博 士 論 文

無脊椎脊索動物カタユウレイボヤにおける
転写開始点の網羅的同定と
プロモーター領域の網羅的解析

東京大学大学院 新領域創成科学研究科
メディカル情報生命専攻

横森　類

目次

第 1 章　序論 1
1.1　本研究の背景・目的 1
1.2　本論文の構成 2

第 2 章　背景 3
2.1　TSS-seq 法 3
2.2　Spliced Leader (SL) トランススプライシング 5

第 3 章　材料と方法 6
3.1　データセット 6
3.2　リードの前処理 6
3.3　TSS およびトランススプライスアクセプター部位（TAS）の同定 6
3.4　TSS クラスター（TSC）と TAS クラスター（TAC）の同定 7
3.5　TSC と TAC の位置 8

第 4 章　結果 10
4.1　TSS と TAS の同定 10
4.2　TSS クラスター（TSC）と TAS クラスター（TAC）の同定 11
4.3　リボソームタンパク質遺伝子の TSS の同定 13
4.4　幅が 1 bp の TSC 16
4.5　CDS と 3′ UTR 上の TSC 24
4.6　既知コアプロモーターモチーフの分布 29
4.7　RP 遺伝子プロモーターの性質 34
4.8　非 RP 遺伝子プロモーターの性質 37
4.9　推定プロモーターの同定 45
4.10　SL *trans*-spliced 遺伝子のプロモーター候補 48
4.11　Non-*trans*-spliced 遺伝子プロモーターと *trans*-spliced 遺伝子プロモーターの比較 . 52
4.12　選択的プロモーターの同定 55

第 5 章　考察 57

謝辞 61

参考文献 62

付録 A　Supplemental Methods 68

A.1 CTGG TSC の除去 68
A.2 幅が 1 bp の TSC の除去 68
A.3 切断された RNA 由来の TSC の除去 68
A.4 右歪曲な TSC 69
A.5 TATA box の探索 69
A.6 TSS 分布のタイプ 69
A.7 相対エントロピー 70
A.8 超幾何分布による検定 70
A.9 クラスターペアの分類 71

付録 B Supplemental Figures 72

付録 C Supplemental Tables 84

第 1 章　序論

1.1　本研究の背景・目的

遺伝子の転写制御を理解するためは、転写開始点およびプロモーターの同定が必要不可欠である。近年のハイスループットな次世代シークエンサーと oligo-capping 法 (Maruyama and Sugano, 1994; Suzuki et al., 1997) や cap trapper 法 (Carninci et al., 1996) を組み合わせることにより、ゲノムワイドに転写開始点（TSS: transcription start site）を決定することが可能となっている (Suzuki et al., 2001; Carninci et al., 2005; Kawaji et al., 2006; van Heeringen et al., 2011)。この方法を用いて、いくつかの研究では網羅的な TSS の同定とそれによって得られたプロモーター領域の解析を行っている (Carninci et al., 2006; Yamamoto et al., 2009; Zhao et al., 2011)。

哺乳類のプロモーター解析では、プロモーターを TSS の分布に従い、sharp-type と broad-type に分類し、それぞれが TATA box と CpG アイランドに関連があることを示している (Carninci et al., 2006)。ここで、sharp-type のプロモーターとは、転写が狭い領域において開始し、TSS の分布が鋭いピークを示すタイプであり、broad-type のプロモーターとは、転写がより広い領域で始まり、TSS が広範囲に分散しているタイプのことである。この 2 つのタイプのプロモーターはショウジョウバエでも確認されているが (Rach et al., 2009; Ni et al., 2010; Hoskins et al., 2011)、ショウジョウバエの broad-type プロモーターは、CpG アイランドではなく、DNA replication-related element (DRE) などの位置が厳格に決まっていないモチーフと関連があることが示されている (Ni et al., 2010)。また、ヒトとショウジョウバエでは、broad-type プロモーターは sharp-type プロモーターよりも正確なヌクレオソームポジショニングを示すことが報告されている。さらにそれに関連して、ヒトの broad-type プロモーターは +1 ヌクレオソームに対応する領域で WW モチーフ（W は A もしくは T）の 10.5 bp の周期分布を示すことも報告されている (Forrest et al., 2014)。また、リボソームタンパク質（RP）遺伝子のプロモーターに対しても解析が行われており、哺乳類とショウジョウバエの RP 遺伝子プロモーターは、シトシン塩基から転写が始まるポリピリミジンイニシエーターモチーフを有し、sharp-type な転写開始点分布を示すことが報告されている (Yoshihama et al., 2002; Perry, 2005; Parry et al., 2010)。

尾索動物カタユウレイボヤは、無脊椎の脊索動物であり、生物学的研究において重要なモデル生物である (Sasakura et al., 2012; Stolfi and Christiaen, 2012)。カタユウレイボヤのドラフトゲノム配列は 2002 年に初めて発表され (Dehal et al., 2002)、その改良版となる Kyoto Hoya (KH) アセンブリが 2008 年に報告された (Satou et al., 2008)。最新の KH モデルによると（version 2013)、カタユウレイボヤゲノムは約 160Mb であり、遺伝子数はおよそ 15,000 個である。このことは、カタユウレイボヤがヒトやマウスなどの脊椎動物よりはるかにコンパクトなゲノムをもっていることを示しており、このコンパクトさのおかげで、遺伝子上流の転写調節配列探索がより簡便なものになっている (Johnson et al., 2005; Kusakabe, 2005; Irvine, 2013)。また、カタユウレイボヤが属する尾索動物は、無脊椎動物の中で我々ヒトに最も近い生物であることが示されてお

り (Delsuc et al., 2006; Putnam et al., 2008)、脊索動物の転写調節プログラムやその進化の解明において重要なモデル生物であると言える。さらに、ホヤの胚発生は非常に速く（約 18 時間で幼生になる)、電気穿孔法によってレポーターコンストラクトを大量の受精卵に一度に導入することが可能であることから、遺伝子上流の転写制御領域の解析が比較的容易であるといった利点もある (Corbo et al., 1997; Takahashi et al., 1999; Di Gregorio and Levine, 2002; Harafuji et al., 2002; Johnson et al., 2004; Kusakabe et al., 2004)。これらの利点は、カタユウレイボヤが遺伝子調節解析において有用なモデル生物であることを示している。しかしながら、カタユウレイボヤにおいてゲノムワイドな正確な転写開始点の同定およびプロモーター領域解析は現在のところほとんど行われていない。

カタユウレイボヤでは、約半分の遺伝子が spliced leader (SL) トランススプライシングを受けると言われている (Vandenberghe et al., 2001; Ganot et al., 2004; Hastings, 2005; Satou et al., 2006)。SL トランススプライシングとは、mRNA の 5′ 末端配列（アウトロン）が、16 塩基から成る短い RNA（SL 配列）に取って代わるという現象である (Vandenberghe et al., 2001)。したがって、トランススプライシングを受けた mRNA は、元々の 5′ 末端を失い、転写開始点の同定が困難となっている。重要なことに、トランススプライシングを受ける遺伝子の中には、トランススプライシングされやすい遺伝子とされにくい遺伝子があると報告されている (Matsumoto et al., 2010)。したがって、トランススプライシングされにくい遺伝子、もしくはトランススプライシングされやすいが高発現している遺伝子は、TSS-seq 法によって転写開始点を同定できる可能性がある。実際、Khare らは TSS-seq 法と理論的には同じ方法を用いて、トランススプライシングされる（*trans*-spliced）遺伝子の一つであるトロポニン I 遺伝子の転写開始点を同定している (Khare et al., 2011)。これらの研究結果は、TSS-seq 法がいくつかの *trans*-spliced 遺伝子の転写開始点の同定にも有用であることを示している。

本研究では、カタユウレイボヤの 5 つ（卵巣、心臓、体壁筋、神経複合体、幼生）のサンプルを用いて、転写開始点およびプロモーター領域を網羅的に同定・解析を行う。また、既にヒトにおいて行われているほとんどのプロモーター解析は CAGE 法を用いているため、本研究ではヒトにおいても TSS-seq 法で得られたデータを用いて再解析を行い、ヒトとカタユウレイボヤ間で比較をすることで非脊椎脊索動物と脊椎動物の類似点・相違点を見いだす。さらに、TSS-seq 法で得られたデータを用いて幾つかの *trans*-spliced 遺伝子の転写開始点の予測を行い、*trans*-spliced 遺伝子プロモーターの特徴を探る。最後に、選択的プロモーターの解析も行う。

1.2 本論文の構成

本論文は以下のように構成される。
第 2 章では、本研究の背景について述べる。
第 3 章では、本研究で用いた手法と材料について述べる。
第 4 章では、本手法で得られた結果を述べる。
第 5 章では、結果に対する考察を述べる。

第 2 章　背景

2.1　TSS-seq 法

TSS-seq 法とは、oligo-capping 法によって得られた mRNA の 5′ 末端配列を次世代シークエンサーを用いて大量にシークエンシングする方法である。また、oligo-capping 法とは、poly(A)+ RNA を BAP（Bacterial Alkaline Phosphatase）および TAP（Tobacco Acid Pyrophosphatase）で処理した後にオリゴヌクレオチドを加えることで、キャップ構造付きの RNA の 5′ 末端に選択的にオリゴヌクレオチドを導入する方法である。以下で、TSS-seq 法について説明する。

まず、Total RNA から dT selection によって選択された ploy(A)+ RNA に BAP 処理を行うことで、5′ 末端にあるリン酸基を除去する。次に TAP 処理を行うことで、キャップ構造を加水分解する。これらの処理によって、キャップ構造をもっていた RNA だけが 5′ 末端にリン酸基を持つようになる。そして、T4 RNA ligase を用いて合成オリゴを加えることで、キャップ構造をもつ RNA に選択的に 5′-oligo を導入する（図 1）。続いて、ランダムプライマーを用いて PCR を行い、cDNA を合成する。最後に、次世代シークエンサーを用いて、5′ 末端配列をシークエンシングする（図 2）。

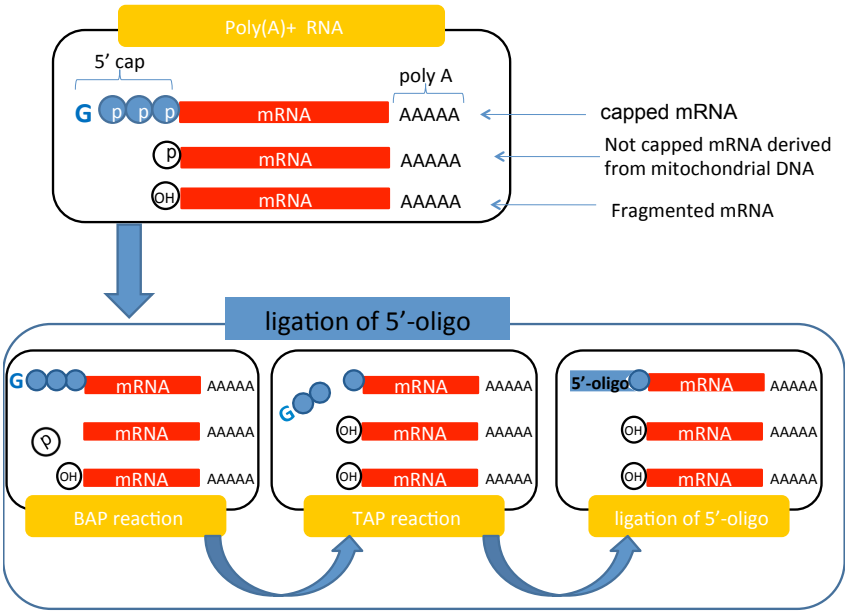

図 1: oligo-capping 法の概略図：Poly(A)+ RNA に BAP および TAP 処理を行うことによってキャップ構造付き RNA の 5′ 末端に選択的に合成オリゴを導入する方法。

