様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　6 月　6 日現在

機関番号：12601
研究種目：若手研究(B)
研究期間：2009～2010
課題番号：21710210
研究課題名（和文）
大規模 5’端配列を用いた転写制御の解明およびデータベースの構築
研究課題名（英文）　Transcriptional regulation analysis and construction of database based on massive 5’-EST sequences
研究代表者
山下　理宇　（YAMASHITA RIU）
東京大学・医科学研究所・特任助教
研究者番号：10401259

研究成果の概要（和文）：
研究者らは、オリゴキャッピング法によって作られた cDNA に基づき転写開始点が決められている転写開始点データベース DBTSS を構築してきた。本研究では、まず大規模シークエンサーによって配列決定されたデータを約 3 億配列追加することにより、DBTSS の拡充を試みた。また、これらの転写開始点と RNA polymerase・ヒストン修飾の ChIP-seq 情報、MNasaeI によるヌクレオソーム情報、ポリソーム分画による RNA-seq 情報などと比較を行った。この結果、転写開始点付近には特徴的なヌクレオソーム構造、ヒストンの H3K4me3 修飾、RNA polmerase の集積が観察され、その傾向は発現量の多いものほど強かった。以上のことから、本研究で作成したデータベース DBTSS が転写制御解析に非常に有用であることが示唆された。

研究成果の概要（英文）：
We have constructed the DataBase of Transcription Start Sites (DBTSS: http://dbtss.hgc.jp/) which contains the information of accurate transcription start sites (TSSs) based on experimentally determined 5’-end clones based on the oligo-capping method. We updated the database, and now it has about 300 million new short sequences. Based on the DBTSS data, we observed significant characteristics difference in major TSSs: namely, highly ordered nucleosome structures, strong RNA polymerase II binding signals, more frequent translation. These results indicate that our database provides a powerful solution not only to discriminate TSCs having clear biological significance from the other possible noise level transcriptions, but also to analyze transcription regulation.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,700,000	510,000	2,210,000
2010 年度	1,800,000	540,000	2,340,000
年度			
年度			
年度			
総　計	3,500,000	1,050,000	4,550,000

研究分野：情報学・生体生命情報学
科研費の分科・細目：　ゲノム科学・ゲノム情報科学
キーワード：ゲノム、転写制御、プロモータ、データベース

１．研究開始当初の背景

研究者らは、転写制御解析の手助けになるものとして、正確な転写開始点を網羅したデータベース、DataBase of Transcription Start Sites (DBTSS) を構築し公開してきたこれには、実験に基づく cDNA の 5’端 EST 配列をゲノムにマッピングした情報が登録されている。2008 年度のアップデートでは、大規模シークエンサーSOLEXA（Illumina）から得られた 25bp 程度の短い 5’ 端配列を 2 種の細胞(HEK293、MCF7)について約 2000 万配列追加してきた。このような大量の 5’端配列には、２つの大きなメリットがある。一つ目は、転写開始点が分かることであり、もう一つは、一つの mRNA から得られる 5’端配列は一つであるので、配列数を数えることによって、より正確な転写産物間の発現量の比較ができることである。研究開始当時、次世代シークエンサーを使った研究が出始めていた頃であり、RNA を対象とする RNA-seq、またヒストン修飾や転写因子を対象にする ChIP-seq 等に応用されていた。しかしながら、転写開始点をターゲットとした次世代シークエンサーを使った配列の報告は多くなかった。本研究では、次世代シークエンサー由来の主に 5’端 EST 配列を用いて、転写制御機構解明のためのデータベース構築とそれを利用した解析を目的とした。

２．研究の目的

本研究では、次世代シークエンサー由来の大量の 5’端 EST 配列の効率的かつ精度の高いマッピング処理の方法を確立する。配列をマッピング後、転写開始点情報を元に転写産物の転写量とプロモータの同定、遺伝子情報との結合を行い、データベースを作成する。さらに、ヌクレオソーム構造や ChIP-seq のデータとの融合を行い、実際に転写量及び発現組織を決める要因について探索する。

