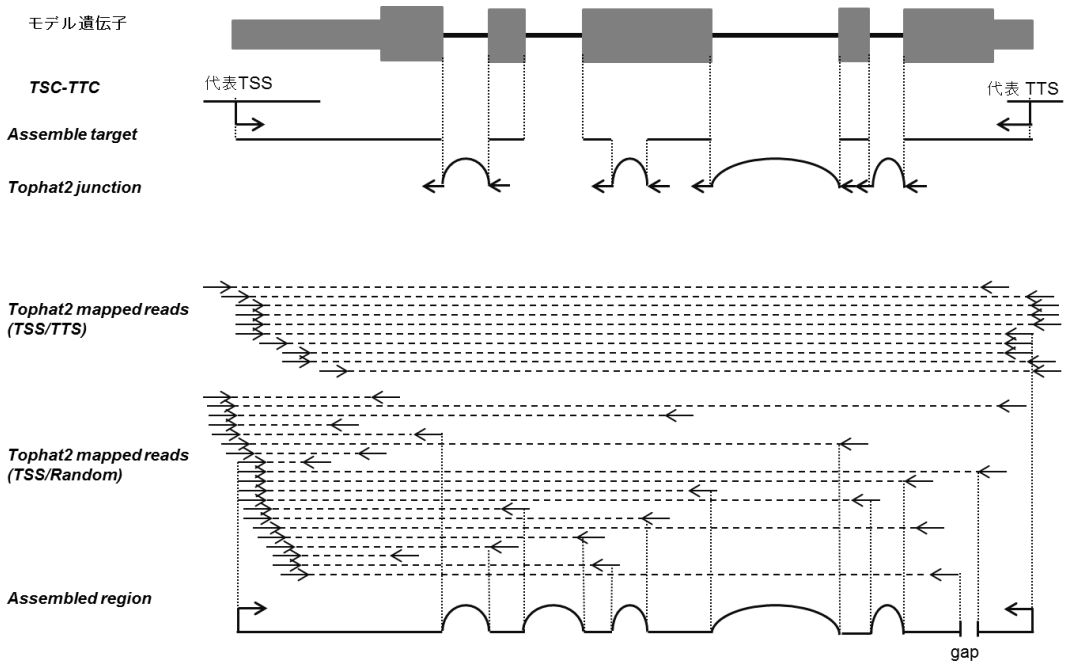

図 2. マップアセンブルスキーム

TopHat v2.0.8b によるマッピング結果を利用してアセンブルを行った。右向きの矢印は TSS タグ、左向きの矢印は TTS もしくは cDNA 内部を示すタグである。TSC と TTC にはさまれたゲノム領域を候補領域とし、転写産物は点線の部分に存在する。

４．結果と考察

４．１．TSS-TTS library 作成

図 1 に示すスキームに基づいて、14 種類のヒト組織由来の RNA と 4 種類のヒト細胞株由来の total RNA100μg を使用して TSS-TTS library の作成を行った。作成した library は次世代シークエンサーillumina Hiseq2000 を使用して塩基配列の決定を行った。101bp の塩基配列を両端から 1,000 万タグ以上を取得した。得られた塩基配列を参照ヒトゲノム配列にマッピングし、RefSeq における遺伝子モデルを指標に遺伝子への対応付けを行った。TSS、TTS の両方についてそれぞれ 500bp の bin でクラスタリングを行い、TSS cluster(TSC)、TTS cluster(TTC)を同定した[19]。シークエンスノイズや、TSC-TTC 間の偽陽性のペアを除くために、5ppm 以上の TSS タグ、TTS タグからなる TSC、TTC のみを解析に使用した。キャップ構造の不正確な置き換えによる不完全長の TSS や、逆転写を行う際に、dT プライマーが mRNA の内部に結合することにより作成される不完全長の cDNA に由来する TTS をこのフィルタリングによって除くことができたと考えている。また環状化の段階でランダムに発生すると考えられる、複数分子の cDNA が結合することに由来する TSS と TTS のキメラについても、同様のフィルタリングにより除くことができたと考えている。

構築された合計 18 の cDNA library から総数 44,902 の TSC-TTC クラスター、平均 8,890 TSC-TTC ペアを同定した。これらのクラスターには、RefSeq 遺伝子 18,808 遺伝子のうち、10,759 遺伝子(25,600 TSC-TTC)が含まれていた。また、既知の 574 lncRNA(818 TSC-TTC)、新規の 5,709 TSC-TTC が同定できた(表 1)。RefSeq 転写産物モデルを元に TSC と TTC の座標のマッピングを行い、それぞれのシークエンスタグのゲノム座標上での妥当性について検証を行った(図 3、図 4)。TSS タグと推定されたシークエンスタグは RefSeq の TSS 近傍にマッピングされ、TTS タグと推定されたシークエ

ンスタグは RefSeq の TTS 近傍にマッピングされた(表 1)。92%の TSC タグは RefSeq 遺伝子の 5'端の上流に位置し、78%の TTC タグは RefSeq 遺伝子の下流に存在していた(図 3、図 4)。これらの結果から本手法により構築された cDNA library を解析することにより、高精度に TSS-TTS 情報を取得することが可能となったと考えられた。1 例を図 5 に示す。

	Library数	取得されたタグ数	RefSeq遺伝子の上流もしくはエクソンにマップされたタグ数(%)	RefSeq遺伝子の下流もしくはラストエクソンにマップされたタグ数(%)	5ppm以上発現していたNM遺伝子数	5ppm以上発現していたNR遺伝子数	複数のTSCを含む遺伝子数	複数のTTCを含む遺伝子数
DLD1	-	8,097,384	95%	74%	7,540	80	243	1,440
HEK293	-	3,367,796	95%	67%	7,768	113	227	1,622
HeLa	-	5,635,548	95%	67%	7,488	91	323	1,550
MCF7	-	5,617,408	94%	64%	7,472	128	396	1,522
Adipose	-	2,414,959	90%	77%	6,076	88	212	1,071
Lung	-	2,533,045	84%	69%	4,545	52	201	759
Ovary	-	3,921,177	93%	81%	5,138	90	247	790
Brain	-	4,068,184	92%	85%	6,440	132	286	948
Breast	-	3,154,960	94%	82%	6,620	101	312	1,145
Colon	-	4,473,789	93%	85%	6,956	88	344	1,029
Heart	-	4,513,913	94%	87%	4,026	56	205	614
Kidney	-	3,378,219	94%	87%	6,074	115	272	948
Liver	-	2,552,292	94%	89%	3,379	36	166	442
Lymph Node	-	3,074,947	94%	88%	4,778	85	247	707
Prostate	-	387,113	92%	72%	2,541	43	83	425
Skeletal Muscle	-	2,430,014	93%	88%	4,647	74	268	690
Testis	-	6,292,637	84%	85%	7,045	384	532	989
Thyroid	-	6,382,769	94%	82%	6,626	106	312	1,155
平均	-	4,211,188	92%	79%	5,983	107	281	1,033
総数	18	80,012,575	91%	76%	10,038	574	2,488	5,096

表 1. TSS-TTS library から取得し解析に利用した TSS タグと TTS タグの統計

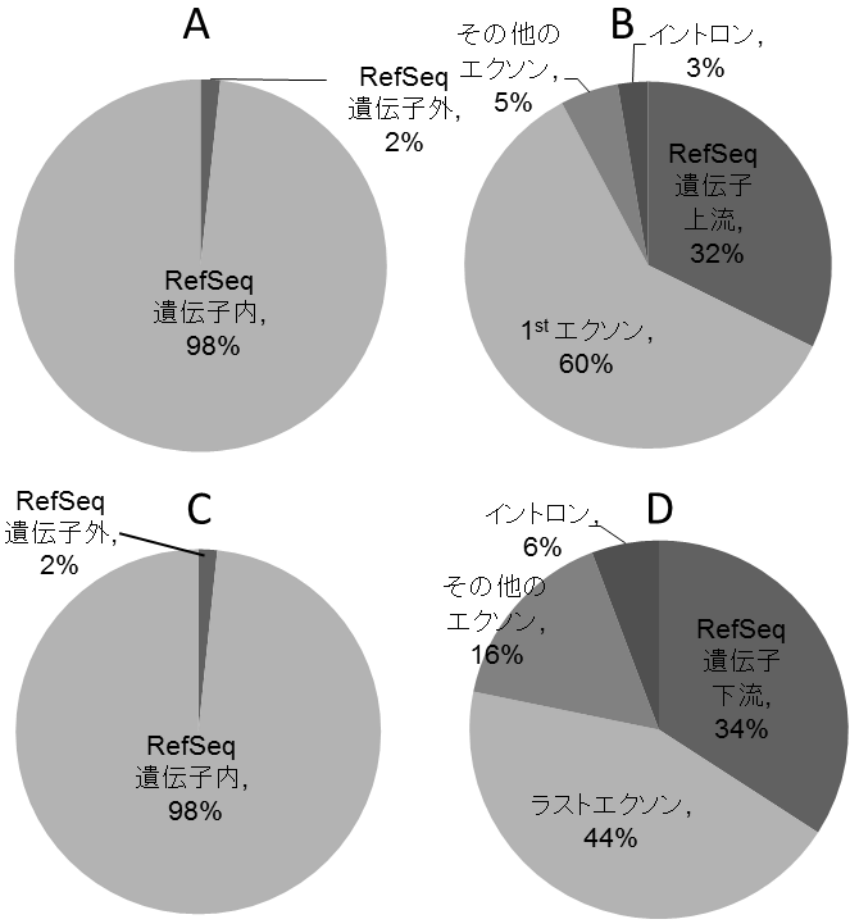

図 3. 取得したタグの分布

A. TSS タグのゲノム上での分布、B. TSS タグの RefSeq 遺伝子上での分布、C. TTS タグのゲノム上での分布、D. TTS タグの RefSeq 遺伝子上での分布

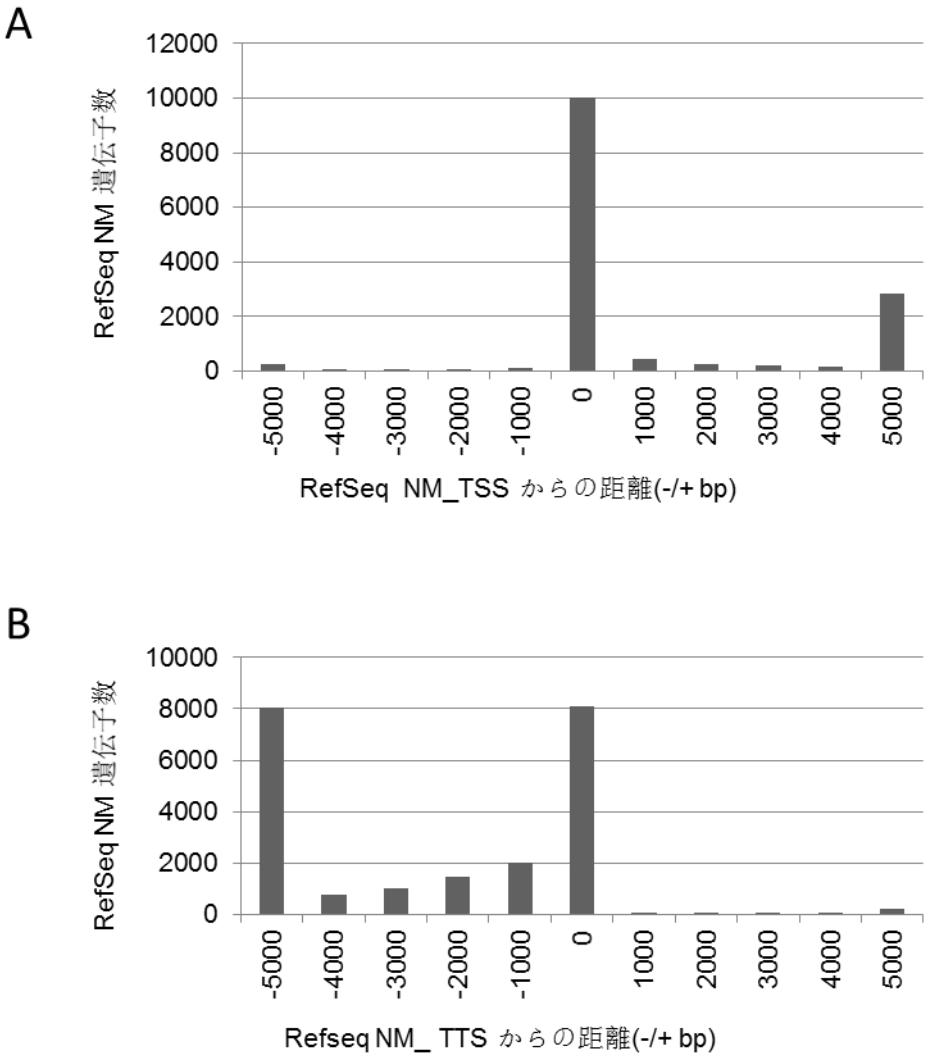

図 4. 取得したタグの RefSeq 遺伝子上での位置

A. RefSeq NM 転写産物の TSS と、TSS-TTS library から取得した TSS タグの距離、

B. RefSeq NM 転写産物の TTS と、TSS-TTS library から取得した TTS タグの距離

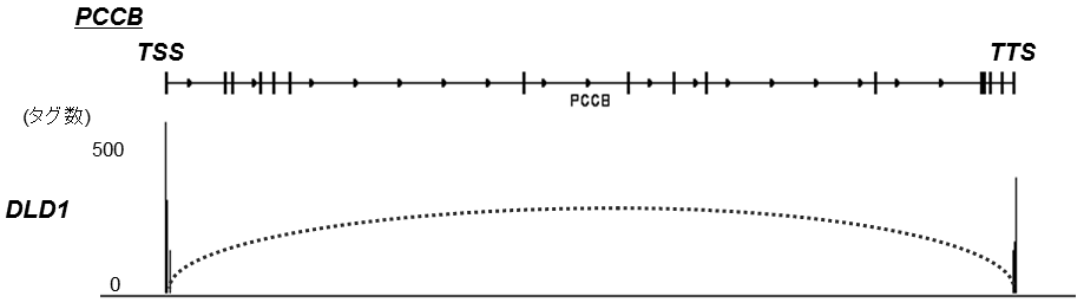

図 5. 取得したクラスターの RefSeq 遺伝子上での分布例

PCCB 遺伝子での TSS タグと TTS タグの分布を示す。左のピークは TSS-TTS library から取得された TSS タグ、右のピークは TTS タグを示している。点線の曲線は TSC-TTC ペアを示している。