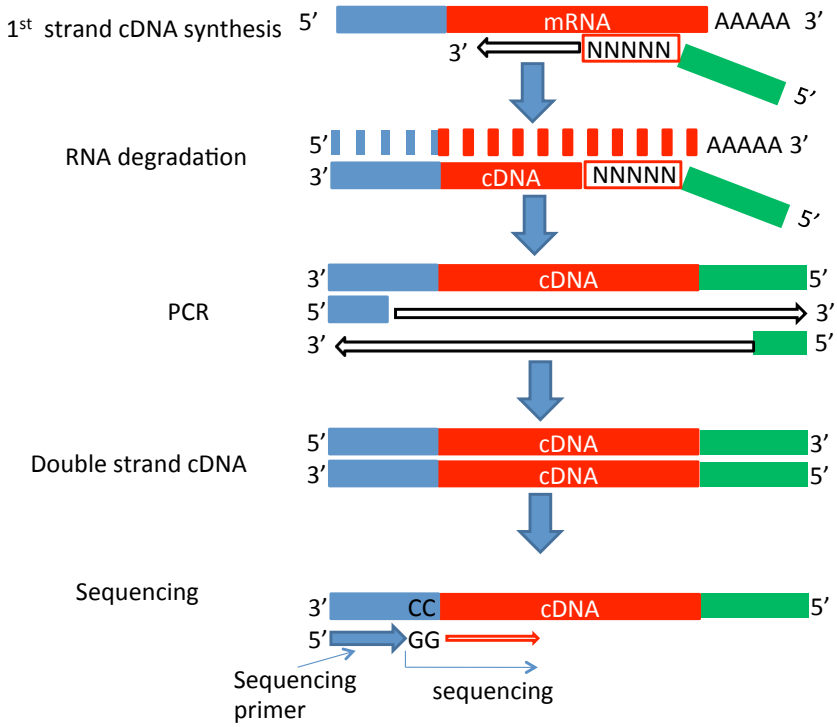

図 2: TSS-seq 法の概略図：oligo-capping 法によって得られた 5′-oligo 付きの RNA をランダムプライマーを用いて PCR によって増幅する。得られた cDNA の 5′ 末端を次世代シークエンサーでシークエンシングすることで、キャップ構造付き RNA の 5′ 末端に由来する配列が大量に得られる。

2.2 Spliced Leader (SL) トランススプライシング

カタユウレイボヤは、spliced leader (SL) トランススプライシングという特殊なスプライシング機構を持っている (Vandenberghe et al., 2001)。SL トランススプライシングとは、図 3A のように、SL RNA と呼ばれる短い RNA のエキソンが mRNA の 5′ 末端配列（アウトロン）と置き換わるという現象である。カタユウレイボヤでは、SL トランススプライシングによって生じた mRNA の 5′ 末端には、16 塩基から成る SL 配列と呼ばれる配列（ATTCTATTTGAATAAG）が付加される。最もよく記述される機能は、図 3B に示したようにオペロンから転写されたポリシストロニックな転写産物をモノシストロニックな転写産物へと分解する機能である。その他には、RNA polymerase I によって転写されたキャップ構造を持たないタンパク質コード RNA にキャップ構造をつけることで翻訳可能にする機能や翻訳効率を上げる機能、5′ UTR 上に存在する不都合なエレメントを除去する機能が提唱されている (Hastings, 2005)。

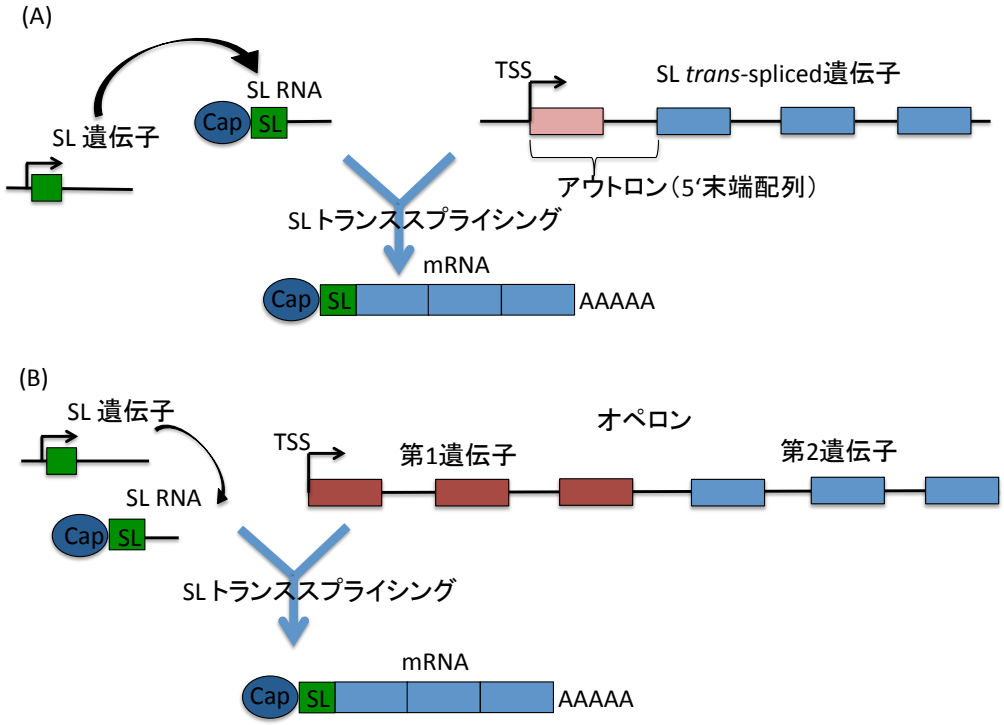

図 3: SL トランススプライシングの概略図。(A) アウトロンとよばれる 5′ 末端配列が SL 配列に取って代わる。(B) オペロンから転写されたポリシストロニックな転写産物をモノシストロニックな転写産物に分解する。

第 3 章　材料と方法

3.1　データセット

カタユウレイボヤにおける転写開始点を同定するために、4 つの異なる組織（卵巣、心臓、体壁筋、神経複合体）と一つの発生ステージ（幼生）に対して TSS-seq 法を適用し、計 5 つのリードデータを取得した。リードデータは 36 nt のシングルエンドリードである。また、ヒトの 15 の異なる組織（脂肪、脳、乳房、結腸、心臓、腎臓、肝臓、肺、リンパ液、筋肉、卵巣、前立腺、精巣、甲状腺、副腎）から得られた TSS-seq リードのマッピングデータを DBTSS (Yamashita et al., 2012) から取得した。

3.2　リードの前処理

取得したカタユウレイボヤのリードデータのフィルタリングを行った。まず、TSS-seq 法によって得られたリードデータから illumina の quality filtering をパスしたリードだけを選択した。続いて、そのリードの中から、最初の 2 塩基が GG で始まるリードだけを選択した。この GG は TSS-seq 法のプライマーの 3′ 末端配列に由来している。そのため、マッピングの際は、この GG は除去された。次に、アダプター配列およびプライマー配列（TCGTATGCCGTCTTCT、AATGATACGGCGACCA）を含むリードを探索し除去した。この探索において、配列間のレーベンシュタイン距離が 3 以下の場合、その配列はアダプターもしくはプライマーと判断した。さらに、リードを rRNA 配列群に対して、BLAST (Zhang et al., 2000) をかけることによって、rRNA 由来のリードを除去した。もし、リードが E-value $< 10^{-4}$ のヒットを持つ場合、そのリードを rRNA 由来のリードと判断した。rRNA 配列群 (SSUParc_115.fasta と LSUParc_115.fasta) は、Silva (Quast et al., 2013) からダウンロードした。最後に、リード中の SL 配列を探索した。ここで、SL 配列とは、16 塩基の SL 配列（ATTCTATTTGAATAAG）とのレーベンシュタイン距離が 3 以下の配列のことを言う。本研究では、SL 配列の有無と位置によって、リードを以下の 3 つのクラスに分類した。(1) SL 配列を持たないリード、(2) 5′ 末端に SL 配列を持つリード、(3) 5′ 末端以外の場所に SL 配列をもつリードである。 1 番目のクラスを SL(−) リード、2 番目のクラスを SL(+) リードと呼ぶ。3 番目のクラスのリードは、トランススプライシングされた mRNA に由来するのかどうか曖昧なため、本研究では用いないこととした。

3.3　TSS およびトランススプライスアクセプター部位 (TAS) の同定

TSS とトランススプライスアクセプター部位（TAS: *trans*-splice acceptor site）を同定するために、SL(−) リードと SL(+) リードを NovoAlign (V2.07.11; http://www.novocraft.com) と MapSplice (version 1.15.2) (Wang et al., 2010) を用いて、レファレンスゲノム上にマッピング

した。カタユウレイボヤのレファレンスゲノムとしては、Kyogo Hoya (KH) アセンブリ (Satou et al., 2008) を用いた。まず、NovoAlign を用いてリードをマッピングした。次に、multiply mapped リードと unmapped リードを MapSplice を用いてリマッピングした。ユニークにマップされた SL(+) リードと SL(−) リードの 5′ 末端をそれぞれ TSS と TAS と見なした。表 1 は、マッピングの際に用いた各ソフトウェアのオプションを示す。

表 1: マッピングの際に用いた各ソフトウェアのオプション

Reads	Novoalign	MapSplice
SL(−) reads	-s -o SAM -l 17 –3Prime	-L 17 -E 2 -m 2
SL(+) reads	-s -o SAM -l 9 –3Prime	-L 7 -E 0 -m 0

3.4 TSS クラスター（TSC）と TAS クラスター（TAC）の同定

TSS の高密度領域である TSC を同定するために、クラスタリングを行った。クラスタリングは二つの異なるクラスタリングからなる。第 1 のクラスタリングでは、全サンプルの TSS をマージして、35-bp のスライディングウィンドウを用いて TSS をクラスタリングすることで、initial TSC を同定した。このスライディングウィンドウを用いたクラスタリングは先行研究でも用いられていたが、この方法には、TSS の位置しか情報として用いていないという欠点が存在する。この欠点のため、真の TSS 間に低頻度のノイズ TSS が存在した場合、得られたクラスターが不自然に大きくそして低密度になることがある。そこで、この問題を解決するために、TSS の頻度を考慮した第 2 のクラスタリングを行うことにした。

第 2 のクラスタリングは、各 initial TSC に対して行われた。各 TSC は、カタユウレイボヤで計 5 個のサンプル、ヒトで計 15 のサンプルを用いるため、カタユウレイボヤおよびヒトで、それぞれ最大 5 と 15 の TSC からなる。まず、第 2 のクラスタリングでは、各サンプルの TSC の高頻度 TSS とピーク TSS を探索した。高頻度 TSS は、その TSC の中で最も高頻度な TSS の頻度より 10 分の 1 以上の頻度がある TSS と定義した。相対的に高い頻度を持つ TSS は、ノイズ TSS ではない可能性が高い。また、高頻度 TSS の中で、最も高頻度な TSS の頻度より 2 分の 1 以上の頻度がある TSS をピーク TSS と定義した。ただし、このステップでは、タグ数が 100 以上の TSC だけを用いた。なぜなら、タグ数が少ない場合、その TSS 分布が真の分布を表していない可能性があり、誤った高頻度 TSS とピーク TSS を導いてしまう可能性があるからである。次に、全サンプルから得られた高頻度 TSS をマージし、スライディングウィンドウを用いてクラスタリングすることで、サブ TSC を同定した。そして、各 TSC を隣接するサブ TSC の中間点で分割した。ただし、もし分割して得られた TSC がピーク TSS を含んでいない場合は、明確に高頻度な TSS をもたない曖昧な TSC と見なし解析には用いないことにした。最後に、すべての initial TSC に

対して、上記 2 つのステップを分割されなくなるまで繰り返した。また、カタユウレイボヤにおいては、4-bp のスライディングウィンドウを用いた第 1 のクラスタリングで TAS のクラスター（TAC）も同定した。

クラスタリング後、クラスターの両側に存在する低頻度のノイズ TSS によって不自然なサイズのクラスターになるのをさけるため、極端な外れ TSS を除外した。各サンプルのクラスターに対し、四分位範囲（IQR）、すなわち、25 パーセンタイル（Q1）と 75 パーセンタイル（Q3）間の距離を計算し、Q1 − 3 × IQR 以下もしくは Q3 + 3 × IQR 以上に存在するタグを極端な外れ値として除外した。