３．研究の方法

本研究では、DBTSS のさらなる拡張とそれを利用したヒトの転写制御情報の網羅的解析を目的とした。具体的には、次の大きく３つの目標を掲げた。１）大量の配列情報の処理方法の確立。大量の 5’EST 配列は、膨大な情報をもたらすが、ある程度のノイズも予想される。それは、配列が短すぎる故、complexity が低くゲノム配列の複数箇所にマッピングされる事があるためである。また、短い断片に SNP が入っていた場合には、マッピング作業が困難になる。さらに、発現量の非常に多い配列の場合、シークエンスの精度が悪く 1 塩基程度のミスマッチが置き、このミスマッチによりゲノムの他の場所に誤ってマッピングされることも考えられる。そこで、まず配列のマッピング方法について検討し、できるだけノイズを省いたデータセットの確立を目指した。２）DBTSS の拡充・プロモータ部分の特徴抽出。現在の DBTSS に、SOLEXA 由来の配列を導入し DBTSS の拡充を図った。さらに、全ての細胞種間で共通に発現している転写産物、ある細胞で固有に発現している転写産物、ある細胞で発現していない転写産物をそれぞれカタログ化することを目指した。３）転写産物量の推定とそれを決定づける要因の探索。転写に関わっているとされている要因を、配列だけでなくエピジェネティックなものも検討するため、ChIP-seq、クロマチン構造、転写因子結合部位等のデータを作成し、DBTSS のデータとの対応を見る。そして、各細胞ごとに発現量を決めている要因は何であるのかを明らかにする。そして、このデータを元に、それらの発現を決めている要因をデータベースと合わせて検討する。

４．研究成果

１）大量の配列情報の処理方法の確立とデータの拡充。

東京大学メディカルゲノムの鈴木らと共同で、次世代シークエンサーSOLEXA 由来の短い大量の mRNA5’端配列を約 3 億配列得た。これらには、DLD1, Bease2B, Ramos, MCF7, TIG, HEK293 のセルライン、また心臓、腎臓、肝臓、胸腺、脳の組織由来の RNA を用いた。一部には Hypoxia と Normoxia など酸素濃度による違いや、IL4 刺激の有無などのデータも取り入れた。これにマウスの NIH3T3 細胞由来のデータを合わせると全部で 32 種類の条件から転写開始点情報を得たことになる。これらのデータをマッピング、情報処理するパイプラインを作成し、得られた結果は、データベース DBTSS(http://dbtss.hgc.jp)に取り入れていった。 転写開始点一カ所ではない例がほとんどだったので、500bp ごとにクラスタリングし TSS cluster (TSC) と定義した。また、実験条件ごとに使われる転写開始点を簡単に比較できるようなツールを合わせて作成し web データベースとして公開した。例えば、図 1 では、cadherin-associated protein の例であるが、一番上流の TSC では腎臓、肝臓、脳での発現が見られるのに対し、

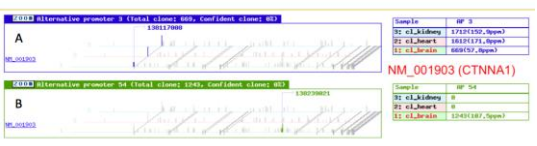

図1 新たに構築したviewerの例

CTNNA1のTSCの様子を示す。主なTSCのうち、もっともよく使われているTSCでは、腎臓、心臓、脳の全てで発現が観察された。これに対し、Bでは、脳でのみ発現が認められた。

第７イントロンに TSC が観察される場合があり、これは脳特異的であった。興味深いことに、この上流約 120bp のところに逆側方向にも転写が脳のみに認められ、その領域は leucine rich repeat transmembrane neuronal 2 の遺伝子領域と重なっていた。さらに、miRNA 等の non-coding RNA の転写開始点・プロモータ配列の検索も出来るような機能を追加した(Yamashita et al 2009)。