４．２．選択的 TSC と TTC の評価

既報の通り[13-15,17]、多数の RefSeq 遺伝子で選択的プロモーターに由来する複数の TSC が確認できた。今回作成した library からは、選択的プロモーターに由来すると考えられる複数の TSC を持つ遺伝子が 2,488 遺伝子、6,944 クラスターが確認された。同様に 5,096 遺伝子において 16,577 クラスターの TTC が確認された(図 6)。図 7. に 1 つの遺伝子内で複数の TSC もしくは TTC が同定された例を示す。複数の TTC を持つ遺伝子が複数の TSC を持つ遺伝子よりも同程度以上存在していることから、選択的 TTC を介する制御が TSC と比較して同じぐらい多様性に富んでいることが示唆された。

TSC 近傍の TATA box や CpG islands[20,21]について複数のクラスター間での差異を検証した。両方の選択的 TSC の近傍に TATA box が存在する例はほぼなく、CpG islands が両方の近傍に存在する例は少なかった（図 8)。一方、polyA 付加シグナルが両方の TTC 近傍に存在する例はまれであった(図 9)。TSC と TTC の組織特異性について検証を行った。TSS-TTS タグカウウントを用いて、Z-score を算出した。Z-score が 2 以上の TSC-TTC を選択したところ、TSC では 16%、TTC では 12%が組織優先的に存在していた(図 10)。TSC と TTC の分布を比較したところ、TSC のほうがより顕著に組織間での偏りがあった(p< 1E-62)(図 11)。転写開始の段階と転写終結の段階で、それぞれに多様な制御が行われていると考えられた。

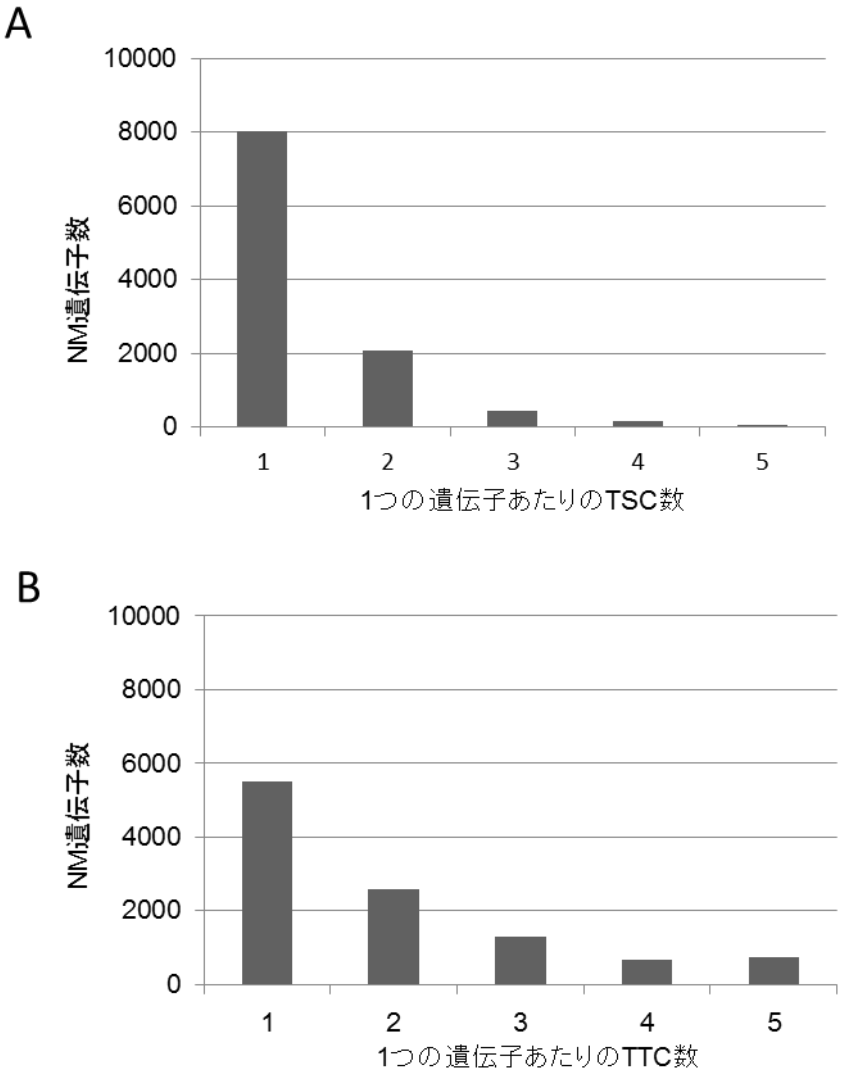

図 6. 複数のクラスターを持つ遺伝子の統計

A. 1 遺伝子あたりの TSC 数、B. 1 遺伝子あたりの TSC 数

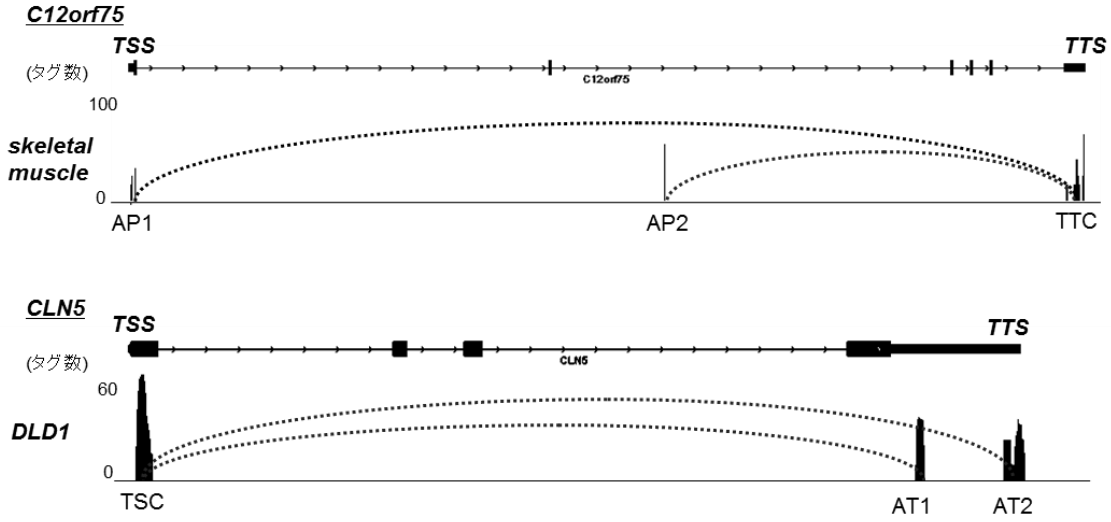

図 7. 一つの遺伝子内で複数の TSC、TTC が存在する例

C12orf75 遺伝子では 2 つの TSC が同定された。CLN5 遺伝子では 2 つの TTC が同定された。AP: alternative promoter region、AT: alternative termination site

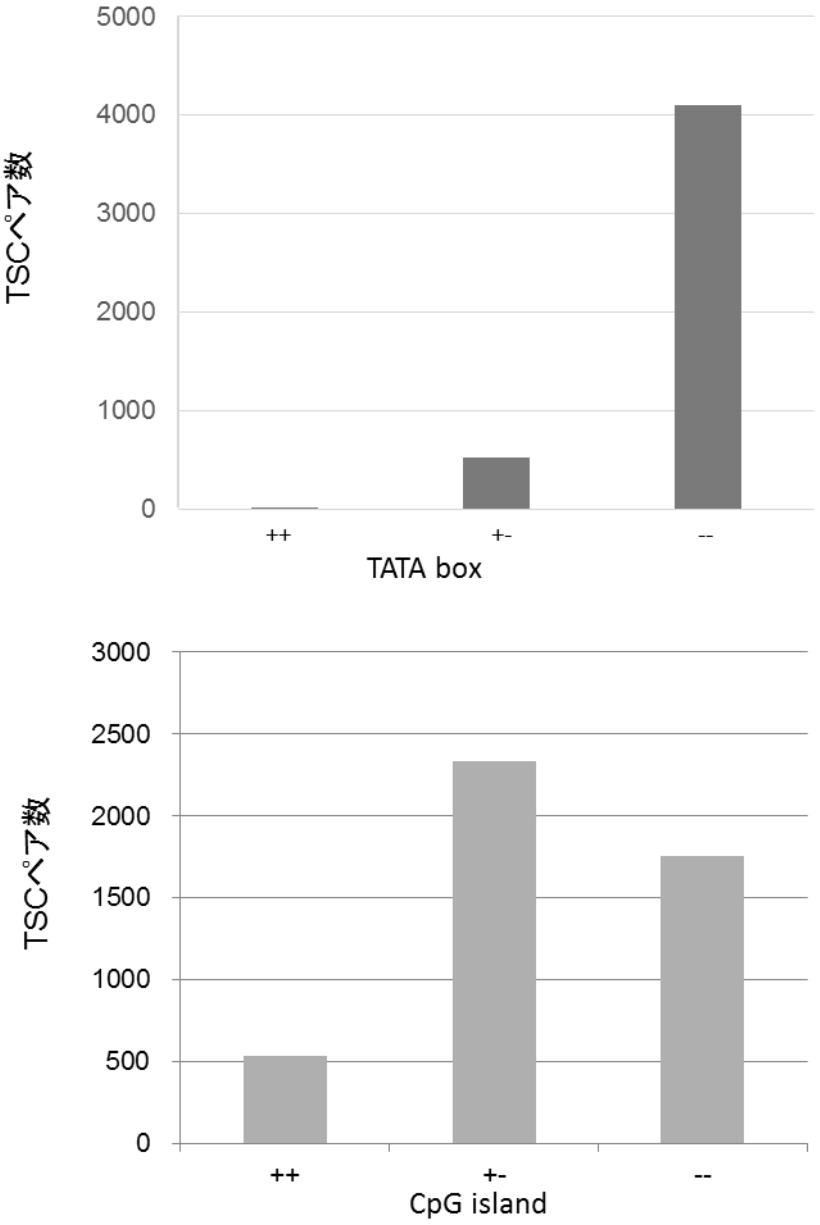

図 8. 1 つの遺伝子内に存在する複数の TSC 近傍のシス因子

A. 1 つの遺伝子内に複数存在する TSC 近傍の TATA box の有無、B. 1 つの遺伝子内に複数存在する TSC 近傍の CpG island の有無を示す。両方の TSC に近傍に存在する場合は++、片方の近傍にのみ存在する場合は+-、両方の近傍に存在しない場合—で示す。