同定した各クラスターは、カタユウレイボヤとヒトにおいてそれぞれ、各サンプル由来の最大 5 個と 15 個のクラスターから成る。各クラスターの代表転写開始点（代表 TSS）および代表トランススプライスアクセプター部位（代表 TAS）（以下、代表点）は、各サンプルから得られた 100 タグ以上のクラスターの最も高頻度なポジションの多数決によって決定した。もし、複数の代表点が存在した場合は、最も上流に位置する代表点を採用した。また、代表点を最も高頻度なポジションとしてもつ各サンプル由来のクラスターの内、最もタグ数の多いクラスターを代表クラスターとして定義した。特別に明示しない限り、TSC の TSS 分布とは、代表 TSC の TSS 分布のことを指す。また、5 パーセンタイルと 95 パーセンタイル間の距離を、TSC の幅として定義した。

3.5 TSC と TAC の位置

カタユウレイボヤでは、TSC と TAC の位置は、KH 遺伝子モデル (version 2013) (Satou et al., 2005) に基づいて決定された。KH 遺伝子モデルには、non-SL、 SL、 ND の 3 つのタイプの転写産物モデルが存在する。non-SL は、トランススプライシングを受けない転写産物であり、その 5′ 末端は TSS を意味する。一方、SL は、トランススプライシングを受ける転写産物であり、その 5′ 末端は TAS を意味する。最後に、ND は、その 5′ 末端が TSS なのか TAS なのか分からない曖昧な転写産物である。

同定した TSC と TAC はその位置によって、TSS、TAS、5′ UTR、CDS、3′ UTR、intron、intergenic の 7 つのカテゴリーに分類された。各クラスターの位置は、コア領域（代表クラスターの IQR）中の最も高頻度なポジションに基づいて決定された。ただし、クラスターが TSS もしくは TAS にオーバーラップしている場合は、CDS とオーバーラップしていない場合に限り、TSS もしくは TAS に分類された。転写産物モデルのオーバーラップのため、クラスターが複数のカテゴリーに分類可能な場合は、(1)TSS、(2)TAS、(3)5′ UTR、(4)CDS、(5)3′ UTR、(6)intron、(7)intergenic の優先順位に基づいて分類した。また、TAC の場合は、(1)TAS、(2)TSS、(3)5′ UTR、(4)CDS、(5)3′ UTR、(6)intron、(7)intergenic の優先順位に基づいて位置を決定した。

ヒトでは、TSC の位置は ReSeq のアノテーションに基づいて、TSS、5′ UTR、CDS、3′ UTR、non-coding RNA の exon (ncRNA)、intron、intergenic の 7 つのカテゴリーに分類された。転写産物モデルのオーバーラップのため、クラスターが複数のカテゴリーに分類可能の場合は、(1) TSS、(2) 5′ UTR、(3) CDS、(4) 3′ UTR、(5) exon(ncRNA)、(6) intron、(7) intergenic の優

先順位に基づいて分類した。

第 4 章　結果

4.1　TSS と TAS の同定

カタユウレイボヤの 5 つのサンプルから得られた TSS-seq リードをゲノム上にマッピングすることによって TSS と TAS を同定した。リードデータは、前処理を受け、SL(−) リードと SL(+) リードに分類された (表 2)。SL(−) リードとは、トランススプライシングを受けなかった mRNA の 5′ 末端に由来するリードで、SL(+) リードとは、トランススプライシングを受けた mRNA の 5′ 末端に由来するリードのことである。これら 2 種類のリードはそれぞれゲノム上にマップされ (表 3)、ユニークにマップされた SL(−) リードの 5′ 末端を TSS、SL(+) リードの 5′ 末端を TAS とした。ヒトの TSS に関しては、DBTSS (Yamashita et al., 2012) から 15 サンプルのデータを利用した。

表 2: リードの前処理。quality1 は illumina のクオリティフィルタングをパスしたリード、with GG は先頭に GG をもつリードを表す。BWM と NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

Sample	raw reads	quality1	with GG	not contaminants	unmapped to rRNA	SL(−) reads	SL(+) reads
ovary	34913433	23951726	21597614	19723497	13628650	7148409	6161390
heart	29518199	22147530	20538561	19233112	18214384	14094602	3919239
BWM	32564064	23437074	21679278	20341295	18080682	13887188	4027858
NC	31628569	23377130	21921787	20966915	19802857	16612659	3018631
larva	26967445	20501993	18482078	17079141	13838575	8859559	4775063

表 3: マッピング結果。表は各サンプルのリード数と全リード数に対する割合を示す。

SL(−) reads

Sample	reads		lowquality or homopolymer		unmapped		mapped		multiply mapped		uniquely mapped	
ovary	7,148,409	100%	89314	1.20%	437,486	6.10%	6,621,609	92.60%	269,960	3.80%	6,351,649	88.90%
heart	14,094,602	100%	111684	0.80%	2,756,007	19.60%	11,226,911	79.70%	510,169	3.60%	10,716,742	76.00%
BWM	13,887,188	100%	140469	1.00%	429,160	3.10%	13,317,559	95.90%	635,947	4.60%	12,681,612	91.30%
NC	16,612,659	100%	135680	0.80%	5,549,562	33.40%	10,927,417	65.80%	395,750	2.40%	10,531,667	63.40%
larva	8,859,559	100%	97619	1.10%	372,453	4.20%	8,389,487	94.70%	540,936	6.10%	7,848,551	88.60%

SL(+) reads

Sample	reads		lowquality or homopolymer		unmapped		mapped		multiply mapped		uniquely mapped	
ovary	6,161,390	100%	280667	4.60%	26	0.00%	5,880,697	95.40%	1,966,452	31.90%	3,914,245	63.50%
heart	3,919,239	100%	118889	3.00%	366	0.00%	3,799,984	97.00%	929,679	23.70%	2,870,305	73.20%
BWM	4,027,858	100%	167763	4.20%	17	0.00%	3,860,078	95.80%	910,913	22.60%	2,949,165	73.20%
NC	3,018,631	100%	89912	3.00%	28	0.00%	2,928,691	97.00%	696,718	23.10%	2,231,973	73.90%
larva	4,775,063	100%	156292	3.30%	106	0.00%	4,618,665	96.70%	1,037,245	21.70%	3,581,420	75.00%

4.2 TSS クラスター（TSC）と TAS クラスター（TAC）の同定

同定した TSS がどのプロモーターに由来するものなのかを決定するために、同定した TSS をクラスタリングすることで、高密度な TSS 領域である TSC を同定した（第 3.4 節 参照)。ただし、タグ数が 100 未満の TSC は今後の解析には用いないことにした。なぜなら、これらの TSC は切断された転写産物や転写機構に本来備わっている曖昧な転写に由来するかもしれないからである (Yamashita et al., 2011)。このクラスタリングと選択によって、カタユウレイボヤとヒトにおいてそれぞれ 9792 個の TSC と 15498 個の TSC を同定した。本研究ではこれら TSC 群を初期の TSC セットとして用いることにした。

カタユウレイボヤの各 TAS には、50 bp 以下離れた近傍ペアが多数存在することが報告されている。この TAS ペアのメジャーな TAS とマイナーな TAS 間の距離の分布は、+3 の位置でピークを示す (Matsumoto et al., 2010)。また、この報告の中では、この短い間隔の選択的な TAS の使用は、スプライシング機構の確率的な側面を反映しており、コードされるタンパク質の構造には影響を与えないであろうと示唆されている。そこで、お互いに近い距離にある TAS を 4-bp のスライディングウィンドウを用いてクラスタリングし、形成されたクラスター（TAC）を一つの TAS と見なすことにした（第 3.4 節 参照)。ただし、タグ数の少ない TAC はエラーによって生じた TSS-seq リードに由来する可能性があるので除去した。また、TAS の特徴である AG motif (Agabian, 1990; Nilsen, 1993) を代表 TAS 上流にもたない TAC は除去された。このクラスタリングと選択によって、5373 個の TAC をカタユウレイボヤにおいて同定した。ほとんどの TAC (88%）は既知の TAS 上に存在していた（表 4)。残り 22% の既知 TAS 上に存在しない TAC は未知の TAS を表していると考えられる。

表 4: カタユウレイボヤにおいて同定された TAC の数。同定された TAC は位置に基づいて 7 つのカテゴリーに分類された（第 3.5 節 参照)。

Location	TACs
TSS	1 (0.0%)
TAS	4748 (88.4%)
5′ UTR	70 (1.3%)
CDS	40 (0.7%)
3′ UTR	8 (0.1%)
intron	104 (1.9%)
intergenic	402 (7.5%)
total	5373 (100%)

4.3 リボソームタンパク質遺伝子の TSS の同定

初期 TSC を同定後、まずカタユウレイボヤのリボソームタンパク質（RP）遺伝子の TSS を探索した。哺乳類やショウジョウバエでは、RP 遺伝子の TSS はポリピリミジンに富む配列のシトシン塩基上に存在することが多く、その分布は鋭いピークを示すことが知られている (Yoshihama et al., 2002; Perry, 2005; Parry et al., 2010)。このように RP 遺伝子の TSS は特殊な性質を持つため、他の遺伝子の TSS よりもより明確に決定できることが期待される。RP 遺伝子 TSS を同定するため、まずカタユウレイボヤにおいて 79 個のヒト RP のオーソロガス遺伝子を同定した。79 個のうち、78 個のオーソロガス遺伝子は BLAST によって同定できたが、*RPL41* のオーソロガス遺伝子は同定できなかった。これは、KH モデルのアノテーションが不完全であり、既知のどのタンパク質配列もヒト RPL41 に類似性を示さなかったからである。そこで、*RPL41* のオーソロガス遺伝子を同定するために、Ribosomal Protein Gene database (RPG) (Nakao et al., 2004) から得られた *Rpl41* の cDNA を BLAST-like アライメントツールである BLAT(Kent, 2002) を用いてゲノム上にマップした。これにより、*Rpl41* が KH.C9.469 の遺伝子座に存在することが分かった。そして、全 RP 遺伝子の上流を調べることで、79 個の RP 遺伝子それぞれに対して、ポリピリミジンに富む配列上に存在し鋭いピークの TSS 分布を示す TSC を発見することに成功した（表 S3)。これらの 79 個の TSC は、各 RP 遺伝子の TSC 候補（TSS 候補）と考えられる。

遺伝子モデルにアノテーションされている 78 個の RP 遺伝子の内、ほとんど全ての遺伝子 (72/78）において、同定した TSC 候補の代表 TSS は既知の TSS と一致していた。このことは、これら 72 の TSC 候補の代表 TSS が RP 遺伝子の真の TSS を表していることを強く示唆している。一方、6 つの RP 遺伝子（*Rps2*、*Rps5*、*Rps21*、*Rpl21*、*Rpl29*、*Rpl37*）に対して同定した TSC 候補の代表 TSS は既知 TSS から 15 塩基以上離れた場所に存在していた。しかしながら、これらの代表 TSS は以下の二つの理由から最も有力な真の RP 遺伝子 TSS であると考えられる。まず一つ目は、これら 6 つの RP 遺伝子の代表 TSS が既知 TSS とは異なり、上述した 72 個の RP 遺伝子の代表 TSS と同様にポリピリミジンに富む配列上に存在しているということである。二つ目は、代表 TSS 上に存在する TSS-seq タグの数が、既知 TSS 上とその近傍に存在するタグの数よりも遥かに多いことである（図 S1-S6)。注目すべきことに、*Rpl21* のプロモーターは既知 TSS からおよそ 1700 bp も離れている場所に存在していた。これは、*Rpl21* がオペロン（KHOP.805）の第 2 遺伝子であることが原因である。*Rpl21* のプロモーター（TSC）は *Rpl21* 遺伝子の 5′ 末端付近には存在せず、オペロンの 5′ 末端付近に存在していた（図 4)。また、*Rpl21* 遺伝子の 5′ 末端にはトランススプライスアクセプター部位を表す TAC が存在していた。これらの結果は、*Rpl21* 遺伝子はポリシストロニックな形で転写され、SL トランススプライシングによってリボソームタンパク質をコードするモノシストロニックな転写産物に分解されることを示唆している。

以上の結果に基づいて、同定した 79 個の TSC は RP 遺伝子プロモーターに由来し、それらの代表 TSS は各 RP 遺伝子の主要 TSS を表すと見なした。そのため本研究ではこれら 79 個の TSC を RP 遺伝子プロモーターの解析に用いる。また、以降の節で行われる TSC のフィルタングは受

