論文の内容の要旨

論文題目　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

氏　　名　　松本　京子

背景

ゲノム上での遺伝子領域を決定し、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)を同定することは重要である。近年 RNA seq により情報が蓄積しつつあるが、TSS、TTS の情報は必ずしも正確ではない。特に、intergenic long non-coding RNAs (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現在の RNAseq 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 RNAseq 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。また、断片化されたタグ情報からは TSS と TTS の間の相関や、全長配列に関する情報、例えば選択的プロモーター由来の転写産物が十分なタンパク質コード領域を有するかどうか、ということを明らかにすることはできない。本研究課題では TSS、TTS を有する完全長 cDNA を環状化させることにより連結させ、次世代シークエンサーを用いて同時に解析する手法を開発した。また、ランダムプライマーを 1st strand cDNA 合成に使用することにより、完全長 cDNA の TSS と cDNA の内部を連結させる手法を開発し、これらの手法をヒトのトランスクリプトーム解析に応用した。

方法

図 1.に示すスキームに基づいて、14 種類のヒト由来の正常組織と 4 種類の細胞株に由来する

total RNA 100μg を使用して、TSS-TTS library、TSS-Random library を作成した。5'端と 3'端それぞれに合成オリゴと dT アダプタープライマー、もしくはランダムプライマーを付加した mRNA の環状化を行った。TSS-TTS library 作成時には 1~5kbp のサイズ分画から回収、精製を行い、TSS-Random library 作成時には、0.5~1kbp、1~2kbp、2~5kbp の 3 つのサイズ分画から回収、精製を行った。単一分子の mRNA に由来する TSS と TTS、もしくは cDNA 内部配列を結合し、ビオチン・アビジン反応を利用して結合部位の DNA 断片の精製を行った。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。配列解析は両端からの 101 塩基を決定した。得られた配列を参照ヒトゲノム配列 RefSeq にマッピングし、TSS cluster (TSC)、および TTS cluster(TSC)を同定、解析に用いた。

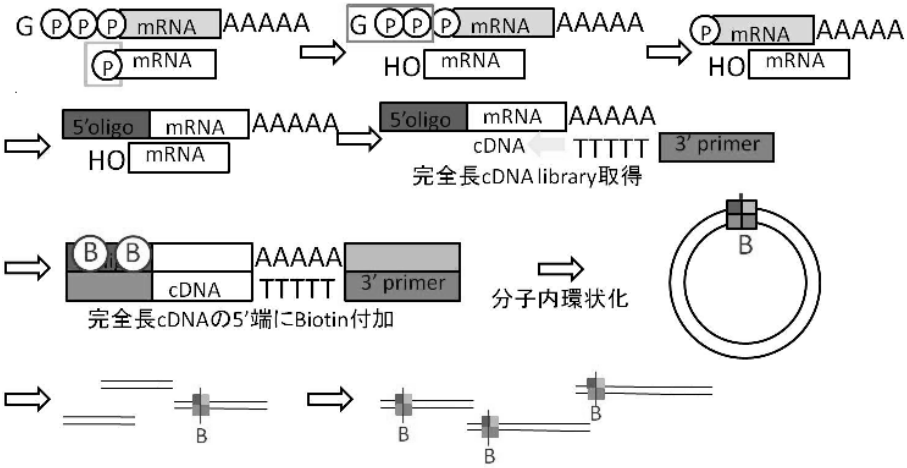

図 1. TSS/TTS library 作成スキーム

オリゴキャッピング法を用いて完全長 cDNA を作成し、分子内環状化を行うことによって TSS と TTS を結合させた。TSS 近傍にビオチンを付加し、Streptavidin beads を用いて TSS と TTS が結合した DNA 断片を回収した。B：ビオチン、P：リン酸基