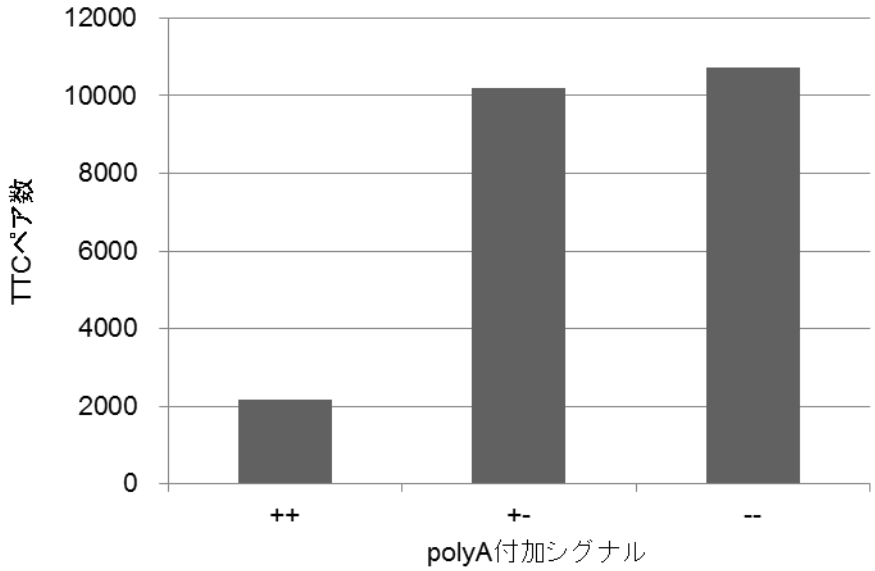

図 9. 1 つの遺伝子内に存在する複数の TTC 近傍の polyA 付加シグナルの分布

1 つの遺伝子内に複数存在する TTC 近傍の polyA 付加シグナルの有無を示している。両方の TTC に近傍に存在する場合は++、片方の近傍にのみ存在する場合は+-、両方の近傍に存在しない場合は—で示す。

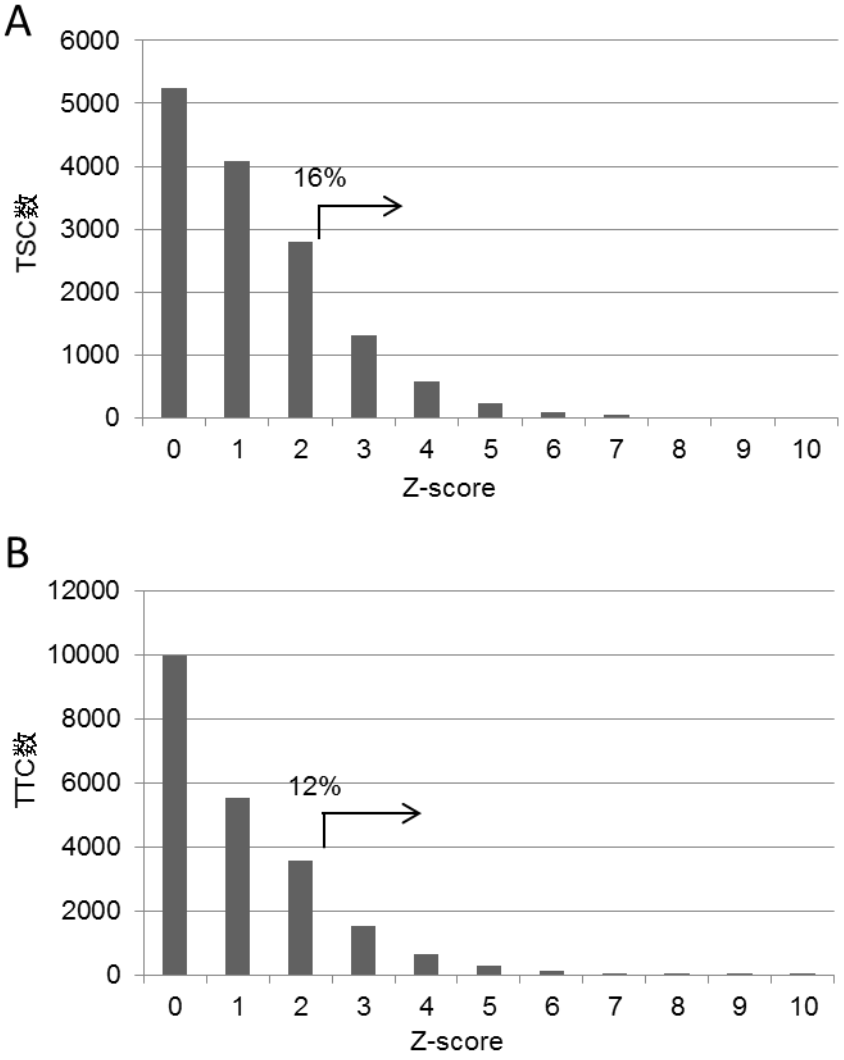

図 10. 同定したクラスターの組織特異性

A. TSC の組織特異性、B. TTC の組織特異性。Z-score は方法に記載した計算式で計算した。

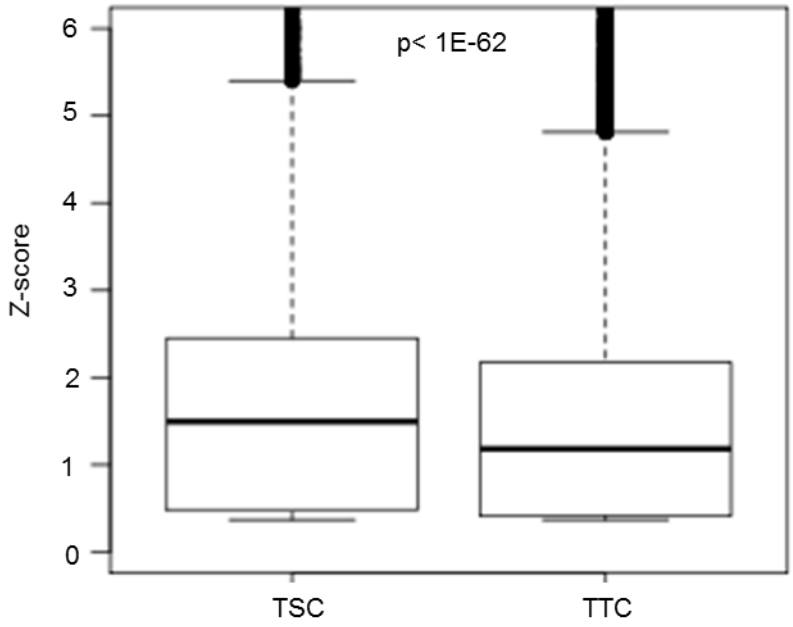

図 11. TSC と TTC の Z-score の分布

Z-score の分布を TSC と TTC で比較した。ウィルコクソン符号順位検定を用いて統計的有意差を計算した。

４．３．Preferred TSC-TTC ペアの同定と評価

一つの遺伝子の中で 2 つ以上の TSC と TTC のペアが存在する場合、TSC と TTC に相関があるかどうかに検証した。TSS-TTS library のそれぞれのタグカウントから、ランダムに選択されることをポワソン分布から仮定して、その帰無仮説として両者の相関を検証した。25,600TSC-TTC のうち 24,833(97%)で、TSC と TTC に統計的に有意な相関($p<0.05$)は見られず、TSC の選択と TTC の選択は独立に行われていることが示唆された。しかしながら、372 遺伝子、767 例で統計的に有為な相関が見られた。これらのペアでは、TSC、TTC に属する 50%以上のタグが特定のペアと連結されていた。以降、“preferred” TSC-TTC と記す。TSC 近傍の TATA box や CpG islands と、TTS 近傍の polyA 付加シグナルの頻度や組み合わせは、全体像と類似していた(図 12,図 13)。

タンパク質をコードする領域が“preferred” TSC-TTC のペアでどの程度異なるかについて解析を行った。“preferred” TSC-TTC はタンパク質コード領域を互いにほとんど共有していなかった。2 つの異なる遺伝子で構成されているかのように、タンパク質コード領域をまったく含んでいない場合も存在した。coding DNA sequences (CDS)を共有している割合は、相関しない TSC-TTC のペアよりも有意に少なかった($p< 3E-78$) （図 14）。図 15 に示すように、PKIA 遺伝子では、TSC-TTC ペア A は TSC-TTC ペア B と完全に離れた位置に存在している。Z-score を計算することにより、2 つのペアの発現パターンの解析を行った。SLC25A27 遺伝子の場合、Unit B は精巣で特異的に発現していたが、Unit A は他の組織で偏りなく発現していた。PKIA 遺伝子の場合、重複しない 2 つのユニットが存在し、それぞれことなる発現パターンを示していた。Unit A は心臓と骨格筋で 4.5 と 3.0 という高い Z-score を示したが、Unit B では高い Z-score が見られなかった（精巣で 0.7）。精巣での Unit A は-2.4、心臓と骨格筋での Unit B の Z-score は 0.3 と 0.2 という結果を示した(図 15)。767 例の TSC-TTC ペアのそれぞれについて Z-score を重ねた(図 16)。図 14 に示した

点線の外部に存在するペアを数えたところ、脳と精巣で特に組織特異的な TSC-TTC のペアの発現が確認された。どのような機能の遺伝子で“preferred” TSC-TTC が使用されているを検証するために、GO term 解析を行った。その結果、GTPase 関連遺伝子が濃縮されていた(p< 8E-06)(表 3)。興味深いことにこれらの TSC-TTC は相互にゲノム上で重複する領域をほとんど持たず、遺伝子内で独立の 2 つのユニットを形成しているように見えた。

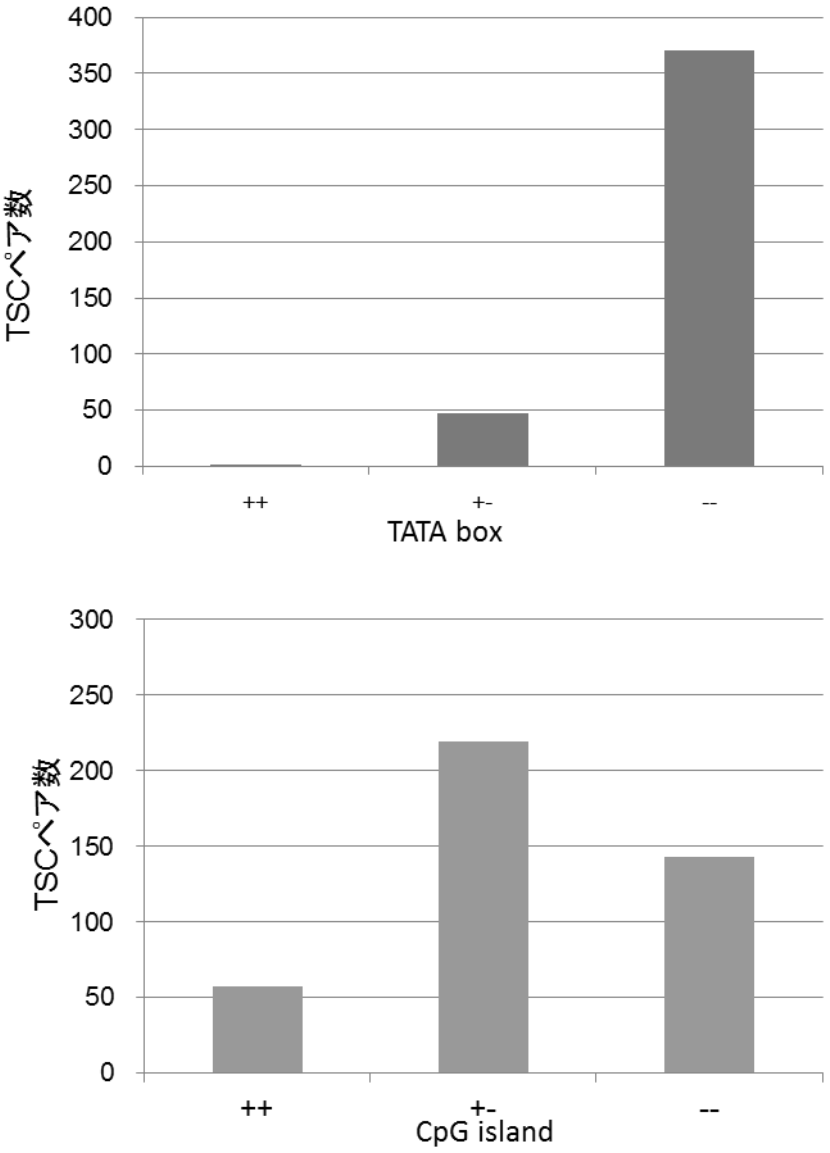

図 12. “preferred” TSC-TTC 間での TSC 近傍のシスエレメント

A. “preferred” TSC-TTC 間での TSC 近傍の TATA box の有無、B. “preferred” TSC-TTC 間での TSC 近傍の CpG island の有無を示す。両方の TSC の近傍に存在する場合は++、片方の近傍にのみ存在する場合は+-、両方の近傍に存在しない場合—で示す。