けないものとする。

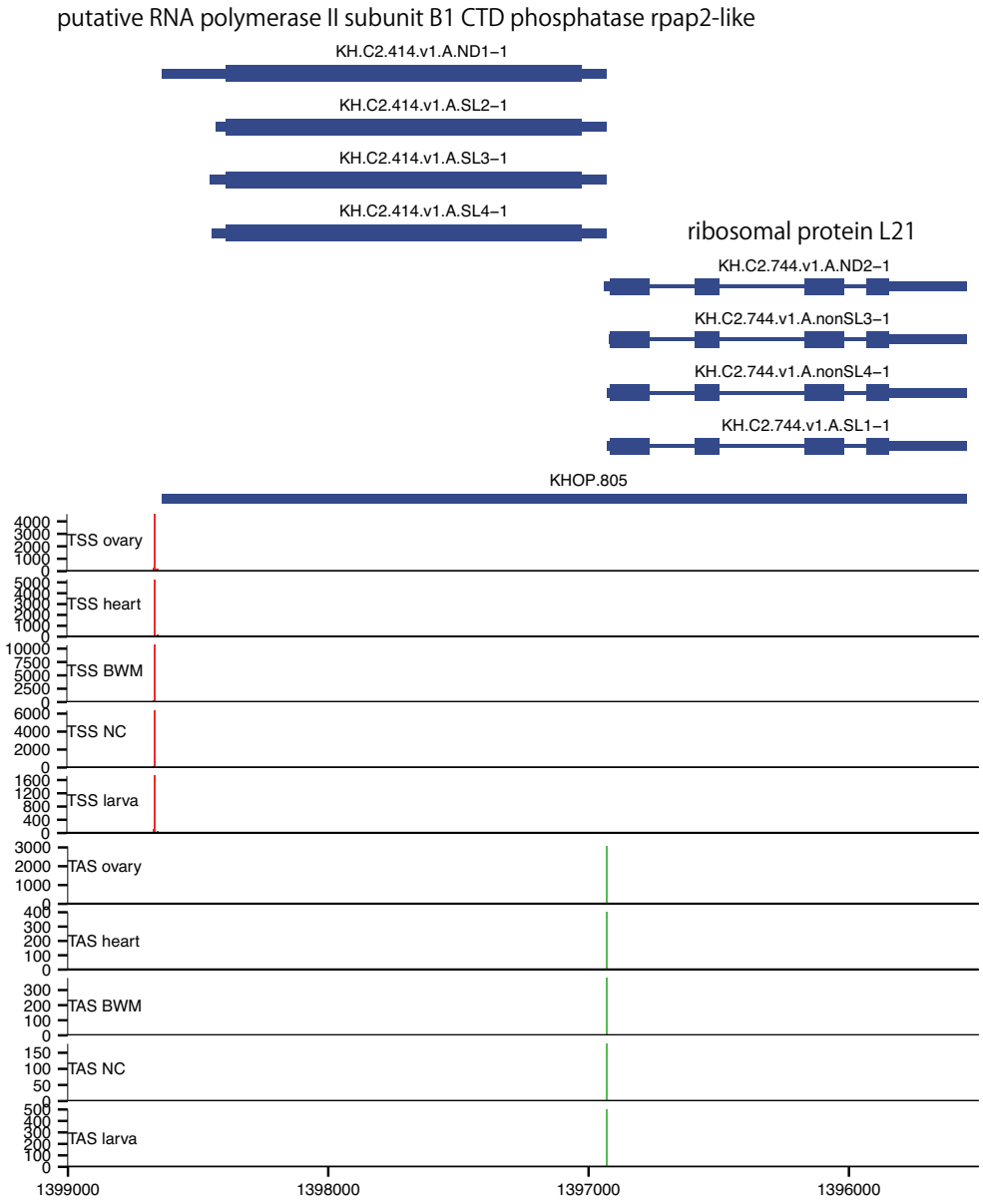

図 4: リボソームタンパク質 L21 遺伝子の TSS。リボソームタンパク質 L21 遺伝子は、二つの遺伝子からなるオペロン（KHOP.805）の下流遺伝子である。上流遺伝子は “putative RNA polymerase II subunit B1 CTD phosphatase rpap2-like” をコードしている。赤と緑のバーがそれぞれ TSS と TAS を表す。y 軸はタグ数を表す。BWM と NC はそれぞれ body wall muscle (体壁筋) と neural complex (神経複合体) を表す。

4.4 幅が 1 bp の TSC

カタユウレイボヤとヒトの両方において、初期 TSC セットの内、多数が遺伝子間領域に存在することが分かった（表 5、6)。注目すべきことに、遺伝子間領域に存在する TSC の多くが幅 1bp であり（図 5A)、それらは明確な CTGG motif を示した（図 5B)。しかしながら、これらの TSC は、TSS-seq における 5′ oligo のミスハイブリダイゼーションによって生じたのものだと考えられた（図 6)。もし、図 6 に示したメカニズムが真ならば、上流 4 塩基が CTGG の TSC はエキソンのアンチセンス鎖に存在するはずである。実際、多くの幅 1 bp の TSC がエキソン（5′ UTRs、CDSs、 3′ UTRs) のアンチセンス鎖上に存在し（図 7)、それらは、期待通りに CTGG モチーフを示した（図 8)。これらの結果は、CTGG TSC が TSS-seq のアーティファクトであることを示している。したがって、CTGG を持つ TSC は今後の解析から除去した（付録 A.1 参照)。

CTGG TSC を除去後も、特にイントロンと遺伝子間領域に幅が 1 bp の TSC が存在していることが分かった（図 9)。注目すべきことに、これらの TSC は、下流に AT-rich 領域を持っていた（図 5C)。イントロンと遺伝子間領域に存在する幅 1 bp の TSC の 15 bp 下流の AT 含量を調べたところ、カタユウレイボヤとヒトにおいてそれぞれ AT 含量が 0.8 以上、0.66 以上の AT-rich な TSC が多く存在することが分かった（図 10)。また、カタユウレイボヤでは、AT-rich な TSC に加えて、逆鎖のスプライスドナー部位の近くに幅 1 bp の TSC が存在していた（図 5D)。

CTGG TSC は TSS-seq による技術的なノイズであることはわかるが、他の 2 つのタイプの幅 1 bp の TSC がノイズであるのか、不定型のプロモーターを表しているのかは明確ではない。これらの TSC はカタユウレイボヤとヒトプロモーターの既知の特徴である PyPu モチーフを持たないので、実験的もしくは生物学的ノイズによるものかもしれない。そこで、本研究ではこれらの幅 1 bp の TSC は除去することとした（付録 A.2 参照)。他の幅 1 bp の TSC は今後の解析にも含めた。

表 5: カタユウレイボヤにおいて同定された TSC と TAC の数。同定された TSC は位置に基づいて 7 つのカテゴリーに分類された（第 3.5 節 参照)。ただし、79 個の RP 遺伝子の TSC は、たとえ既知 TSS 上に存在していなくても便宜上「TSS」に含むこととした。

Location	TSCs
TSS	2097 (21.4%)
TAS	122 (1.2%)
5′ UTR	420 (4.3%)
CDS	1623 (16.6%)
3′ UTR	1459 (14.9%)
intron	721 (7.4%)
intergenic	3350 (34.2%)
total	9792 (100%)

表 6: ヒトにおいて同定された TSC の数。同定された TSC は位置に基づいて 7 つのカテゴリーに分類された（第 3.5 節 参照)。

Location	TSCs
TSS	5207 (33.6%)
5′ UTR	1452 (9.4%)
CDS	1622 (10.5%)
3′ UTR	1312 (8.5%)
exon(ncRNA)	273 (1.8%)
intron	1650 (10.6%)
intergenic	3982 (25.7%)
total	15498 (100%)

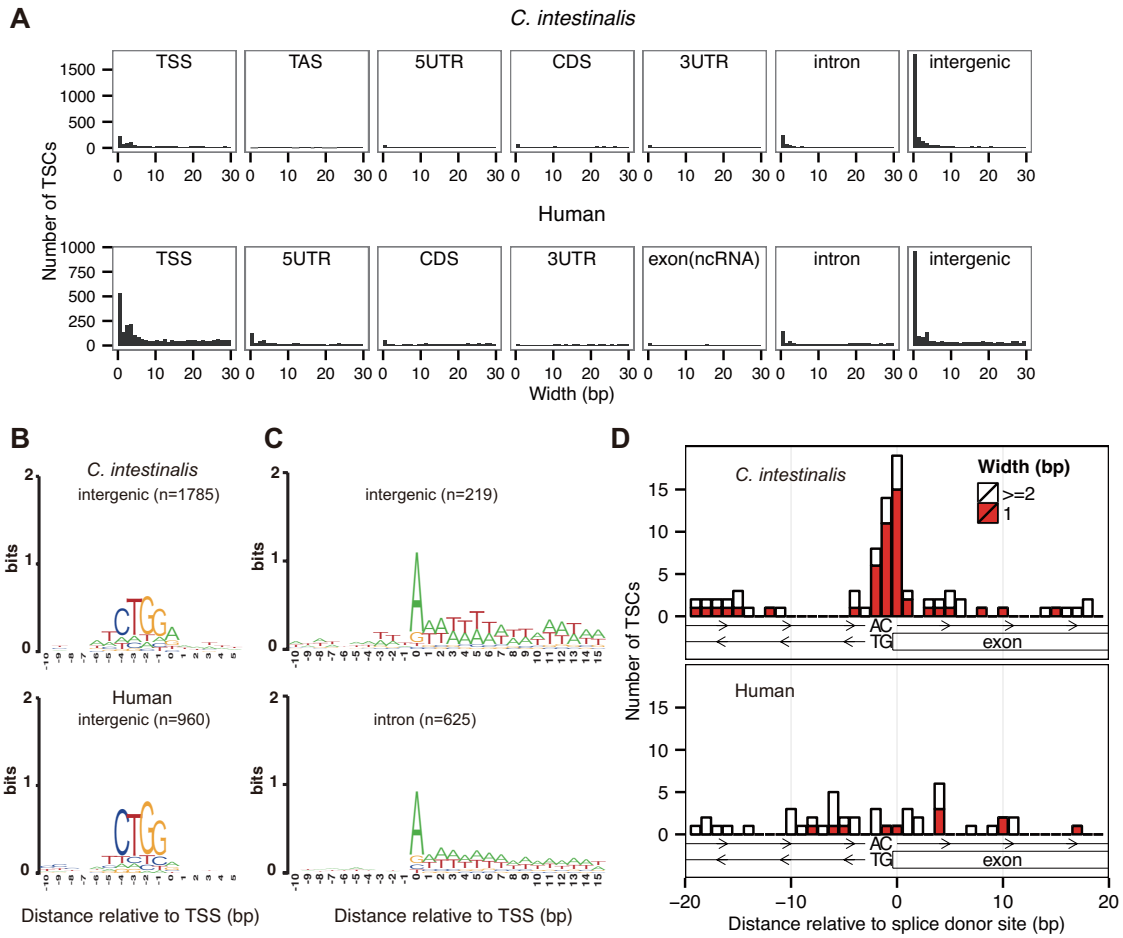

図 5: 幅 1 bp の TSC。(A) TSC の幅の分布。(B) 遺伝子間領域にある幅 1 bp の TSC の sequence logo。 括弧内の数は TSC の数を表す。(C) イントロンおよび遺伝子間領域に存在する幅 1 bp の TSC の sequence logo（図はカタユウレイボヤの TSC)。 (D) 逆鎖のスプライスドナー部位付近にある幅 1 bp の TSC。スプライスドナー部位付近にある幅 1 bp の TSC の数を調べた。赤と白のバーはそれぞれ、幅 1 bp の TSC とその他の TSC の数を表す。