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(H3K4me、H3K4me4、H3K9Ac、H3K9me3、H3K27Ac、H3K27me3、H3K36)、polymerase Ⅱの開始(polII)と伸張(polSII)、クロマチンインシュレーター複合体の構成体(Rad21、CTCF)に対する抗体を使用して Chromatin immunoprecipitation(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。標準的な ChIP seq 解析プログラム MACS を標準のパラメーターで使用してピークのコールを行った。

結果と考察

得られたタグ配列を用いて正確な TSC、TTC の同定を試みた。92%の TSC タグは RefSeq 遺伝子の 5'端の上流に位置し、79%の TTC タグは RefSeq 遺伝子の下流に存在していた。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(表 1)。図 2A に本手法により同定された TSC

及び TTC の例を示す。18 種類の library から総 44,902 TSC-TTC 、平均 8,890 TSC-TTC を同定した。これらは全 RefSeq 遺伝子 18,808 遺伝子のうち、10,759 遺伝子(25,600 TSC-TTC)に対応していた。また、既知の 574 lncRNA(818 TSC-TTC)、新規に 5,709 TSC-TTC が同定できた。

	library 数	取得したタグ数	TSS の上流または 1st エクソンにマップされたタグ数(%)	TTS の下流または last エクソンにマップされたタグ数(%)
平均	-	4, 211, 188	92%	79%
総数	18	80, 012, 575	91%	76%

表 1. 本研究で取得し解析を行った TSS タグと TTS タグの統計

今回作成した library から取得した結果から、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 個確認できた(図 2B)。同様に 5,096 の遺伝子において複数の TTC が確認できた。クラスターの総数は、TSC は 6944 個、TTC は 16,577 個であった。TSS については TTS のほうがより多様性に富んでいた。近傍に TATA box が両者に存在するものが 60%であった。両者が CpG island 上に存在するのは 11%であった。TSC については両者ともに polyA 付加シグナルを有するものが 9%であった。TSC、TTC いずれにおいても多様なシス因子が認められた。転写開始反応だけでなく転写終結反応においても多様な制御が行われていることが示唆された。実際、16%の TSC 及び 12%の TTC について組織特異的な発現パターンが観測された。

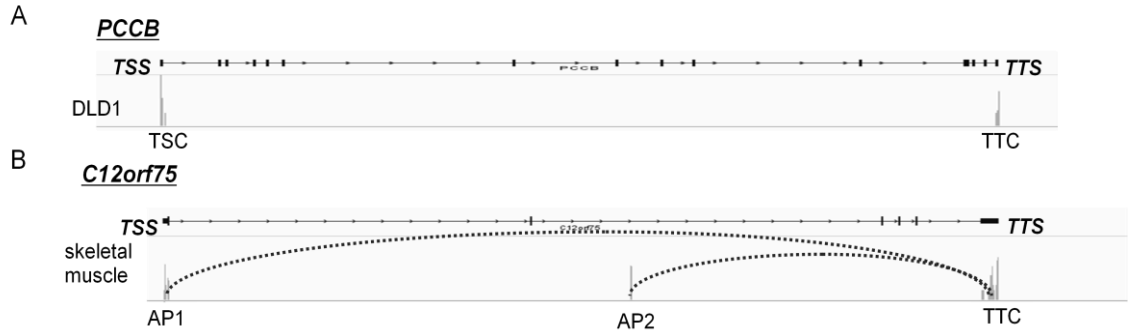

図 2. TSS/TTS library から取得したクラスターの遺伝子上での分布

A. TSS/TTS library で同定した TSC、TTC の例、B. 複数の TSC が確認された遺伝子の例

AP：alternative promoter 由来の TSC

一つの遺伝子の中で 2 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ついて検証を行った。25,600 TSC-TTC のうち 24,833 TSC-TTC (97%)で統計的に有意($p<0.05$)な相関がみられなかった。このことから、TSS の選択と TTS の選択は独立に行われていることが示唆された。例外的に統計的に有意な相関が見られた 372 遺伝子、767 TSC-TTC には GTPase 関連遺伝子が濃縮されていた($p< 8E-06$)。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。ChIPseq 解析の結果から、これらのユニットのそれぞれについて TSS 近傍には HeK4me3 及び polII のピークが、転写領域には H3K36me3 のピークが存在していた。興