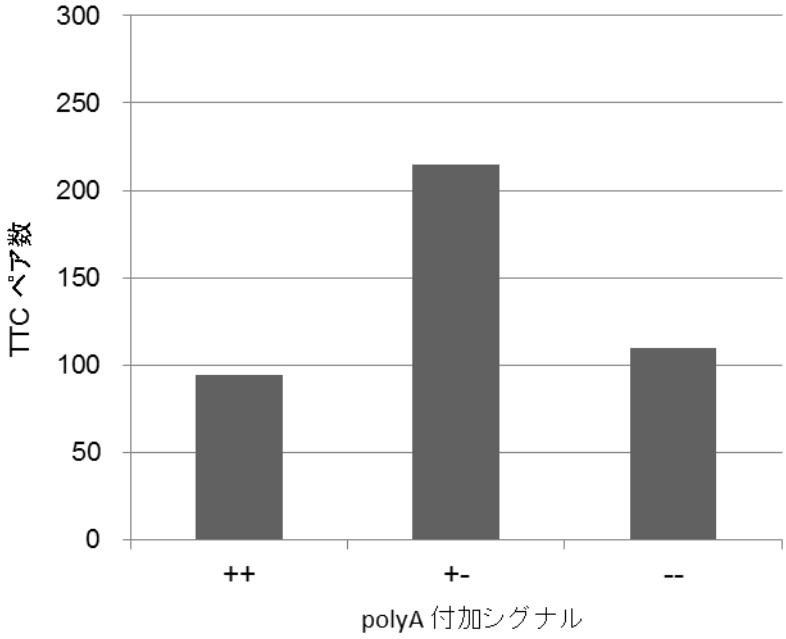

図 13. “preferred” TSC-TTC 間での TTC 近傍の polyA 付加シグナルの分布

“preferred” TSC-TTC 間での TTC 近傍の polyA 付加シグナルの有無を示している。両方の TTC に近傍に存在する場合は++、片方の近傍にのみ存在する場合は+-、両方の近傍に存在しない場合は—で示す。

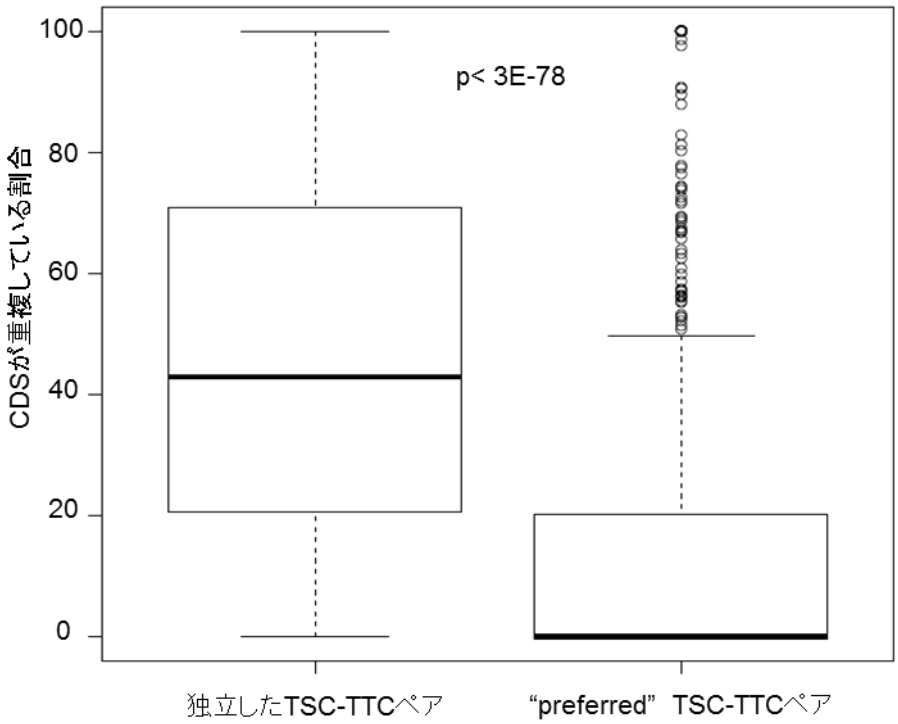

図 14. 1 つの遺伝子内での 2 つの TSC-TTC ペアの CDS 領域の重複

左には独立した TSC-TTC ペア間での CDS 領域の重複、右には”preferred”TSC-TTC ペア間での CDS 領域の重複を示している。ウィルコクソン符号順位検定を用いて統計的有意差を計算した。

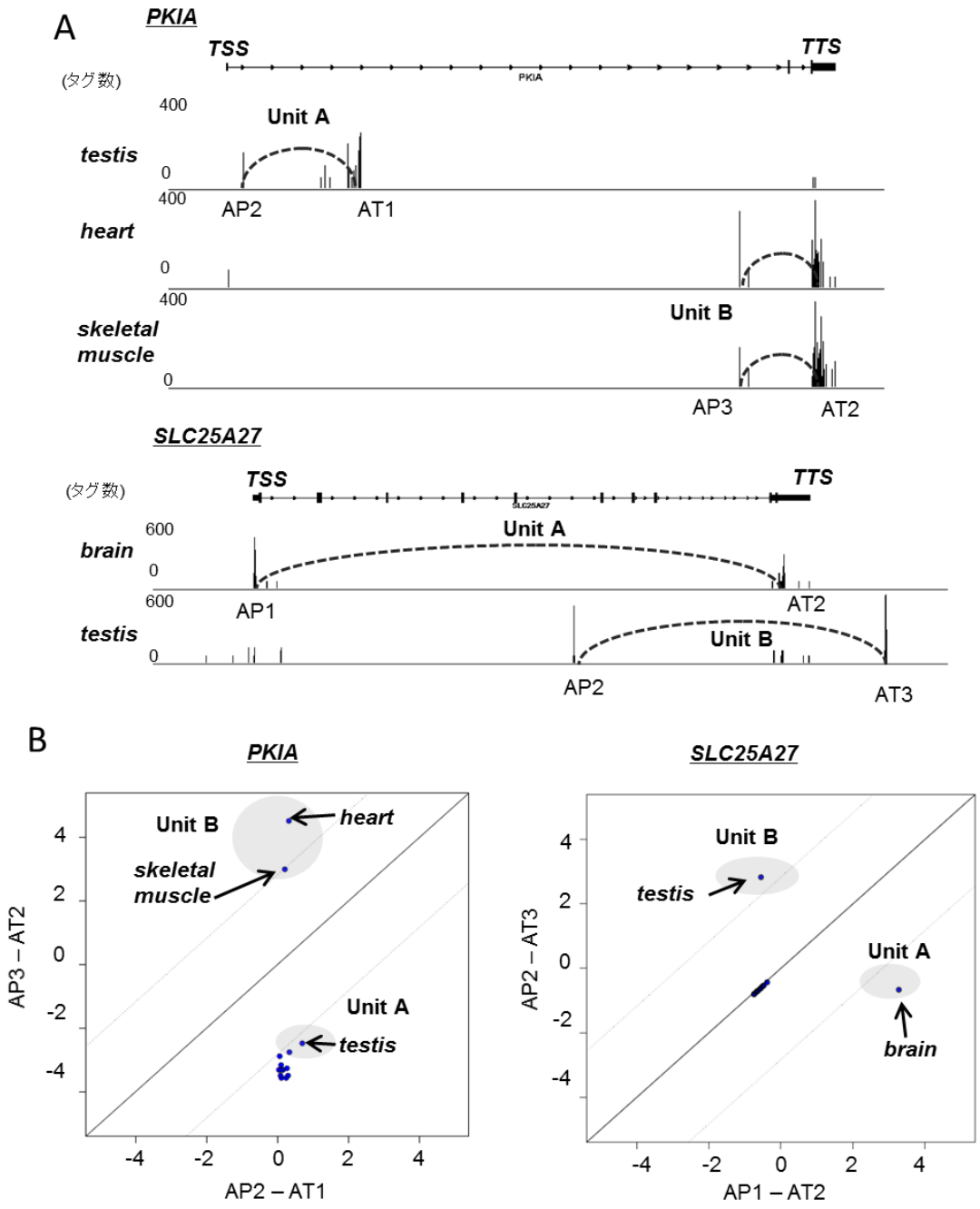

図 15. ”preferred”TSC-TTC ペアの例

A. PKIA 遺伝子と SLC25A27 遺伝子での”preferred”TSC-TTC ペアの例、B. PKIA 遺伝子と SLC25A27 遺伝子でのそれぞれの TSC-TTC Unit の Z-score の分布

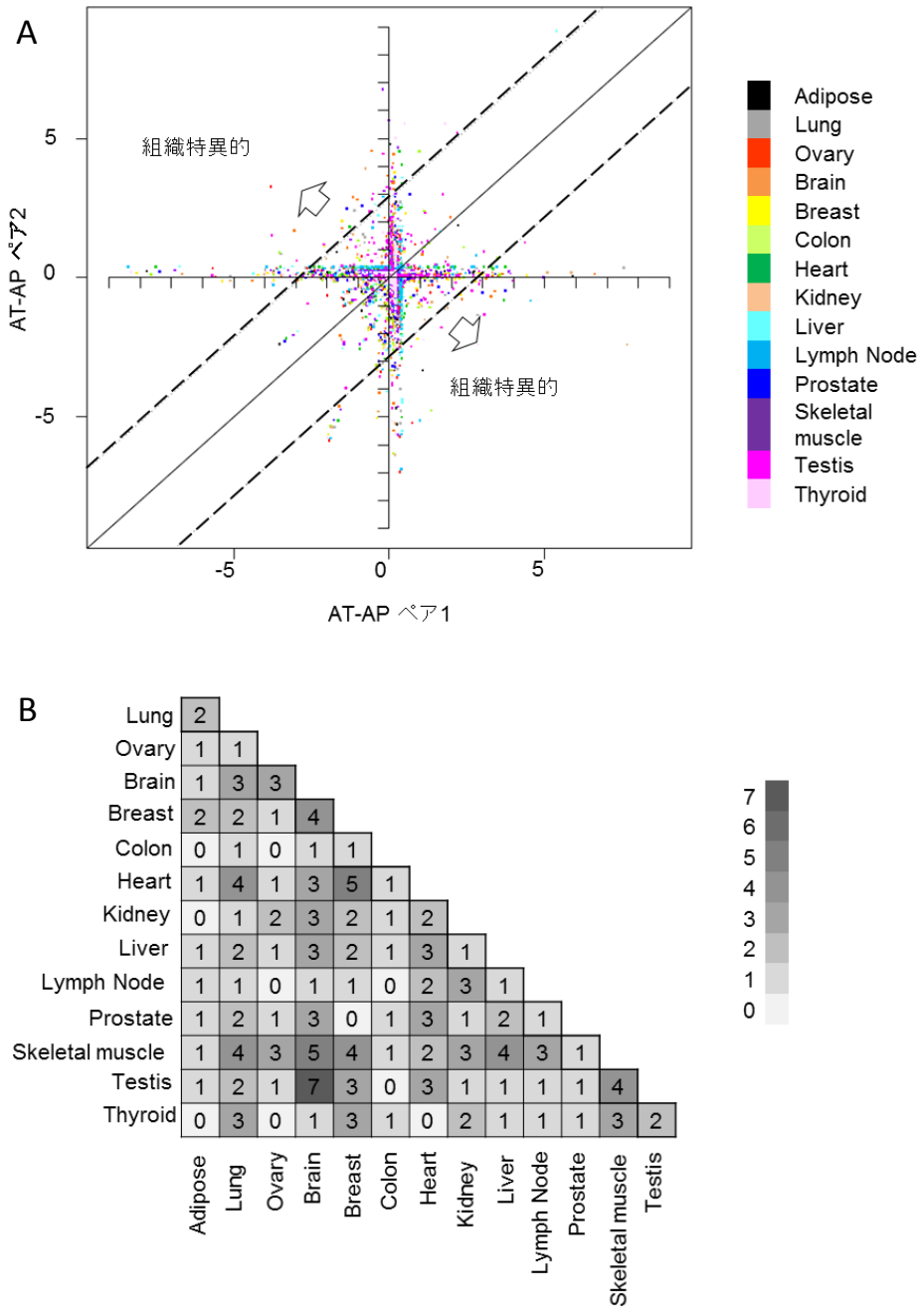

図 16. Z-score 分布

A. 全ての組織における全ての”preferred”TSC-TTC ペアの Z-score、B. Z-score が 2 以上を示す組織特異性の高い”preferred”TSC-TTC ペアの発現数

GO ID	GO term	Number of genes with Preferred TSC-TTC	Number of genes in total population	p-value
0005096	GTPase activator activity	15	195	8E-06
0043547	positive regulation of GTPase activity	12	134	1E-05
0004871	signal transducer activity	13	268	0.002
0045892	negative regulation of transcription, DNA-dependent	17	406	0.003

表 2. ”preferred”TSC-TTC ペアを多く含む GO term

超幾何分布を利用して GO term の濃縮の計算を行った。

４．４．転写構造とクロマチン構造の関係

転写構造とクロマチン状態の関係を検証するために、ヒストン修飾(H3K4me1、H3K4me4、H3K27Ac、H3K27me3、H3K36)、polymeraseⅡ複合体の開始(polII)と伸張(CTD-PS2)、クロマチンインシュレーター複合体の構成部分(Rad21、CTCF)に対する抗体を使用して ChIP seq を 4 種類の細胞株を使用して行った。平均 3400 万シークエンスタグを取得し、解析に使用した(表 3)。MACS の標準のパラメーターを使用してピークのコールを行った。