２）大規模シークエンサーデータに基づく転写制御領域解析

DBTSS に登録された転写開始点の情報を元に、TSC 付近の配列をプロモータとした。さらにそのプロモータ部分に関して、RNA polymerase, ヒストン修飾の ChIP-seq 情報、MNasaeI によるヌクレオソーム情報、ポリソーム分画による RNA-seq 情報などと比較を行った（図 2)。これらの解析で使った台規模シークエンサー由来の配列の総計は 2 億 7600 万配列に上る。この結果、発現量の多い TSC では、プロモータの転写開始点付近に特徴的なヌクレオソーム構造が観察された。またヒストンの H3K4me3、H3Ac 修飾も TSC の発現量によって強くなる傾向が観察された。(図 3)。さらに、プロモータ付近には RNA polymerase の集積が観察され、転写が起こっていることが推測された。興味深いことに、発現量が少ない(<5ppm)TSC でも RNA polymerase が結合していると推定される場合があり、そのうち約 20%では、他の細胞・組織で 5ppm 以上の発現が観察された。これは、RNA polymerase が存在してはいるが、arrest され転写が起きていない可能性が考えられる。また、発現量の高い遺伝子のうち 80%近くがポリソームでの mRNA の存在が確認され、これらのほとんどの遺伝子が翻訳されている可能性を示唆していた。しかしながら、転写されているのにもかかわらず、ポリソーム分画でのｍRNA が観察されない遺伝子も存在し、これらは mRNA レベルでの翻訳制御が行われている可能背が示唆された。また、一つの遺伝子にいくつかのプロモータがある選択的プロモータを持つ遺伝子群が少なくとも 43%の遺伝子で観察された（Yamashita et al 2010)。以上のような傾向は non-coding RNA の転写開始点に対しても同様に観察された(Sathira et al 2010,)。

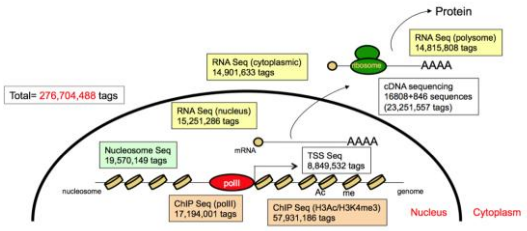

図2 用いた大規模シークエンサーのデータ

DLD1細胞に対して、RNA-seq, TSS-seq, Nucleosome-seq, polysome分画のRNA-seq, ヒストン修飾、PolIIのChIP-Seq, 等、合計約2.8億配列のデータを用いた。

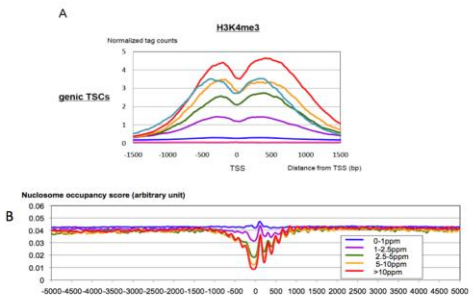

図3 転写開始点付近の特徴

転写開始点付近の特徴を示した。A.ヒストンH3K4me3の例。転写量が多いTSC由来の物ほどメチル化が強く観察された。B. ヌクレオソーム構造。転写開始点付近では、ヌクレオソーム構造の顕著な低下とその付近にしっかりとした構造が観察された。

3) 転写開始点 DBTSS の応用

DBTSS のデータ基に発現ベクターを調整し、下流にルシフェラーゼをつないで細胞内での発現量を測定した。また、プロモータ部分に存在する転写因子結合部位を予測して、それらの存在と発現量の推定を行った（Irie et al 2011)

マウスの RNAK の下流で動いている遺伝子群の転写開始点を DBTSS で推定し、その部分から上流 1000 下流 200 塩基を取り出し、既知の転写因子結合部位を推定した。その結果、STAT1 の転写因子結合部位が有意に濃縮していることがわかり、これらは実験的にそのパスウェイを確認することができた（Oshima et al in press)。

５．主な発表論文等

〔雑誌論文〕（計 6 件）

1. Ohshima D, Qin J, Konnno H, Hirosawa A, Shiraishi T, Yanai H, Shimo Y, Akiyama N, Yamashita R, Nakai K. Inoue J. RANK signaling induces interferon-stimulated genes in the fetal thymic stroma. BBRC. in press.

2. Irie T, Park SJ, Yamashita R, Seki M, Yada T, Sugano S, Nakai K, Suzuki Y. Predicting promoter activities of primary human DNA sequences. Nucleic Acids Res. 2011 Apr 12.

3. Yamashita R, Sathira NP, Kanai A, Tanimoto K, Arauchi T, Tanaka Y, Hashimoto SI, Sugano S, Nakai K, Suzuki Y. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. Genome Res. 2011 Mar 3.