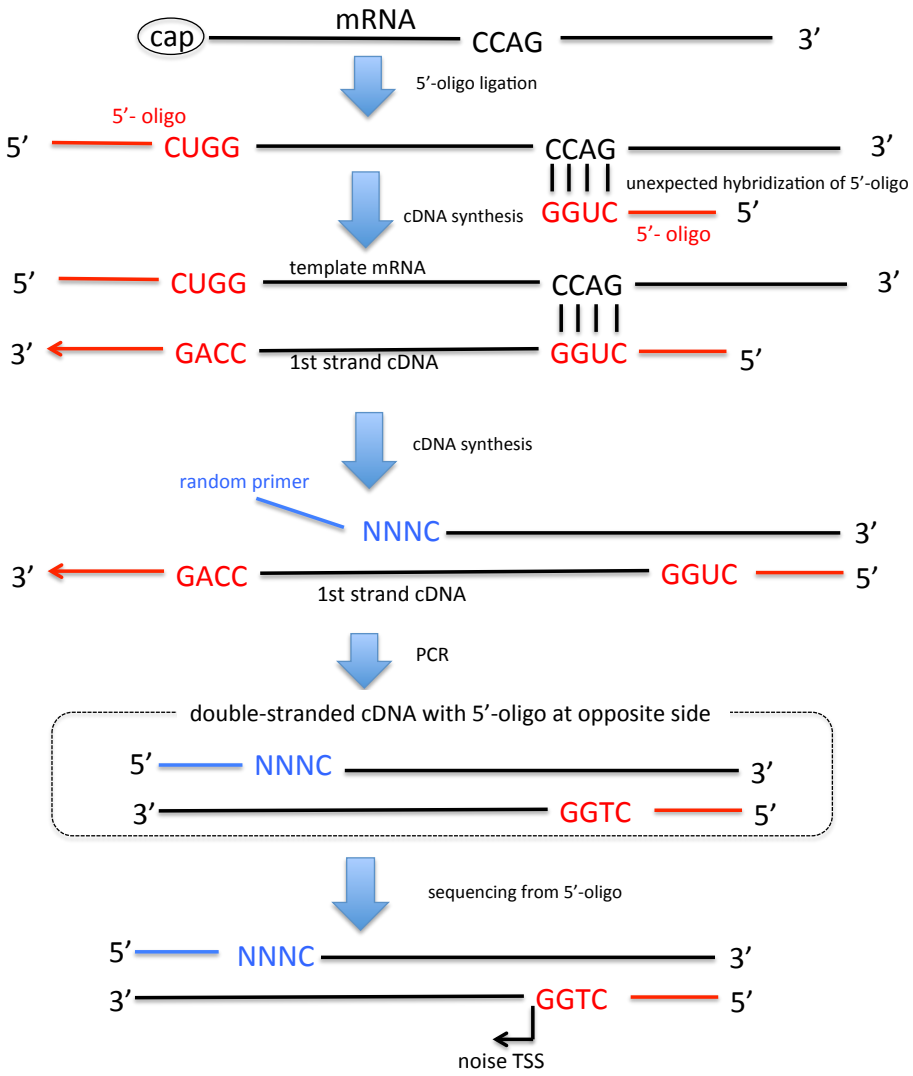

図 6: CTGG TSC の生成メカニズム。5′-oligo capping 中に、5′-oligo 配列は mRNA 上の CCAG にハイブリダイズすることができる。cDNA 合成後、この期待していないハイブリダイゼーションは、真の 5′ 末端の反対側に 5′-oligo 配列をもつ二本鎖 cDNA を生み出す。その後のシークエンシングによってその反対側の配列の由来するリードが得られる。このリードは mRNA のアンチセンス鎖にマップされ、マップされたポジションの上流 4 塩基は CTGG となる。

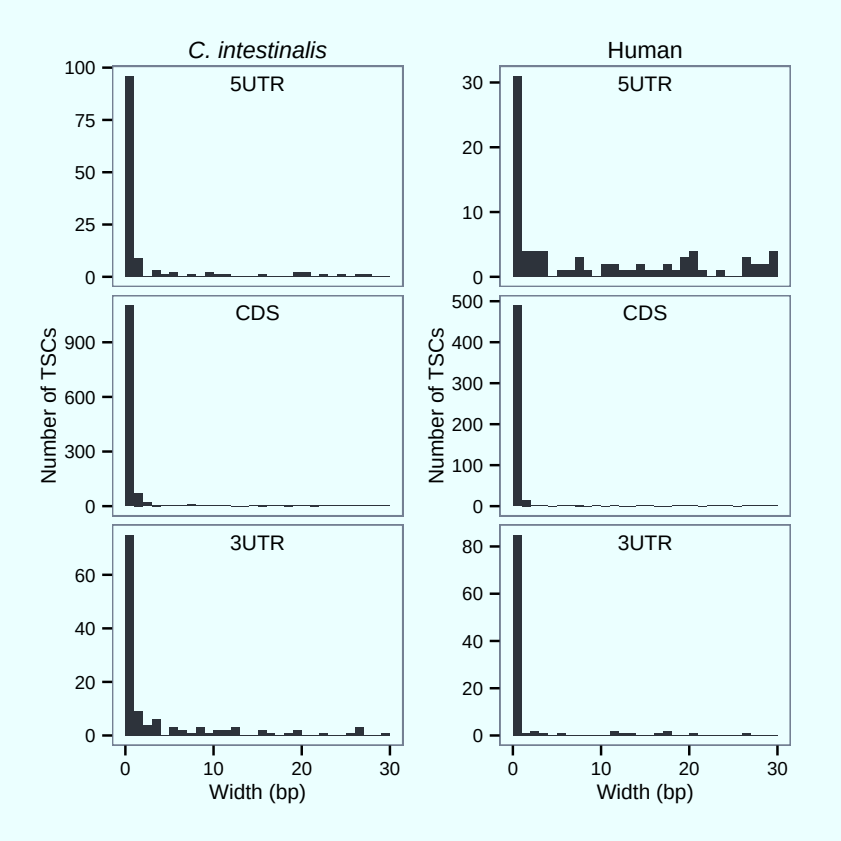

図 7: エキソンのアンチセンス鎖上にある TSC の幅の分布。TSC をアンチセンス鎖における位置に従って分類した。多くの幅 1 bp の TSC が 5′ UTR や CDS、3′ UTR のアンチセンス鎖上に存在していた。

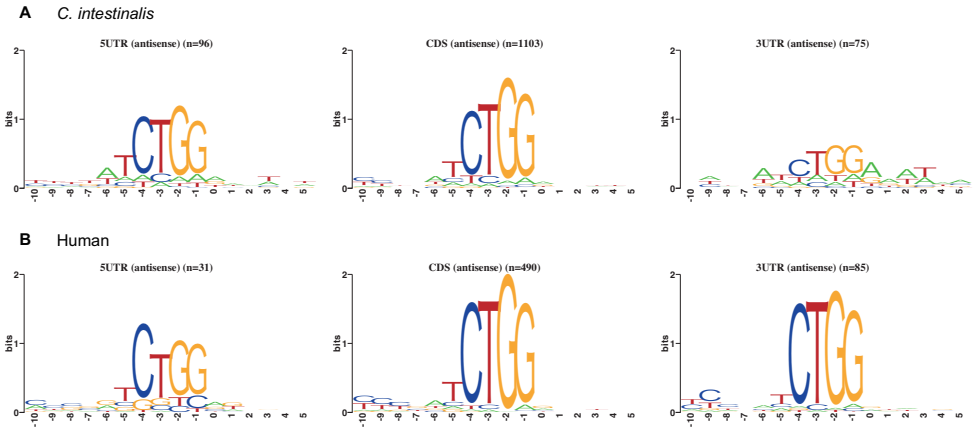

図 8: 5′ UTR、CDS、3′ UTR のアンチセンス鎖に存在する TSC の sequence logo。x 軸は TSS からの距離を表す。括弧内の数字は TSC の数を表す。

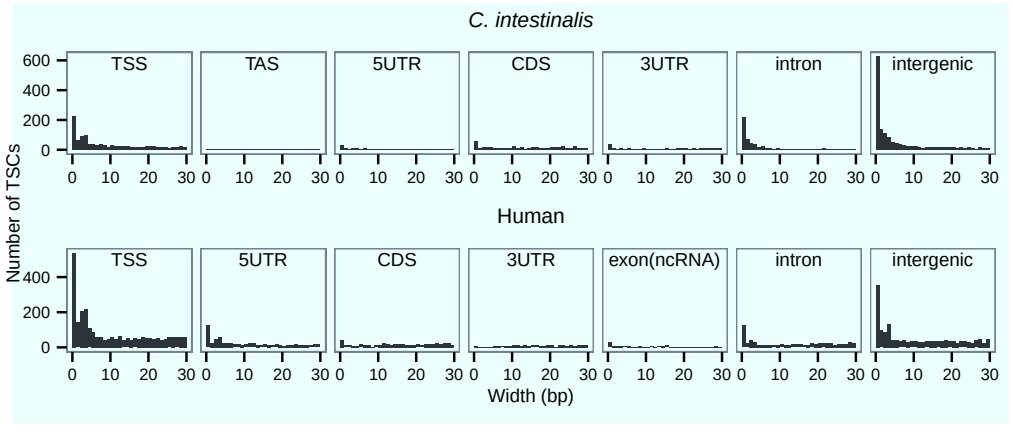

図 9: CTGG TSC 除去後の TSC の幅。TSC アンチセンス鎖における位置に従って分類した。x 軸と y 軸はそれぞれ、TSC の幅と TSC の数を表す。

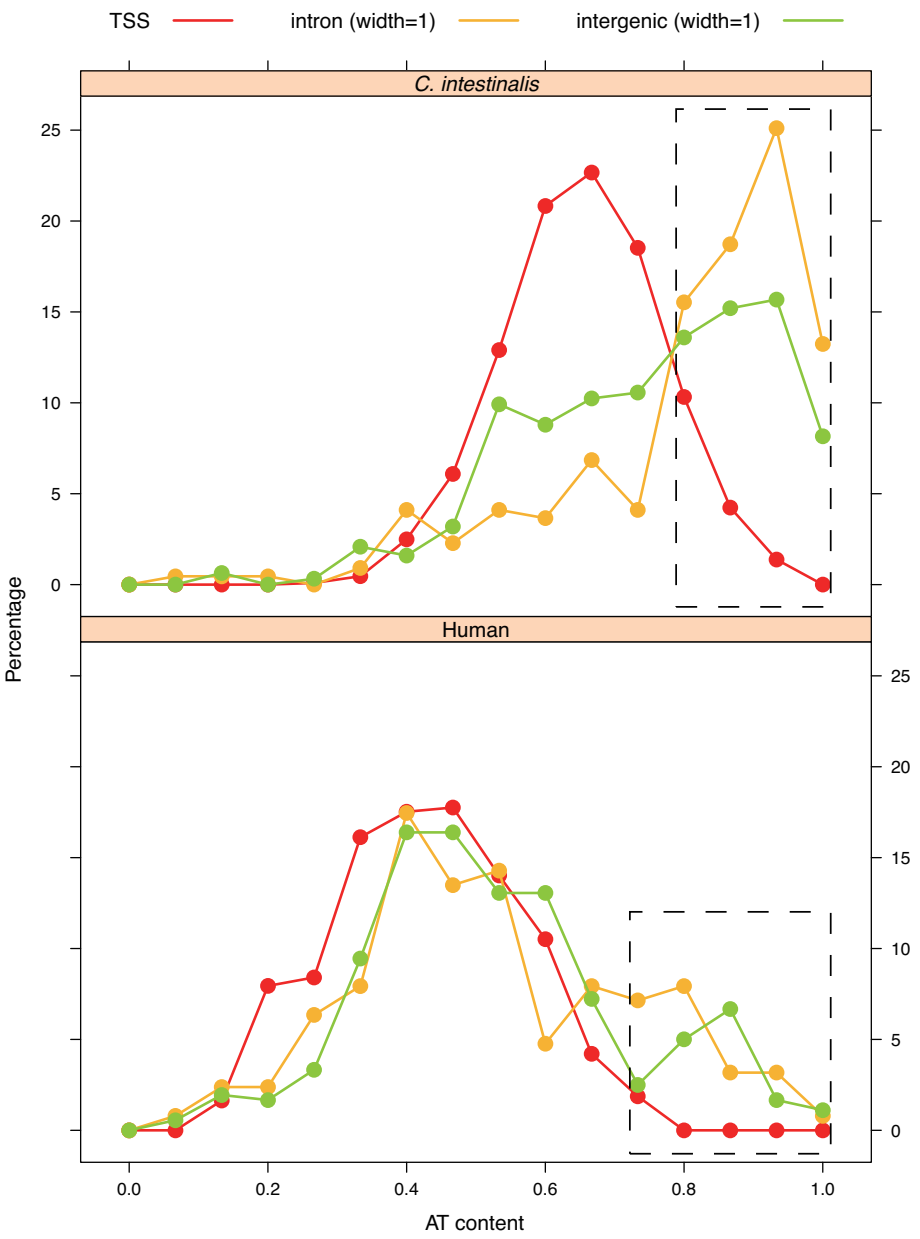

図 10: イントロンおよび遺伝子間領域に存在する幅 1 bp の TSC の下流 15 bp の AT 含量。+1 から +15 の下流 15 bp の AT 含量を調べた。ただし、位置 0 は TSS の位置を表す。イントロンと遺伝子間領域に存在する幅 1 bp の TSC の下流 15 bp の AT 含量を既知 TSS に位置する TSC の AT 含量と比較した。カタユウレイボヤとヒトにおいてそれぞれ 0.80 以上の AT 含量をもつ TSC と 0.66 以上の AT 含量をもつ TSC を多数発見した。