図 17. に示すように TSC と TTC の間の転写領域について ChIP seq パターンの解析を行った。既知の知見どおり、活性化された転写領域の TSC の上流にエンハンサーマーカーである H3K27Ac のピークが見られた[24,25]。H3K4me3 と Pol II のピークは TSC 上に確認できた[26-30]。H3K36me3 は転写領域で顕著に見られ[31]、H3K27me3 はこれらの領域には存在しなかった[29,32]。CTD-PS2 のピークは TTC の部分に蓄積していた[33,34]。これらのシグナルは一般的に発現量の低い転写産物に対しては顕著ではなかった(図 17)。同様の結果は、用いたすべての細胞株において観察された。これらの結果は、TSC-TTC のデータがクロマチンの特徴と相関していていることを示唆している。ChIP seq データに加えて、もしくは ChIP seq データの代わりに転写アノテーションデータを利用することが可能であると考えている。ただし、TSC-TTC の情報は直接にトランスクリプトーム解析から得られたものであり、高い解像度を持つ。その結果、細胞種における転写領域の正確な同定に有利であると考えている。

“preferred” TSC-TTC とクロマチンの特徴の相関についても同様の解析を行った。これらについても TSC-TTC とクロマチン状態の間に、全 RefSeq の転写産物の場合と類似した相関が観察された(図 18)。“preferred” TSC-TTC が特定の細胞種で実際に転写されている転写産物を表していることを示唆している。一つの遺伝子の中で TSC-TTC のペアのゲノム

の領域を分離し、“preferred”な使用を説明する可能性のある要因を検証した。Rad21[35,36]と CTCF[37,38]の ChIP seq のピークを検証してみると、同じ遺伝子の中で、TSC-TTC のペアがある領域ではそのほかのイントロン領域よりも濃縮されていた(表 4、図 19)。この結果が TSC-TTC 相関の原因なのか、結果なのかは明らかではない。しかし、“preferred” TSC-TTC からなる転写産物は、生物学的ノイズや実験によるアーティファクトではなく、決定論的な方法で制御されていると考えている。

		DLD1	HEK293	HeLa	MCF7
H3K4me1	使用したシークエンスリード数	29,677,167	39,535,295	23,434,100	33,209,524
	MACS で取得したピーク数	91,412	106,309	100,902	78,056
H3K4me3	使用したシークエンスリード数	15,279,900	9,926,213	22,520,220	20,653,434
	MACS で取得したピーク数	16,917	14,172	13,829	15,698
H3K27Ac	使用したシークエンスリード数	9,904,703	32,718,928	62,651,043	31,876,276
	MACS で取得したピーク数	24,332	35,877	56,492	39,646
Pol II	使用したシークエンスリード数	14,471,858	9,926,213	139,395,594	35,905,171
	MACS で取得したピーク数	32,865	25,175	11,395	23,836
H3K36me3	使用したシークエンスリード数	42,143,304	45,931,584	34,927,046	13,828,184
	MACS で取得したピーク数	101,691	58,624	76,265	47,560
CTD-PS2	使用したシークエンスリード数	40,176,827	46,185,168	38,011,086	20,263,552
	MACS で取得したピーク数	24,686	14,379	7,288	9,965
Rad21	使用したシークエンスリード数	49,777,780	55,483,641	39,470,089	7,724,200
	MACS で取得したピーク数	43,108	37,792	77,372	5,634
CTCF	使用したシークエンスリード数	40,743,100	29,032,630	24,818,196	46,715,322
	MACS で取得したピーク数	52,241	50,027	51,401	30,208
H3K27me3	使用したシークエンスリード数	15,754,505	47,865,947	42,484,670	20,988,401
	MACS で取得したピーク数	6,326	97,553	70,776	51,686

表 3.ChIP seq 結果統計

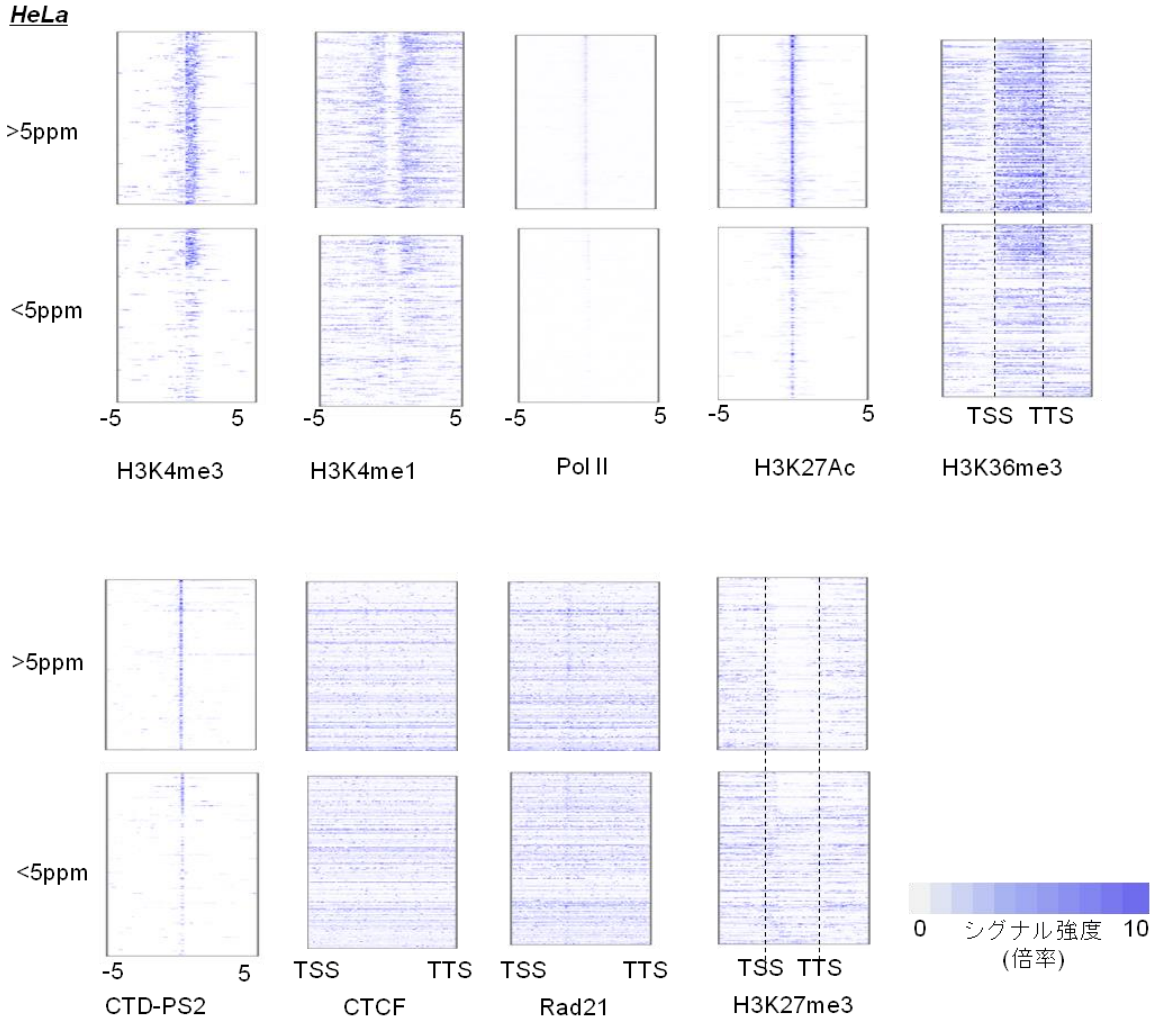

図 17.ChIP seq シグナル強度

HeLa 細胞における ChIP seq のピークのシグナル強度を色で示している。Pol II、H3K4me1、H3K4me3、H3K27Ac については RefSeq 遺伝子の 5'端から 50kb 以内、CTD-PS2 については RefSeq 遺伝子の 3'端から 50kb 以内に存在するピークについて示している。H3K36me3 と H3K27me3 については点線で示した TSC と TTC で囲まれた領域を基準にしている。

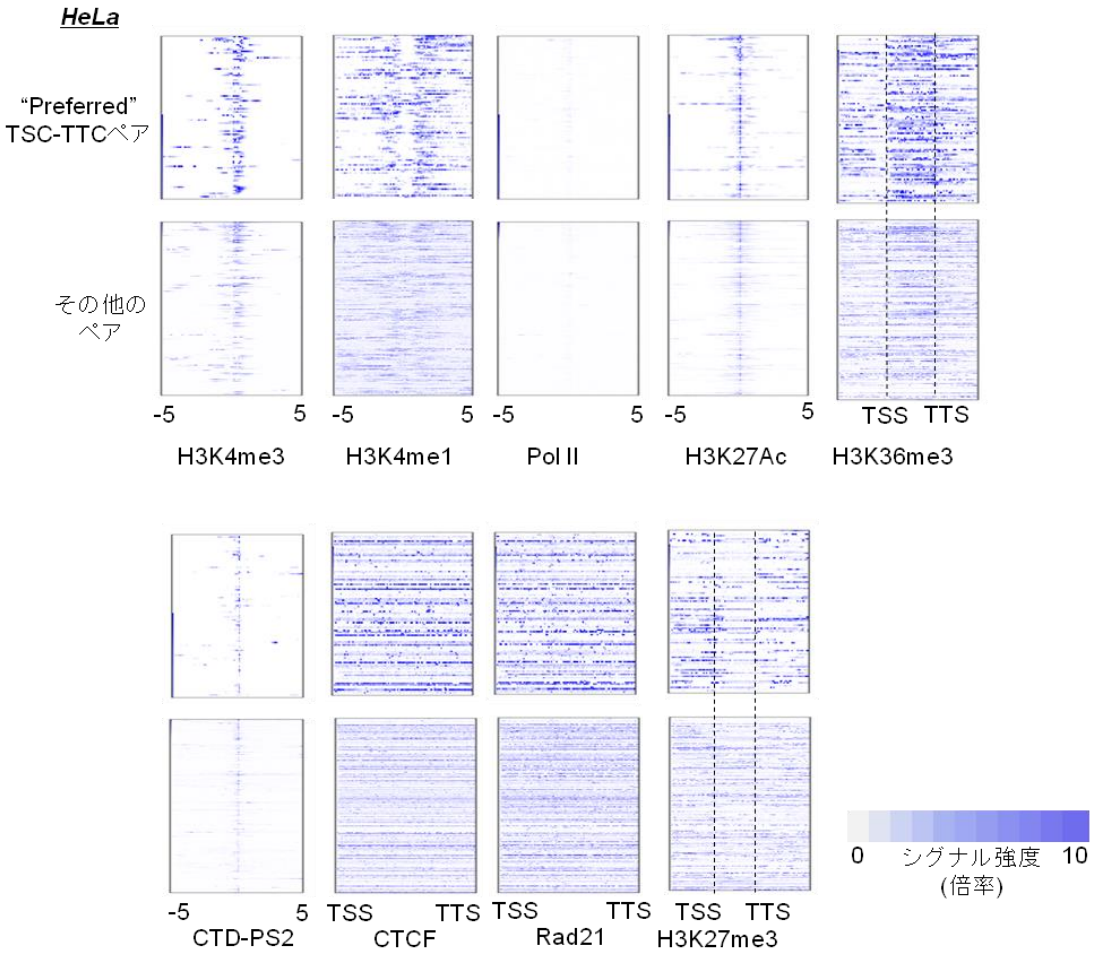

図 18. “preferred” TSC-TTC ペアでの ChIP seq シグナル強度

HeLa 細胞における“preferred” TSC-TTC ペアでの ChIp seq のピークのシグナル強度を色で示している。Pol II、H3K4me1、H3K4me3、H3K27Ac については RefSeq 遺伝子の 5’端から 50kb 以内、CTD-PS2 については RefSeq 遺伝子の 3’端から 50kb 以内に存在するピークについて示している。H3K36me3 と H3K27me3 については点線で示した TSC と TTC で囲まれた領域を基準にしている。

	CTCF	Rad21	CTCF & Rad21
“preferred” TSC-TTCの間	12 / 28 (42%)	14 / 28 (50%)	12 / 28 (42 %)
同じ遺伝子内の他のイントロン上	1 / 28 (4%)	1 / 28 (4%)	1/ 28 (4%)