味深いことにユニット間の領域に Rad21 及び CTCF のピーク領域が濃縮していた(図 3)。また、隣接した 2 つの遺伝子をまたぐ TSC-TTC が 174 箇所同定され、遺伝子融合転写産物の存在が示唆された。これらの場合にも 1 つの遺伝子内で観察されたようにそれぞれの遺伝子について H3K4me3、polII、H3K36me3 のピークが観察され、融合転写産物が観察されない組織では遺伝子関領域に Rad21 と CTCF が濃縮されていた。遺伝子内、及び遺伝子間において多様な転写産物を生成することにより遺伝子機能の多様化が実現されている可能性が示唆された。

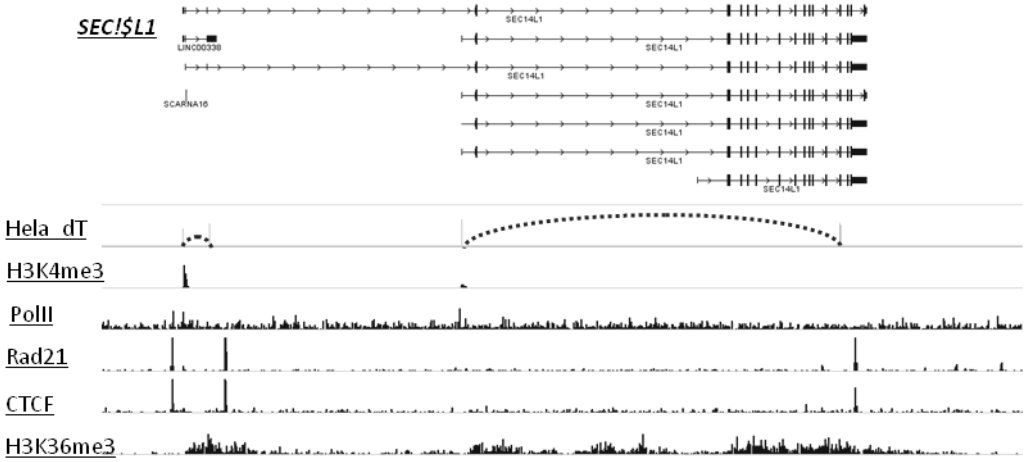

図 3. TSC-TTC とクロマチン構造の関係

TSC と TTC が存在する領域の近傍に Rad21 と CTCF のピークが存在する。

また、本手法を用いてがん細胞でしばしがん原性を有する 2 つの異なる遺伝子間での融合遺伝子転写産物を同定することを試みた。肺がん細胞株である LC2AD での CCDC6/RET 融合遺伝子、MCF7 細胞での BCAS4/BCAS3 融合遺伝子などが同定でき、RT-PCR 法により検証することが可能であった。

TSS/Random library のタグを利用して、選択的プロモーターに由来する転写産物の構造の再構築を試みた。個々の TSC ごとにそのペアとなるランダムタグをアセンブルした。その結果、新規の選択的プロモーターに由来する 2,292 の TSCs 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、十分なタンパク質コード領域を保持していた。lncRNA についても同様の手法でアセンブルを行った。399 個の lncRNA についてアセンブルを行うことが可能であった。

結語

TSS/TTS library と TSS/Random library の作成法を開発した。本法は環状化完全長 cDNA library をヒトのトランスクリプトーム解析に応用した初めての試みである。本手法を用いることにより、多様なヒトのトランスクリプトーム構造を明らかにすることができたことは意義深い。これは近年の断片化された転写産物情報を収集する RNAseq を補完するものとして重要であると考えている。