4.5 CDS と 3′ UTR 上の TSC

遺伝子間領域だけでなく、初期 TSC セットの多数の TSC が CDS と 3′ UTR 上にも存在していることが分かった（表 5)。TSC が存在する CDS 群のほとんど（80%）において、半分以上の領域が TSC にカバーされていた（図 11A)。このことは、TSS-seq リード（タグ）が CDS 上の広範囲に分布していることを示している。また、全体の CDS の多くの領域が TSC でカバーされているが、イントロン領域はカバーされていない転写産物モデルを多数発見した（図 11B)。多くの転写産物モデルにおいて TSS-seq タグは特異的に CDS 領域に分布しているようである。CDS と 3′ UTR 上に存在する TSC は PyPu モチーフとは顕著に異なり、−1 position が PyPu モチーフと同じく比較的ピリミジン塩基で保存されている一方で、0 position の保存度が低かった。(図 11C)。また、これらの結果はヒトにおいても同様に観察された（図 12)。以上の結果から、CDS や 3′ UTR などのエキソン上に存在する TSC の多くは、切断された RNA 由来の偽の TSC である可能性が示唆された。転写された RNA がエキソン上で広範囲に切断されたと仮定すると、その切断された RNA の 5′ 末端に由来するタグは、TSS 下流に存在するエキソン領域上に広くマップされるはずであり、これによってエキソン上に TSC が形成される。さらに、エキソン上の TSC がスプライスアクセプター部位近傍にピークを頻繁に持つことを発見し（図 13)、それらの多くが図 14 のように右歪曲な TSS 分布を示すことも分かった。この結果は、アクセプター部位の近くが他の場所と比較して切断されやすいことを示唆しているのかもしれない。本研究では、エキソン上に存在する TSC の多くは切断された RNA 由来の TSC である可能性があるため可能な限り除去することとした（付録 A.3 参照)。

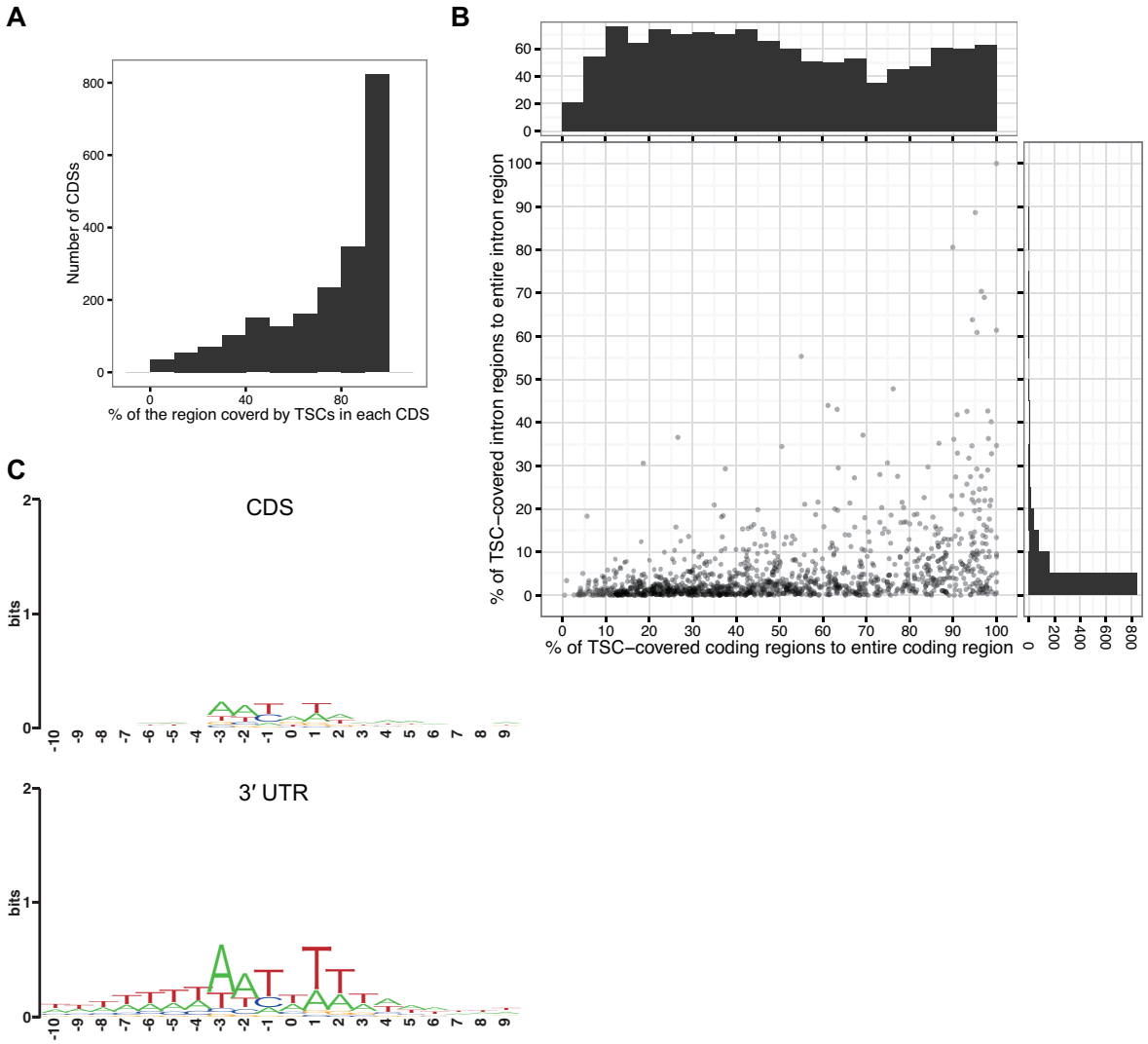

図 11: カタユウレイボヤにおける CDS と 3′ UTR に存在する TSC の解析。(A) TSC でカバーされる領域の割合。少なくとも一つの TSC が存在する CDS に対して、TSC でカバーされる領域の割合を計算した。およそ 80% の CDS において、半分以上の領域が TSC でカバーされていた。(B) 全体の CDS とイントロン領域に対する TSC でカバーされる領域の割合。少なくとも 1 つの TSC を CDS 上にもつ転写産物モデルに対して、全体の CDS と全体のイントロンに対する TSC でカバーされる領域の割合を調べた。各ドットは少なくとも 1 つの TSC を CDS 領域にもつ転写産物モデルを表す。イントロン領域を持たないモデルは図の中に含まれない。(C) CDS と 3′ UTR 上の存在する TSC の sequence logo。x 軸は、代表 TSS からの距離を表す。

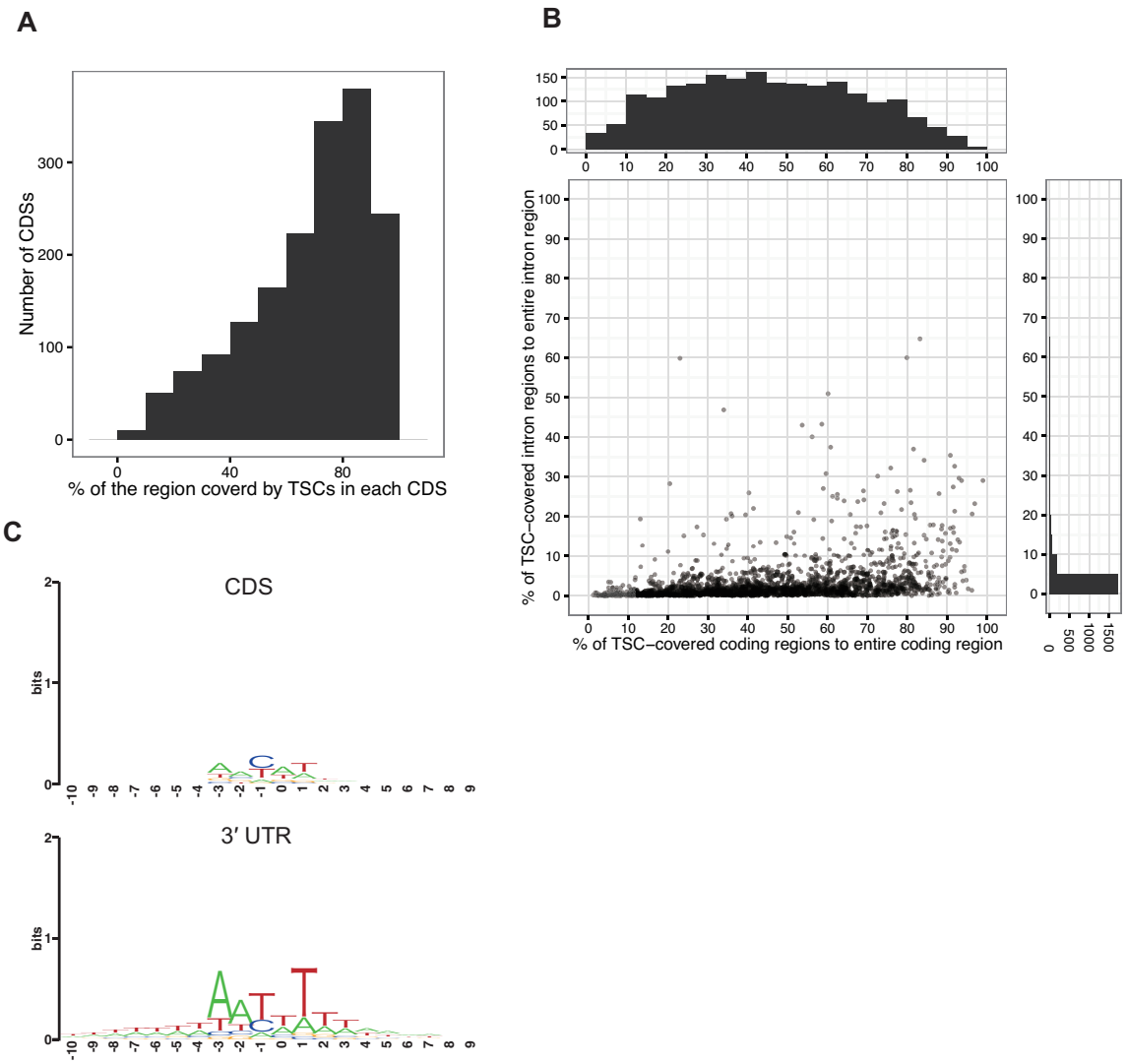

図 12: ヒトにおける CDS と 3′ UTR に存在する TSC の解析。(A) TSC でカバーされる領域の割合。少なくとも一つの TSC が存在する CDS に対して、TSC でカバーされる領域の割合を計算した。(B) 全体の CDS とイントロン領域に対する TSC でカバーされる領域の割合。少なくとも 1 つの TSC を CDS にもつ転写産物モデルに対して、全体の CDS と全体のイントロンに対する TSC でカバーされる領域の割合を調べた。各ドットは少なくとも 1 つの TSC を CDS 領域にもつ転写産物モデルを表す。イントロン領域を持たないモデルは図の中に含まれない。(C) CDS と 3′ UTR 上の存在する TSC の sequence logo。x 軸は、代表 TSS からの距離を表す。