表 4. CTCF と Rad21 の結合割合

“preferred” TSC-TTC で囲まれた領域での CTCF と Rad21 の ChIP seq でのピーク頻度

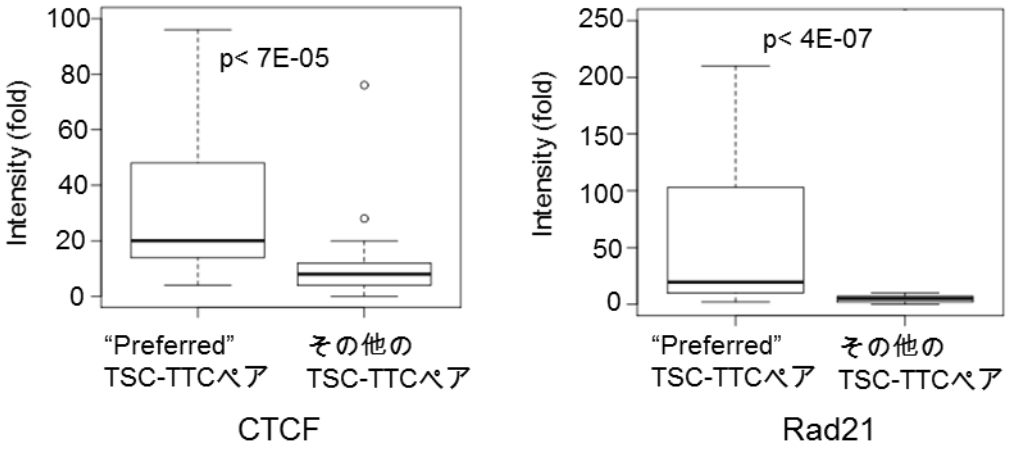

図 19. CTCF と Rad21 の結合状態

“preferred” TSC-TTC ペアで囲まれた領域での CTCF と Rad21 の ChIP seq でのピーク強度と、独立した TSC-TTC ペアで囲まれた領域での CTCF と Rad21 の ChIP seq でのピーク頻度を示している。ウィルコクソン符号順位検定を用いて統計的有意差を計算した。

４．５．多様な転写領域と融合遺伝子転写産物の同定における TSC-TTC 情報の応用

解析を行う中で、隣接しているが異なる RefSeq 遺伝子間で、TSC と TTC が結合されている例が見出された。このような箇所を 174 箇所同定し、クロマチンの状態を検証した。“preferred” TSC-TTC と同様に、2 つの隣接する RefSeq 遺伝子に特徴的なクロマチン状態を同定した(図 20)。H3K36me3 の結合は 2 つの遺伝子の間で高い値を示した。Rad21 と CTCF の結合サイトについても検証を行った。隣接する 2 つの RefSeq 遺伝子が多く転写されている場合、同じ長さの他の遺伝子間領域よりもピークは少なかった（表 5、図 21）。これらの結果から、“preferred” TSC-TTC と類似の遺伝子の転写の多様性が起こっている可能性が示唆された。多様な転写は、遺伝子内だけではなく、遺伝子間でも起こっている可能性があると考えられた。

がん細胞での 2 つの異なる遺伝子の結合とみなされる融合遺伝子の同定について TSC-TTC の情報を応用した。近年、融合遺伝子は様々ながんで起こる染色体異常の結果作成される転写産物であり[39,40]、発癌の引き金となる例が報告されている[41,42]。LC-2/AD 、MCF7、DLD1、HeLa といったがん細胞株で融合転写産物の探索を行った。肺がん細胞株である LC-2/AD では、同じ染色体上の異なるストランド上に 1Mb 以上離れて存在する 2 つの遺伝子、CCDC6 の TSC と RET の TTC が物理的に結合していることが見出された[43]。実際、この融合遺伝子は肺腺がんの患者の 3%で報告されており[44]、抗がん剤の標的として注目されつつある。MCF7 細胞で BCAS4/BCAS3 融合遺伝子についても同定された（図 22）[45-47]。TSC と TTC の領域に primer を設計し、RT-PCR で検証を行ったところ、両方とも融合遺伝子の発現が確認された。これらの結果から、TSC-TTC library の結果から、融合遺伝子の同定・評価を行うことが可能であると考えている。

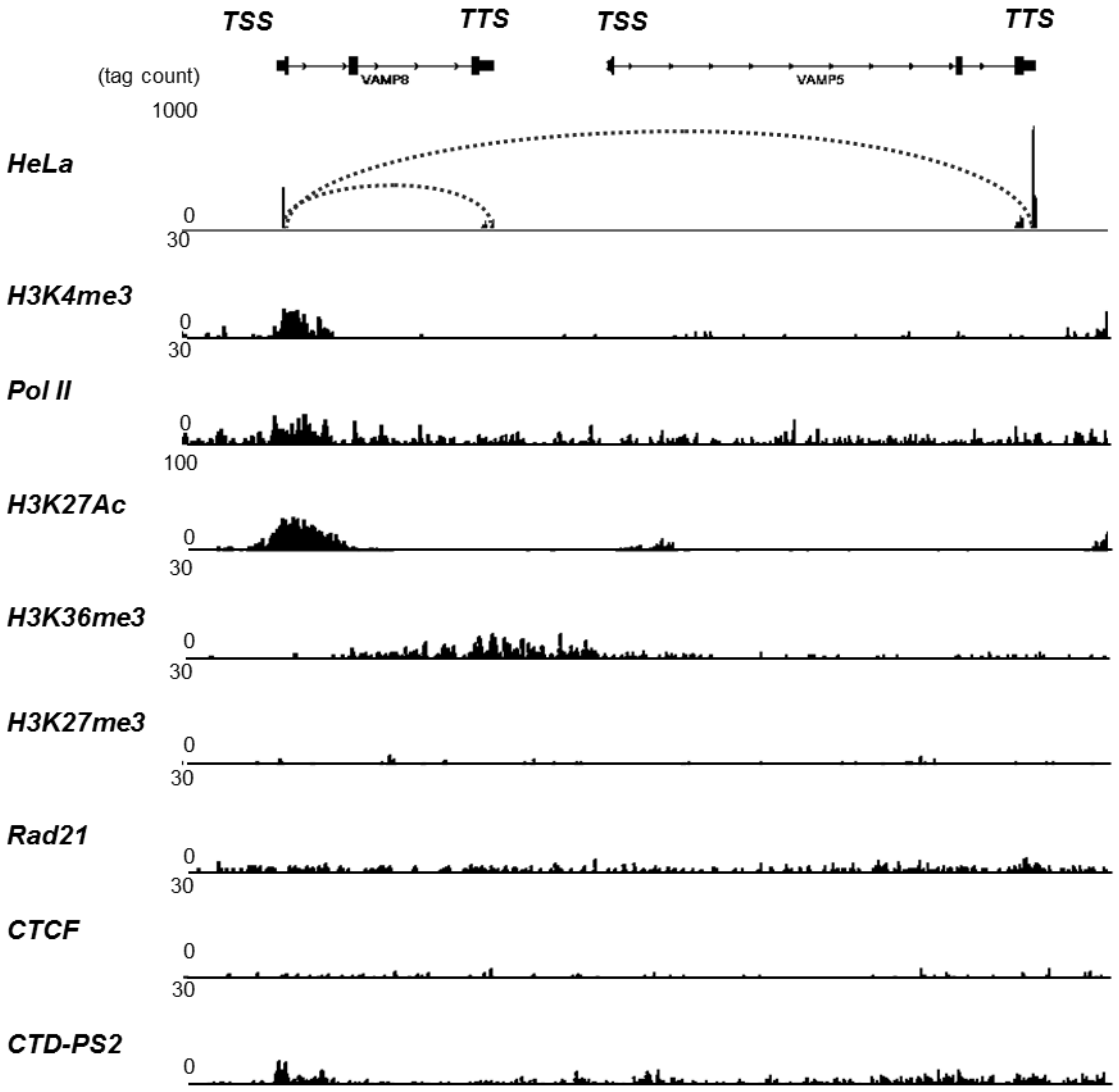

図 20. 隣接する遺伝子での TSC-TTC ペアが確認できる箇所でのクロマチン状態

ピークはそれぞれの library でのタグ数を示している。点線は TSS-TTS library での TSC-TTC ペアを示している。

	CTCF	Rad21	CTCF & Rad21
5ppm以上	3 / 22 (13 %)	6/ 22 (27%)	3 / 22 (13 %)
5ppm未満	46 / 150 (30%)	52 / 150 (34%)	42 / 150 (27%)

表 5. 隣接する遺伝子間での CTCF、Rad21 の ChIP seq ピーク頻度

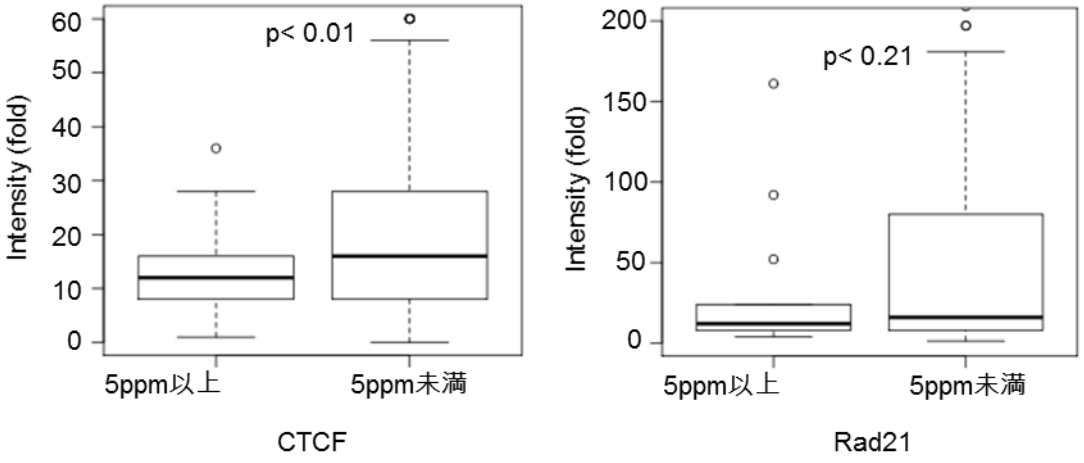

図 21. 隣接する遺伝子間での CTCF、Rad21 の ChIP seq ピーク分布

ウィルコクソン符号順位検定を用いて統計的有意差を計算した。

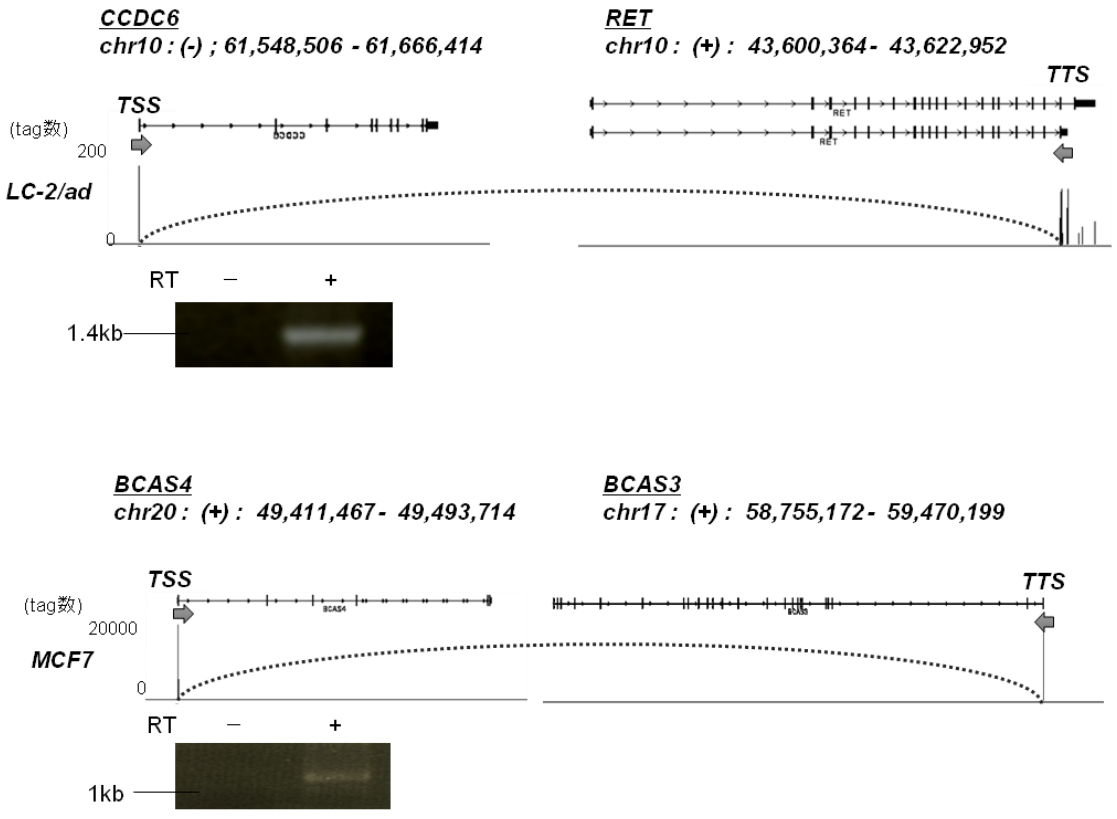

図 22. がん細胞株で同定できた融合遺伝子の例

LC-2/AD 細胞における CCDC6-RET 融合遺伝子と、MCF7 細胞における BCAS4-BCAS3 融合遺伝子の例を示している。それぞれ RT-PCR で検証した。