4. Tanaka Y, Yamashita R, Suzuki Y, Nakai K. Effects of Alu elements on global nucleosome positioning in the human genome. BMC Genomics. 2010 May 17;11:309.

5. Sathira N, Yamashita R, Tanimoto K, Kanai A, Arauchi T, Kanematsu S, Nakai K, Suzuki Y, Sugano S.Characterization of Transcription Start Sites of Putative Non-coding RNAs by Multifaceted Use of Massively Paralleled Sequencer. DNA Res. 2010 Apr 17.
6. Yamashita R, Wakaguri H, Sugano S, Suzuki Y, Nakai K. DBTSS provides a tissue specific dynamic view of Transcription Start Sites. Nucleic Acids Res. 2009 Nov 12.

〔学会発表〕(計 16 件)
口頭発表
1. Riu Yamashita et al: "New features of DBTSS (DataBase of Transcription Start Sites) and promoter analysis." 1st international conference on bioinformatics and systems biology, India Jan. 2010
2. 山下理宇："A Database of Transcription Start Sites (DBTSS) for transcriptional regulation and its application" Pathogen genome analysis with next generation sequence technologies, Sapporo, Feb.21,2011
3. 山下理宇（鈴木穣　代理）："ヒトのトランスクリプトーム解析"　第 148 回農林交流センターワークショップ　次世代シーケンサーを利用したゲノム解析の実際理化学研究所　筑波　2010 年 9 月
4. Riu Yamashita et al. "Introduction and application of DBTSS (Database of Transcription Start Sites) for Transcriptional regulation analysis" BioSoft 1011. Beijing, China. (Mar. 23-25, 2011)
5. 山下理宇："転写制御機構解明のためのデータベース DBTSS の構築とその利用"　卓越した若手研究者の自立促進プログラム成果発表会、東京大学医科学研究所、東京、2011 年 3 月 30 日
6. Riu Yamashita et al. "Extension of DBTSS with SOLEXA sequences and analysis of bi-directional promoters." CBI 学会年会 東京　2010 年 9 月
7. 山下理宇："転写・翻訳制御のゲノムワイドな解析"　卓越した若手研究者の自立促進プログラム成果発表会、東京大学、東京、2009 年 9 月
8. 山下理宇："転写制御機構解明のためのデータベース DBTSS の構築とその利用"　卓越した若手研究者の自立促進プログラム成果発表会、東京大学医科学研究所、東京、2011 年 3 月 30 日
9. Riu Yamashita et al. "Extension of DBTSS with SOLEXA sequences and analysis of bi-directional promoters." CBI 学会年会 東京　2010 年 9 月
10. 山下理宇："転写・翻訳制御のゲノムワイドな解析"　卓越した若手研究者の自立促進プログラム成果発表会、東京大学、東京、2009 年 9 月

ポスター発表
1. Riu Yamashita et al, "An Application of DBTSS (Database of transcription start sites) for transcriptional regulation analysis", International Symposium "Towards Comprehensive Understanding of Immune Dynamism" Mar.1-2, 2011, Osaka, Japan
2. Riu Yamashita et al, "Analysis of bi-directional promoters based on a huge number of transcription start sites data", JSBI 学会年会, Dec.13-15, 2010, Kyushu, Japan
3. 山下理宇他、"New features of DBTSS (DataBase of Transcription Start Sites)" 分子生物学会、2010 年 12 月 9-10 日, 神戸
4. Riu Yamashita et al, "Extension of Database of Transcription Start Sites (DBTSS) with a Large Number of Sequences and Analysis of Bi-directional Promoters", Biocuration 2010 Oct.11-14, 2010, Tokyo, Japan
5. Riu Yamashita et al. "Extension of DBTSS with SOLEXA sequences and analysis of bi-directional promoters." CBI 学会年会 2010 年 9 月 15-17 日, 東京
6. 山下理宇他、"New features of DBTSS (DataBase of Transcription Start Sites)" 分子生物学会、2009 年 12 月 9-12 日, 横浜

〔その他〕
ホームページ
http://dbtss.hgc.jp

６．研究組織
(1)研究代表者
山下　理宇（YAMASHITA RIU）
東京大学・医科学研究所・特任助教
研究者番号：
10401259