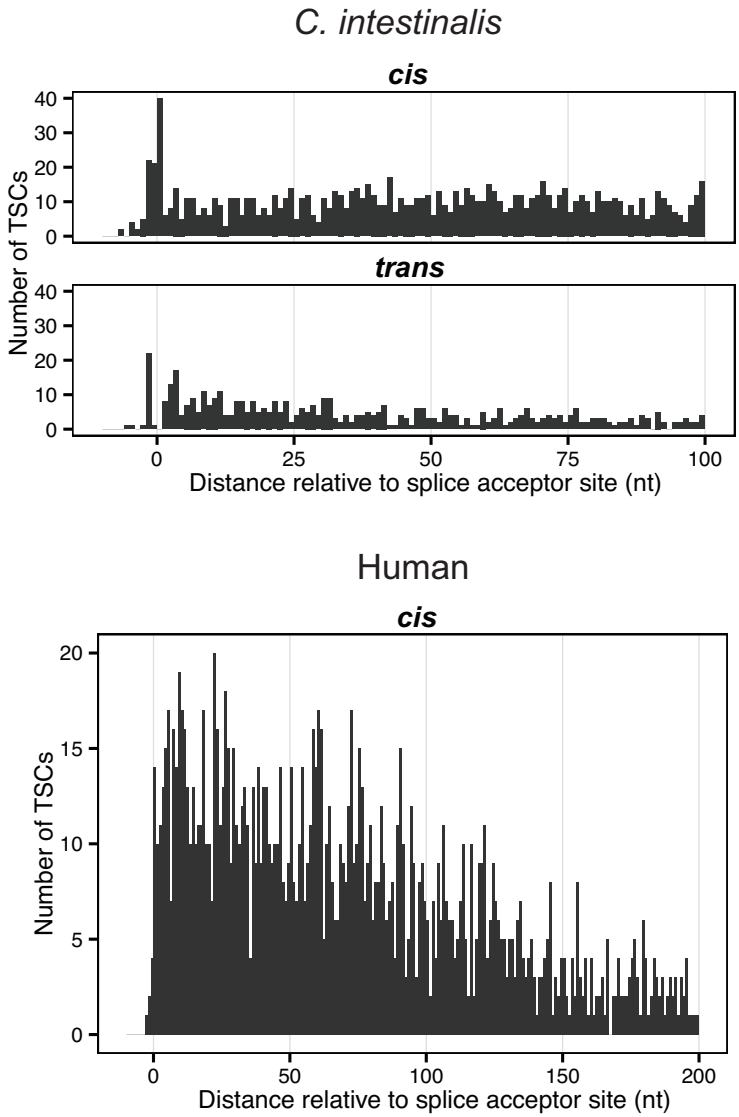

図 13: エキソン上の TSC のピーク位置。エキソン上に存在する各 TSC に対して、最も頻度の高い TSS（ピーク）の位置を調べた。x 軸と y 軸はそれぞれ、シスもしくはトランススプライスアクセプター部位からの距離と TSC の数を表す。TSC のピークは頻繁にスプライスアクセプター部位付近に位置していた。

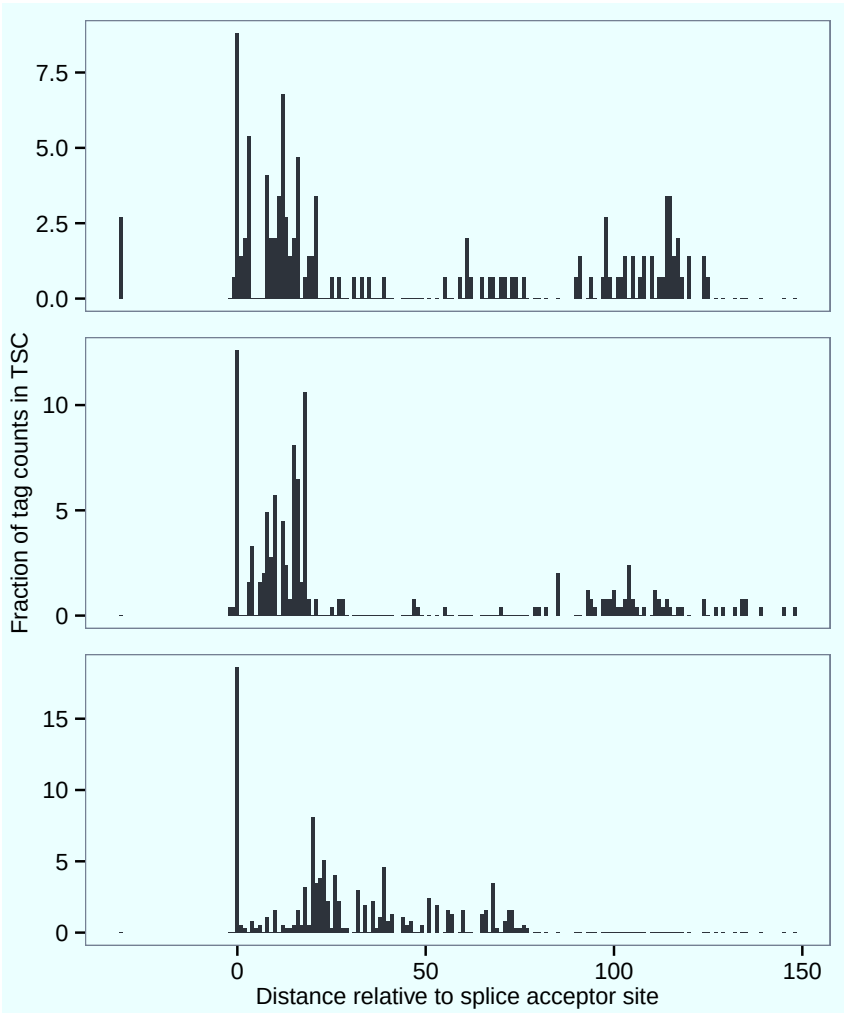

図 14: スプライスアクセプター部位付近にピークをもつ右歪曲な TSC の例。x 軸と y 軸はそれぞれ、スプライスアクセプター部位からの距離と TSC 中に含まれるタグの割合を表す。

4.6 既知コアプロモーターモチーフの分布

発見した 3 タイプの幅 1 bp の TSC と切断された RNA 由来だと思われる TSC を除去した後、カタユウレイボヤにおいて 79 個の RP TSC に加えて 1844 個の既知 TSS 上に存在する TSC を得た（表 7)。これら 1844 個の TSC は遺伝子モデルの既知 TSS 上に存在するため、信頼できる TSC であり既知のプロモーターを表していると考えられる。そこで、これら 79 個の RP 遺伝子を含む計 1923 個の TSC を用いてプロモーターの性質を調べることとした。同様に、ヒトにおいて 5073 個の既知 TSS 上に存在する TSC を得た（表 8)。これらの TSC は、カタユウレイボヤとヒトのプロモーターの性質を調べるために用いられた。ほとんど全ての TSC の幅は 100 bp 以内であり（図 15)、この結果は哺乳類のプロモーターを調べた先行研究の結果 (Forrest et al., 2014) と一致した。

カタユウレイボヤプロモーターにおいて、最もよく知られているコアプロモーターモチーフである TATA box の位置を調べた。TATA box の位置は TRANSFAC の MATCH を用いて予測された（付録 A.5 参照)。カタユウレイボヤでは、ヒトと同様に TATA box は −32 から −29 の位置に高頻度に存在していた（図 16)。この結果に基づいて、TATA box を −32 から −29 の位置に持つプロモーターを TATA-containing プロモーターと定義した。また、TATA box 以外の既知コアプロモーターモチーフ（BRE^{d}、BRE^{u}、DPE、DCE S_{I}、DCE S_{II}、DCE S_{III}、MTE、 XCPE1）(Juven-Gershon et al., 2008) と DRE、motif 1、6、7 に対しても同様に調べてみたが、TATA box のように既知の位置に明確なピークは示さなかった（図 17)。

表 7: カタユウレイボヤにおける 3 タイプの幅 1 bp の TSC と切断された RNA 由来と思われる TSC 除去後の TSC の数。TSC は位置に基づいて 7 つのカテゴリーに分類された（第 3.5 節参照)。ただし、79 個の RP 遺伝子の TSC は、たとえ既知 TSS 上に存在していなくても便宜上「TSS」に含むこととした。

Location	TSCs
TSS	1923 (44.0%)
TAS	57 (1.3%)
5′ UTR	260 (6.0%)
CDS	0 (0.0%)
3′ UTR	0 (0.0%)
intron	487 (11.2%)
intergenic	1639 (37.5%)
total	4366 (100%)

表 8: ヒトおける 3 タイプの幅 1 bp の TSC と切断された RNA 由来と思われる TSC 除去後の TSC の数。

Location	TSCs
TSS	5073 (44.3%)
5′ UTR	1336 (11.7%)
CDS	0 (0.0%)
3′ UTR	0 (0.0%)
exon(ncRNA)	244 (2.1%)
intron	1549 (13.5%)
intergenic	3247 (28.4%)
total	11449 (100%)