４．６．転写産物の構造決定について TSC-TTC/Random データの応用

正確な転写産物構造を決定するために TSC-TTC の間の内部構造を明らかにしようと試みた。TSS-TTS library と同様に、同じ出発材料 RNA を使用して TSS-Random library を作成した。ランダムプライマーを付加された二本鎖 cDNA は 3 つのサイズ分画、0.5~1kbp(dR0.5)、1~2kbp(dR1)、2~5kbp(dR2)、から抽出、精製を行い、mate pair library を作成した。解析に使用した TSS-Random タグの統計を表 6.に示す。ゲノムにマップされた座標の統計的解析から、取得された 5'端は、RefSeq 転写モデルの 5'端のエクソンから 500bp 以内に 92%がマップされていることから、TSS を表していると考えられた。RefSeq 転写モデルにマッピングされたタグの TSS と cDNA 内部断片の距離はサイズ分画を行った cDNA の長さに相関していた(図 23)。18 の library から得られた結果を合わせると、塩基レベルで 90%以上がタグによりカバーされる RefSeq 遺伝子が 49%存在し、74%の RefSeq 遺伝子については、少なくとも 1 タグ以上が転写モデルの全てのエクソンにマッピングされていた(図 24)。

TSS-Random library のタグを利用して cDNA 配列のアセンブルを行った。特に、選択的プロモーターに由来する転写産物の構造を別々に再構築を試みた。異なる TSC を分けるのに TSS のタグを使用し、転写産物の構造を決定するのにペアとなっている TSS-Random library のタグを利用した。通常の RNA seq で利用される程度のタグ数が得られなかったために、アセンブリに *de novo RNA* アセンブリアプローチではなく[48-50]、ゲノムベースのアプローチを行った[51,52]。結果、RefSeq の 5'端と重複しない選択的プロモーターに由来する 2,292 の TSC の下流の転写産物が TSC-TTC の間のゲノム領域で 95%以上のカバレージでアセンブルを行うことができた。アセンブルされた転写産物の平均の長さは 1,232bp で、タンパク質をコードしている可能性が十分あると考えられた(表 7、図 25)。

遺伝子間の lncRNA のアセンブルについても同様の方法を使用し、RefSeq の NR に当てはまる lncRNA や、新規の遺伝子間の TSC-TTC で囲まれた領域に存在する lncRNA の構造を決定した。RefSeq の NR 遺伝子のアセンブル成功率と、遺伝子間 TSC-TTC は成功率は類似していた。これらは、RefSeq の NM 遺伝子の場合よりも低かった。これは発現量の低さが原因であると思われる。しかし、363 個の RefSeq の NR と、36 個の遺伝子間の推定上の lncRNA のアセンブリが可能であった(表 7、図 25)。完全なシークエンスが性格に決定されていない選択的プロモーター由来の転写産物や lncRNA の生物学的役割をさらに推察するのにこういった library 作成による塩基配列の決定は必要であると考えている。

	Library 数	取得されたタグ数	NM遺伝子の上流もしくは1stエクソンにマップされたタグ数(%)	90%以上マップされたNM遺伝子数	全てのエクソンがマップされたNM遺伝子数	90%以上マップされたNR遺伝子数	全てのエクソンがマップされたNR遺伝子数
DLD1	0.5 1 2	7,374,699 5,731,468 6,948,401	92%	425	1,053	87	313
HEK293	0.5 1 2	3,976,025 4,094,137 8,511,193	94%	335	1,430	72	353
HeLa	0.5 1 2	14,180,262 5,284,007 8,639,636	93%	442	2,040	67	345
MCF7	0.5 1 2	7,275,826 12,813,315 5,259,460	93%	338	1,728	73	357
Adipose	0.5 1 2	4,460,047 3,723,208 5,081,920	87%	593	1,768	48	277
Lung	0.5 1 2	4,116,414 6,741,772 15,416,440	85%	734	1,490	58	225
Ovary	0.5 1 2	3,522,169 5,027,249 6,962,983	92%	879	1,781	76	361
Brain	0.5 1 2	5,333,608 6,988,354 12,433,092	86%	1732	2,704	93	473
Breast	0.5 1 2	3,561,190 6,728,231 15,849,297	86%	756	2,033	113	447
Colon	0.5 1 2	2,739,746 5,931,356 10,413,600	92%	657	1,826	78	367
Heart	0.5 1 2	5,149,486 5,929,058 8,087,642	90%	983	2,496	50	335
Kidney	0.5 1 2	5,600,064 8,488,442 12,285,999	91%	1193	2,550	66	377
Liver	0.5 1 2	2,038,009 7,049,135 10,612,484	93%	387	1,304	37	183
Lymph Node	0.5 1 2	2,770,236 4,822,871 7,336,121	82%	399	1,511	52	303
Prostate	0.5 1 2	3,386,518 3,498,422 7,484,706	88%	113	786	21	123
Skeletal Muscle	0.5 1 2	3,838,335 4,127,464 6,645.,758	87%	727	1,528	40	237
Testis	0.5 1 2	4,581,230 4,018,607 3,860,240	77%	1086	3,308	123	660
Thyroid	0.5 1 2	2,905,811 4,775,469 4,326,014	92%	282	1,275	27	211
平均	-	6,668,394	89%	647	1,807	66	332
総数	54	254,188,750	93%	9,254	1,2168	497	1,596

表 6. TSS-random library から取得し解析に利用した TSS タグと cDNA 内部タグの統計

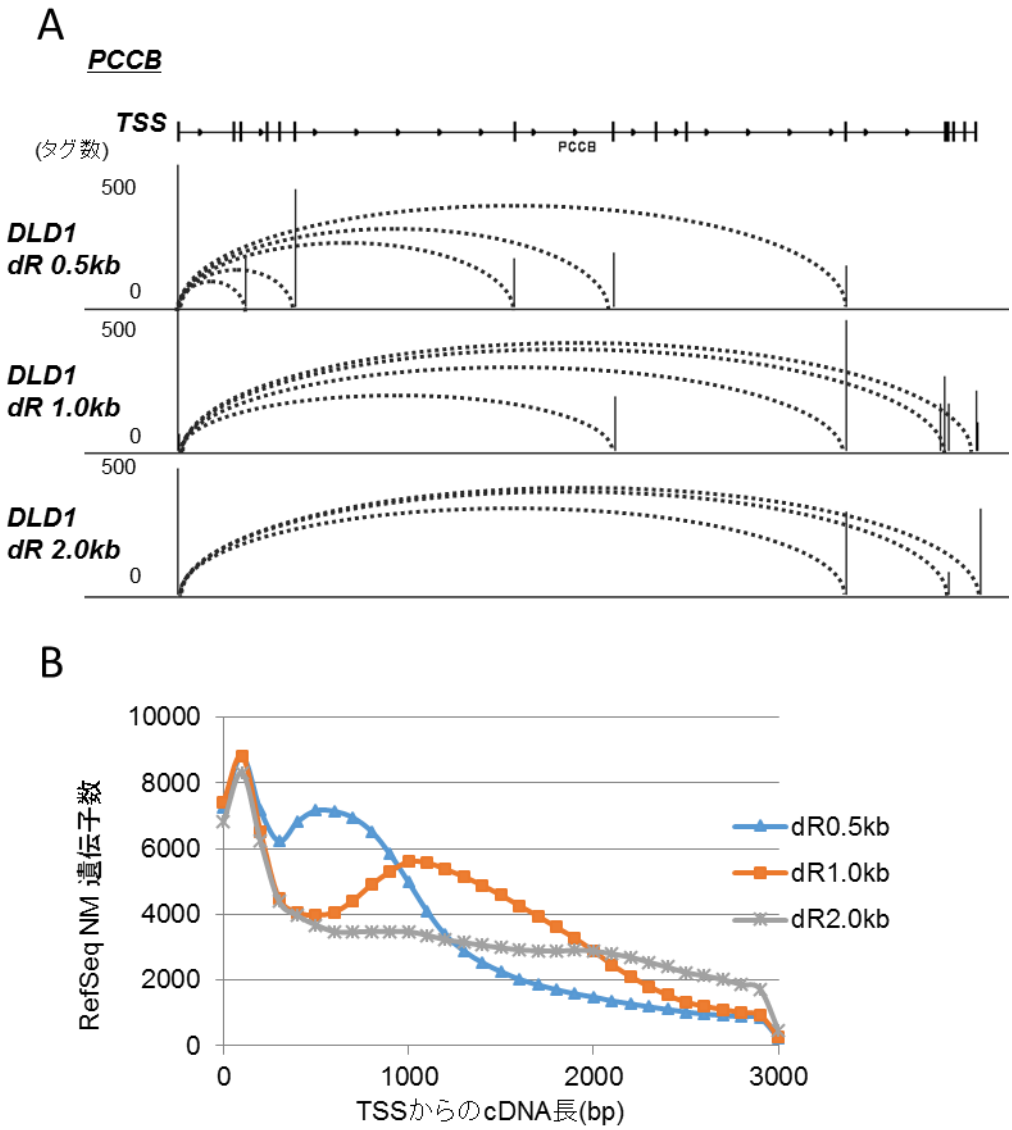

図 23.TSS-Random library での cDNA 内部タグの RefSeq 遺伝子内での分布

A. PCCB 遺伝子における TSS-Random library で同定した TSS-Randam タグの例、B. 3 種類の TSS-Random library の RefSeq 遺伝子での cDNA 長

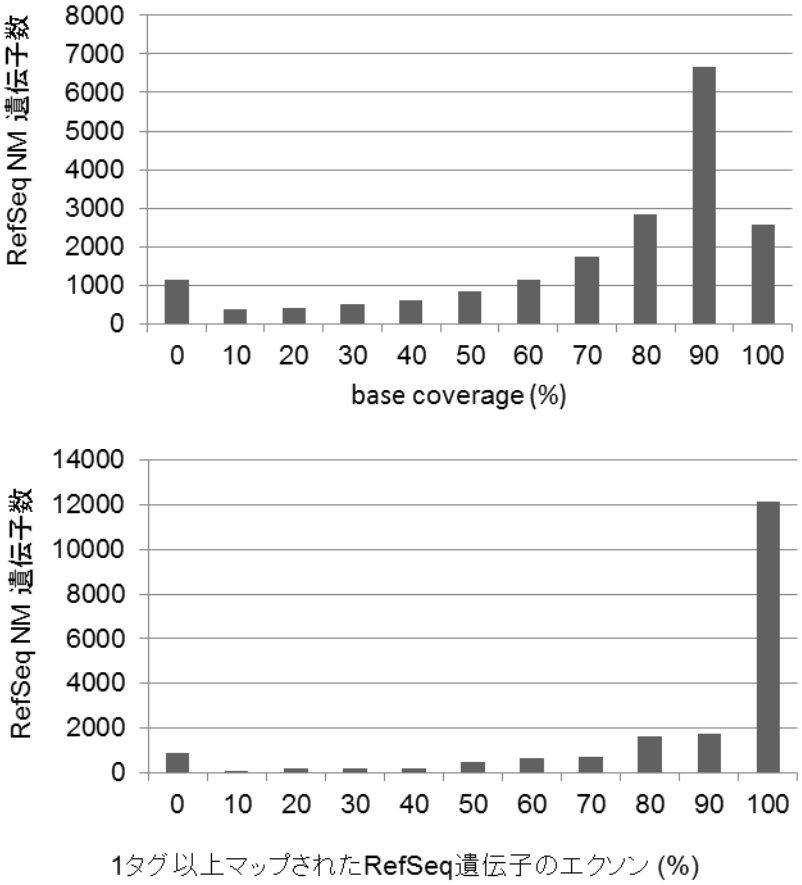

図 24.cDNA 内部タグでカバーできた RefSeq 遺伝子の領域

TSS-Random library から取得できたタグの RefSeq NM 遺伝子に対するカバー率。A. 塩基ごとのカバー率。B. エクソンごとのカバー率。

	アセンブルされたRNA数	アセンブルされたRNAの平均長(bp)
RefSeq遺伝子の選択的プロモーター由来の転写産物	2,193	1,261
NR遺伝子	63	733
Intergenic RNA	36	317

表 7. TSS-Random library タグのアセンブル結果

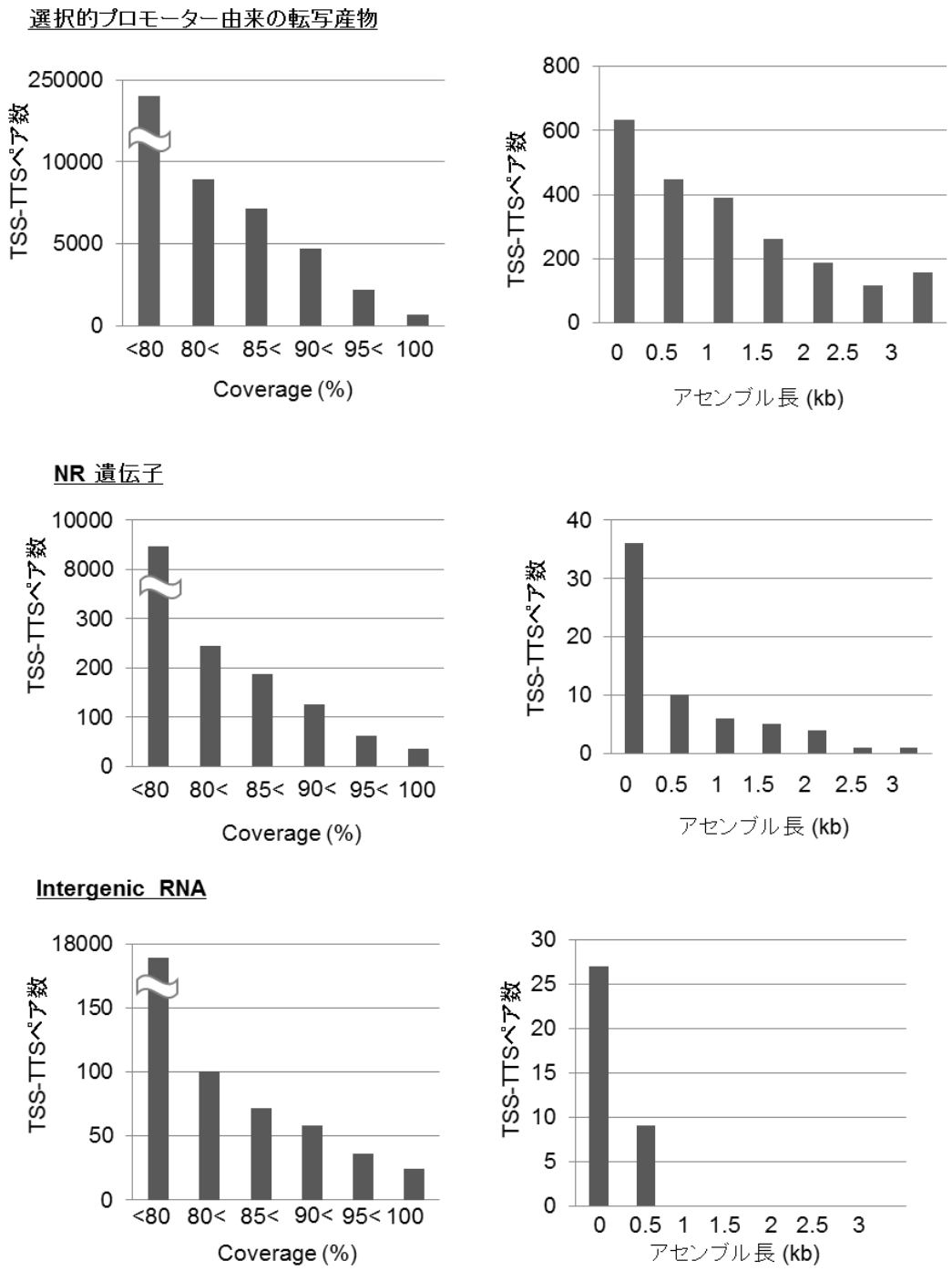

図 25. TSS-Random library タグのアセンブル結果の分布

各種類の RNA について TSS-Random library タグのアセンブル結果を示している。左のグラフはカバー率、右のグラフは 1RNA あたりのアセンブル長を示している。

５．結語

TSS-TTS library と TSS-Random library の作成と評価について論じた。類似の方法論がすでに報告されているが、ヒトのトランスクリプトーム解析に応用したのは本研究が最初の報告である。一連の完全長 cDNA Mate Pair library 作成により、TSC と TTC の相関について検証を行った。TSC と TTC の相関解析から、転写単位が、一つの遺伝子、もしくは異なる遺伝子の結合分子を部分に分けることがあることが見出された。多様な転写産物は完全長 cDNA シークエンス解析で報告されているが、近年の RNA seq からは報告例がない、これは断片化された転写産物を表すだけで情報としては不十分であることに起因すると考えている。

また、今回のトランスクリプトーム解析により、転写モデルの同定が可能であった。この方法論を用いなければ選択的プロモーターに由来する転写産物の構造を決定することは実質的に不可能であった。また lncRNA の転写産物については、一般に発現量が低く、RNA seq は TSS、TTS の決定には不適である。今回の研究はヒトだけではなく他の生物種においてもゲノムのアノテーションに有効である。近年の RNA seq を使用したトランスクリプトーム解析を補足するものとして、完全長 TSS-TTS/Random library は実用性があると考えている。

６．謝辞

本研究は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において行われた。研究、本論文の作成に指導を賜った菅野純夫博士、鈴木穣博士、渡邊学博士に深く感謝いたします。

私の実験の協力、特に塩基配列決定を中心に手伝っていただきました、菅野研究室シークエンスチームの阿部佳澄氏、今村聖実氏、西澤理絵氏、登坂真紀子氏、昆布恵美氏に深く感謝いたします。また解析について貴重な助言をしていただいた堀内映実氏、若栗浩幸に深く感謝いたします。

最後に、研究の遂行にあたり、終始温かく励ましてくれた菅野研究室の方々に心から感謝いたします。

７．参考文献

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

2 Suzuki Y, YoshitomoNakagawa K, Maruyama K, Suyama A, Sugano S. 1997. Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. Gene 200(1-2): 149-156.

3 Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M et al. 2006. CAGE: cap analysis of gene expression. Nature Methods 3(3): 211-222.

4 de Hoon M, Hayashizaki Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques 44(5): 627-+.

5 Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biology 10(7): 21.

6 Muro EM, Herrington R, Janmohamed S, Frelin C, Andrade-Navarro MA, Iscove NN. 2008. Identification of gene 3 ' ends by automated EST cluster analysis. Proceedings of the National Academy of Sciences of the United States of America 105(51): 20286-20290.

7 Fox-Walsh K, Davis-Turak J, Zhou Y, Li HR, Fu XD. 2011. A multiplex RNA-seq strategy to profile poly(A(+)) RNA: Application to analysis of transcription response and 3 ' end formation. Genomics 98(4): 266-271.

8 Hoque M, Ji Z, Zheng DH, Luo WT, Li WC, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3 ' region extraction and deep sequencing. Nature Methods 10(2): 133-139.

9 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Research 22(9): 1775-1789.

10 Rinn JL, Chang HY. 2012. Genome Regulation by Long Noncoding RNAs. In Annual Review of Biochemistry, Vol 81, Vol 81 (ed. RD Kornberg), pp. 145-166. Annual Reviews, Palo Alto.

11 Sun L, Zhang L, Liu H. 2012. Prediction of Long Non-Coding RNAs Based on RNA-Seq. Progress in Biochemistry and Biophysics 39(12): 1156-1166.

12 Ilott NE, Ponting CP. 2013. Predicting long non-coding RNAs using RNA sequencing. Methods 63(1): 50-59.

13 Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F et al. 2012. Landscape of transcription in human cells. Nature 489(7414): 101-108.

14 Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research 22(9): 1760-1774.

15 Sanyal A, Lajoie BR, Jain G, Dekker J. 2012. The long-range interaction landscape of gene promoters. Nature 489(7414): 109-U127.

16 Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF et al. 2010. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science 330(6012): 1787-1797.

17 Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao YA, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nature Methods 7(7): 521-U557.

18 Tanimoto K, Tsuchihara K, Kanai A, Arauchi T, Esumi H, Suzuki Y, Sugano S. 2010. Genome-wide identification and annotation of HIF-1alpha binding sites in two cell lines using massively parallel sequencing. Hugo J 4(1-4): 35-48.

19 Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto S, Sugano S, Nakai K, Suzuki Y. 2011. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res 21(5): 775-789.

20 Smale ST, Kadonaga JT. 2003. The RNA polymerase II core promoter. Annual Review of Biochemistry 72: 449-479.

21 Ponjavic J, Lenhard B, Kai C, Kawai J, Carninci P, Hayashizaki Y, Sandelin A. 2006. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. Genome Biology 7(8).

22 Larsen F, Gundersen G, Lopez R, Prydz H. 1992. CPG ISLANDS AS GENE MARKERS IN THE HUMAN GENOME. Genomics 13(4): 1095-1107.

23 Wang Y, Leung FCC. 2004. An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20(7): 1170-1177.

24 Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America 107(50): 21931-21936.

25 Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470(7333): 279-+.

26 Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karlsson EK, Kulbokas EJ, Gingeras TR et al. 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120(2): 169-181.

27 Kim TH, Barrera LO, Zheng M, Qu CX, Singer MA, Richmond TA, Wu YN, Green RD, Ren B. 2005. A high-resolution map of active promoters in the human genome. Nature 436(7052): 876-880.

28 Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130(1): 77-88.

29 Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K et al. 2006. Control of developmental regulator's by polycomb in human embryonic stem cells. Cell 125(2): 301-313.

30 Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. Nature Genetics 39(12): 1507-1511.

31 Kizer KO, Phatnani HP, Shibata Y, Hall H, Greenleaf AL, Strahl BD. 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3K36 methylation with transcript elongation. Molecular and Cellular Biology 25(8): 3305-3316.

32 Kirmizis A, Bartley SM, Kuzmichev A, Margueron R, Reinberg D, Green R, Farnham PJ. 2004. Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. Genes & Development 18(13): 1592-1605.

33 Komarnitsky P, Cho EJ, Buratowski S. 2000. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes & Development 14(19): 2452-2460.

34 Mayer A, Lidschreiber M, Siebert M, Leike K, Soeding J, Cramer P. 2010. Uniform transitions of the general RNA polymerase II transcription complex. Nature Structural & Molecular Biology 17(10): 1272-+.

35 Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P. 2012. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Research 22(11): 2163-2175.

36 Mehta GD, Kumar R, Srivastava S, Ghosh SK. 2013. Cohesin: Functions beyond sister chromatid cohesion. Febs Letters 587(15): 2299-2312.

37 Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CWH, Ye C, Ping JLH, Mulawadi F et al. 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature Genetics 43(7): 630-U198.

38 Holwerda SJB, de Laat W. 2013. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philosophical Transactions of the Royal Society B-Biological Sciences 368(1620).

39 Froehling S, Doehner H. 2008. Chromosomal abnormalities in cancer. New England Journal of Medicine 359(7): 722-734.

40 Dalmasso C, Broet P. 2011. Detection of chromosomal abnormalities using high resolution arrays in clinical cancer research. Journal of Biomedical Informatics 44(6): 936-942.

41 Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, Fujiwara S-i, Watanabe H, Kurashina K, Hatanaka H et al. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448(7153): 561-U563.

42 Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T, Sakamoto H, Tsuta K, Furuta K, Shimada Y et al. 2012. KIF5B-RET fusions in lung adenocarcinoma. Nature Medicine 18(3): 375-377.

43 Matsubara D, Kanai Y, Ishikawa S, Ohara S, Yoshimoto T, Sakatani T, Oguni S, Tamura T, Kataoka H, Endo S et al. 2012. Identification of CCDC6-RET Fusion in the Human Lung Adenocarcinoma Cell Line, LC-2/ad. Journal of Thoracic Oncology 7(12): 1872-1876.

44 Suzuki M, Makinoshima H, Matsumoto S, Suzuki A, Mimaki S, Matsushima K, Yoh K, Goto K, Suzuki Y, Ishii G et al. 2013. Identification of a lung adenocarcinoma cell line with CCDC6-RET fusion gene and the effect of RET inhibitors in vitro and in vivo. Cancer Science 104(7): 896-903.

45 Barlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, Kallioniemi A. 2002. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes & Cancer 35(4): 311-317.

46 Ruan YJ, Ooi HS, Choo SW, Chiu KP, Zhao XD, Srinivasan KG, Yao F, Choo CY, Liu J, Ariyaratne P et al. 2007. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using paired-end diTags (PETs). Genome Research 17(6): 828-838.

47 Hahn Y, Bera TK, Gehlhaus K, Kirsch IR, Pastan IH, Lee B. 2004. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed

sequence databases. Proceedings of the National Academy of Sciences of the United States of America 101(36): 13257-13261.

48 Martin J, Bruno VM, Fang ZD, Meng XD, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z. 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. Bmc Genomics 11: 8.

49 Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ et al. 2010. De novo assembly and analysis of RNA-seq data. Nature Methods 7(11): 909-U962.

50 Surget-Groba Y, Montoya-Burgos JI. 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Research 20(10): 1432-1440.

51 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7): 621-628.

52 Will S, Yu M, Berger B. 2013. Structure-based whole-genome realignment reveals many novel noncoding RNAs. Genome Research 23(6): 1018-1027.