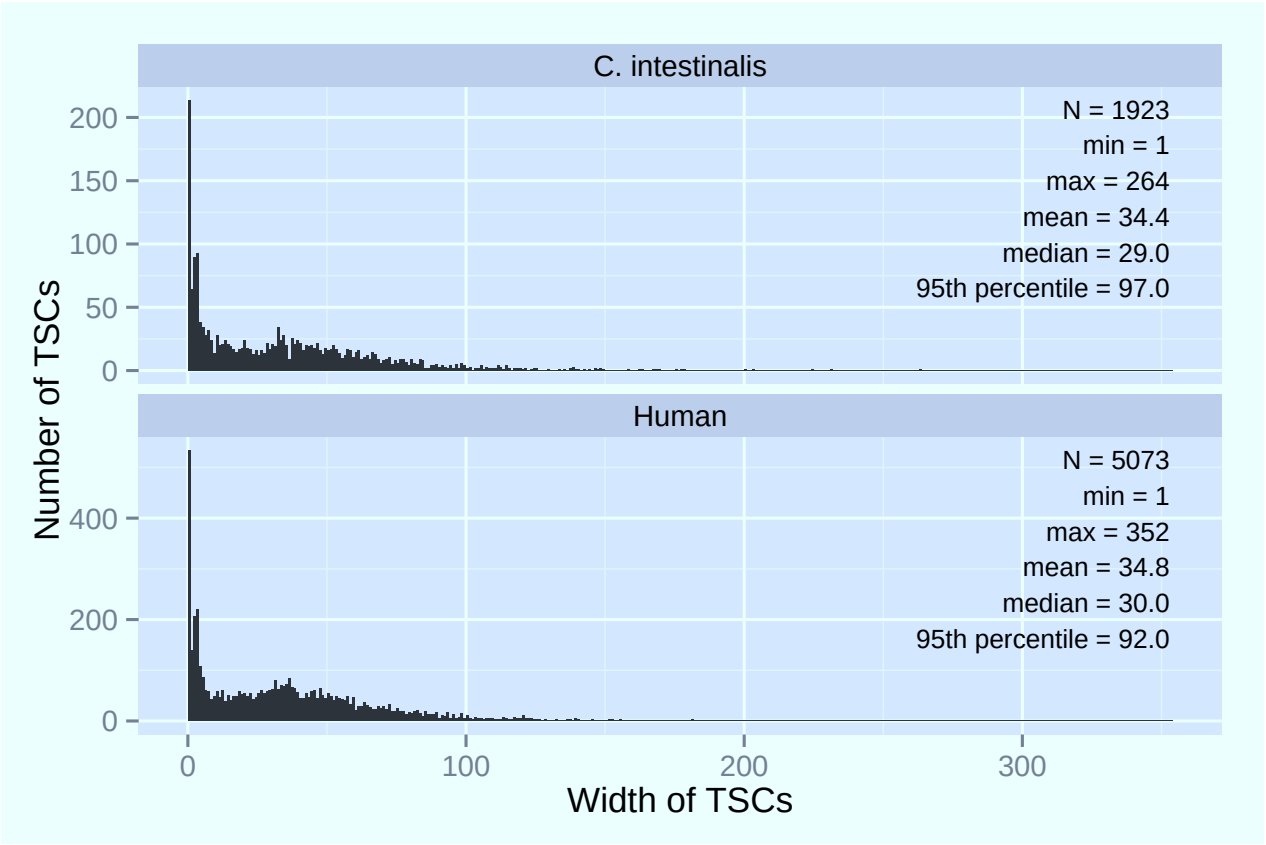

図 15: 既知 TSS 上に存在する TSC の幅の分布。

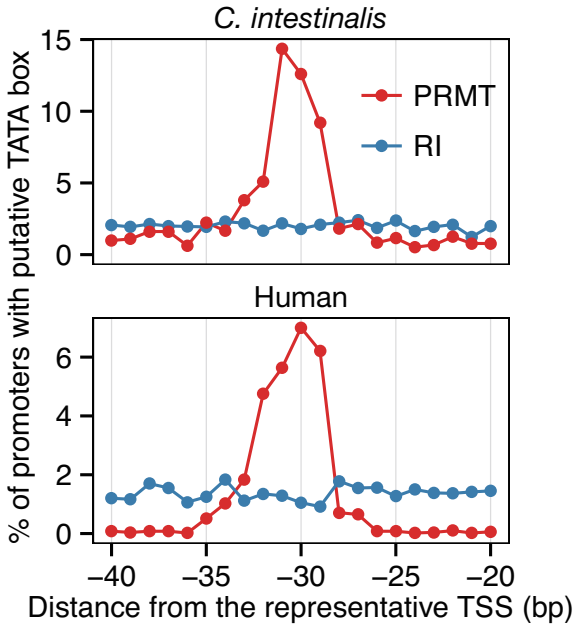

図 16: TATA box の位置。推定された TATA box の位置をコアプロモーター領域において調べた。x 軸と y 軸はそれぞれ、代表 TSS からの距離と各位置に TATA box をもつプロモーターの割合を表す。プロモーター配列（PRMT）だけでなく、ランダムに取得された遺伝子間領域（RI）の配列における推定された TATA box の位置も示した（青線）。

図 17: コアプロモーターモチーフの分布。カタユウレイボヤのコアプロモーター領域（−60 to +39）上のコアプロモーターモチーフを FIMO (Grant et al., 2011) を用いて探索した（デフォルトパラメータを使用）。JASPAR に登録されている既知コアプロモーターモチーフ（TATA-box、BRE^{d}、BRE^{u}、DPE、DCE S_{I}、DCE S_{II}、DCE S_{III}、MTE、XCPE1）の位置重み行列（PWM）を探索に使用した。DRE と motif 1、6、7 の PWM は文献 (Ni et al., 2010) から得た。また、コアプロモーター（PRMT）と同数のランダム遺伝子間配列（RI）に対しても同様に探索を行った（青線）。探索の結果、3 つの DCE モチーフはコアプロモーター配列上に存在していなかったため図には示していない。x 軸と y 軸はそれぞれ、代表 TSS からの距離と各位置にモチーフをもつプロモーターの割合を表す。

4.7 RP 遺伝子プロモーターの性質

第 4.3 節で同定した 79 個の RP 遺伝子の TSC を用いてカタユウレイボヤの RP 遺伝子プロモーターの性質を調べた。また、ヒト RP 遺伝子の TSS は既に解析が行われているので、それらをヒト RP 遺伝子プロモーターの TSS として用いることとした (Perry, 2005)。

第 4.3 節で述べたように、79 個の RP 遺伝子全てに対して、シャープな TSS 分布を持ちピリミジンに富む配列上に存在する TSC を同定した（図 18A、図 S7-S8)。ただし、*Rplp1* プロモーターだけが幾分異なる TSS 分布を示した（図 18A)。このプロモーターはお互いに 30 bp ほど離れた 2 つのポリピリミジンイニシエーターモチーフを持っており、転写は両方のポリピリミジンイニシエーターモチーフのシトシン塩基から始まっていた（図 18A)。このタイプの RP 遺伝子プロモーターはヒト RP 遺伝子プロモーター（e.g., *RPL39*）でも観察されており、離れたポリピリミジンイニシエーターモチーフ内の異なるシトシンから転写が始まる (Yoshihama et al., 2002)。

同定された 79 個の RP 遺伝子プロモーター全てがポリピリミジンモチーフを持っており、*Rpl22* プロモーター以外において代表 TSS はシトシン塩基であった。*Rpl22* プロモーターでは、代表 TSS はポリピリミジンモチーフのシトシン塩基の 2 塩基下流のチミン塩基に位置していた（図 19)。また、カタユウレイボヤのポリピリミジンイニシエーターモチーフは、ヒトと同様によく保存されていた（図 18B)。カタユウレイボヤとヒト間で RP 遺伝子プロモーターのイニシエーターモチーフを比較したところ、カタユウレイボヤの方がより厳格に保存されているようであった。カタユウレイボヤでは、中心にある −1 から +4 の 6 塩基のポリピリミジン配列が非常によく保存されており、特に +3 の位置は強くチミン塩基で保存されていた。

RP 遺伝子プロモーターの TATA box の有無を調べた。ヒトでは、16 個の RP 遺伝子プロモーターが TATA box を持っていたが、カタユウレイボヤでは 79 個の内 2 個しか TATA box を持っていなかった。また、TATA-less RP 遺伝子プロモーターの上流 30 付近の AT の豊富さを調べたところ、ヒトでは TATA-less RP 遺伝子プロモーターは、TATA-containing RP 遺伝子プロモーターと同様に −30 付近で高い AT 含量を示した（図 18C)。このことは、ヒト RP 遺伝子プロモーターが既存研究でも報告されているように TATA box もしくは TATA-like な配列を持っていることを示唆している (Yoshihama et al., 2002; Perry, 2005)。一方、カタユウレイボヤでは TATA-less RP 遺伝子プロモーターは −30 付近で高い AT 含量を示さなかった（図 18C)。

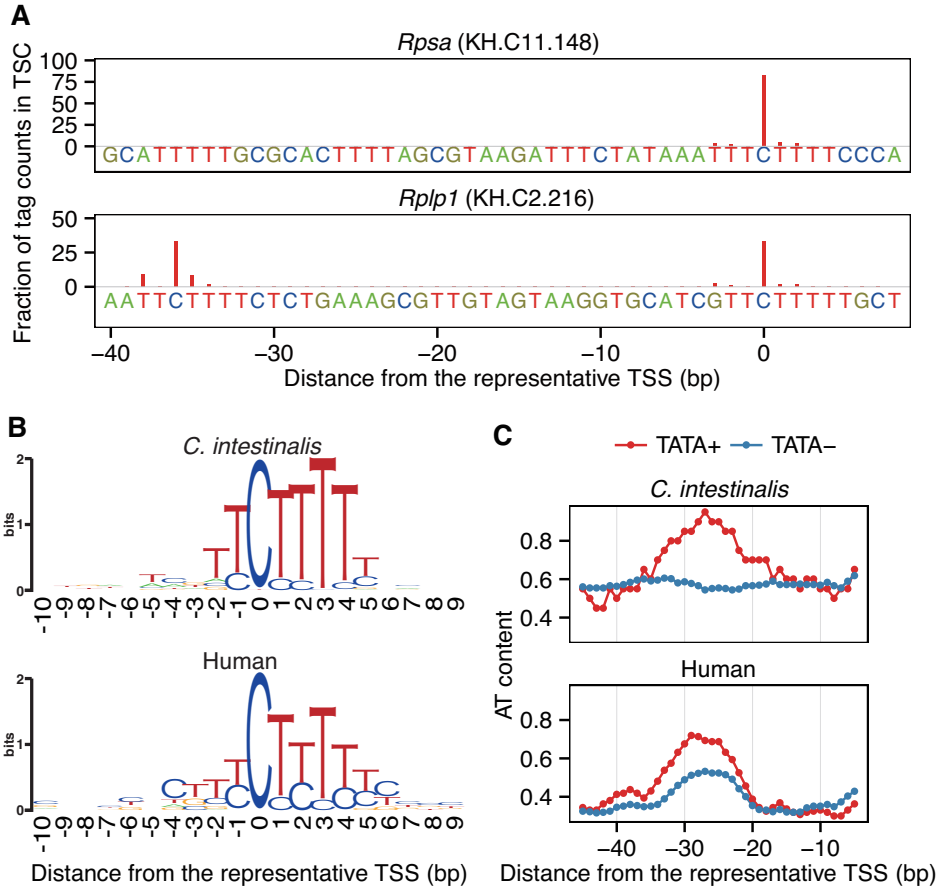

図 18: RP 遺伝子プロモーターの性質。(A) RP 遺伝子プロモーターの TSS 分布。例として、*Rpsa* と *Rplp1* プロモーターの TSS 分布を示した。(B) RP 遺伝子プロモーターのポリピリミジンイニシエーターモチーフ。(C) RP 遺伝子プロモーターの AT 含量の分布。AT 含量は 10-bp スライディングウィンドウを用いて計算された。TATA+ と TATA− はそれぞれ TATA-containing RP 遺伝子プロモーターと TATA-less RP 遺伝子プロモーターを表す。

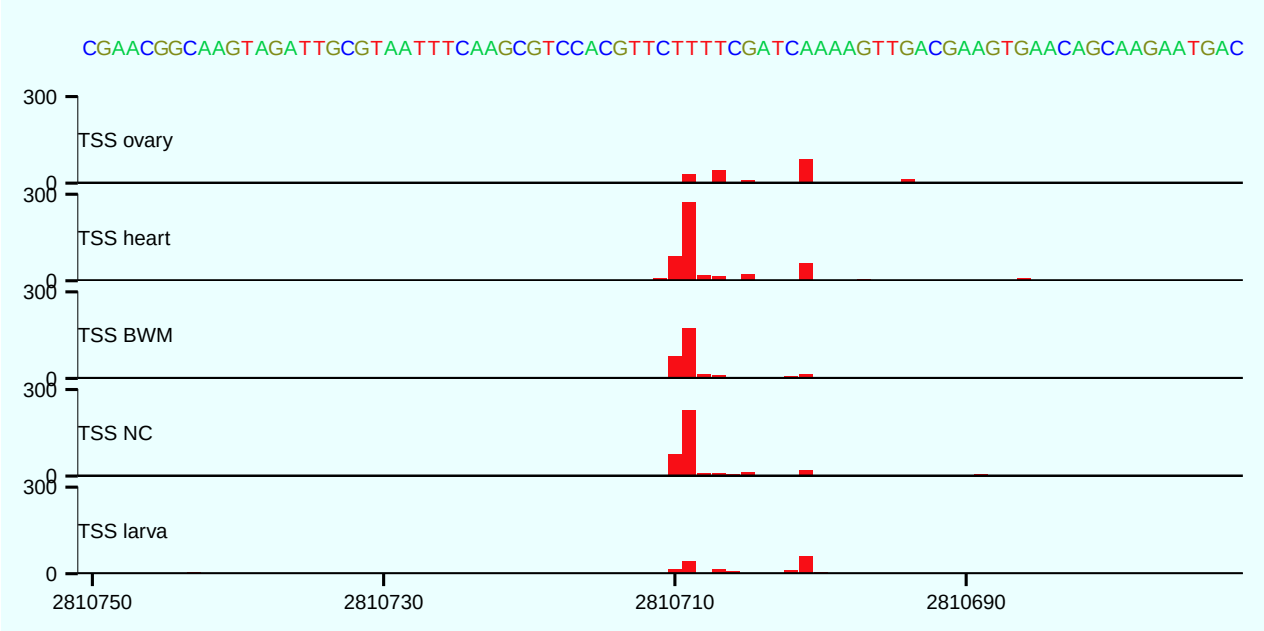

図 19: *Rpl22* プロモーターの TSS。赤いバーは TSS を表し、y 軸はタグ数を表す。BWM, body wall muscle (体壁筋); NC, neural complex (神経複合体).

4.8 非 RP 遺伝子プロモーターの性質

コアプロモーターエレメントと TSS 分布の観点から非 RP（non-RP）遺伝子プロモーターの性質を調べた。この解析は、既知 TSS 上に存在する TSC の内、non-RP 遺伝子の既知 TSS 上に存在する 1844 個の TSC を用いて行われた。また、ヒトに対しても non-RP 遺伝子の既知 TSS 上に存在する 5000 個の TSC を用いて同様の解析を行った。まず、TSS 分布と TATA box の関連を調べるために、non-RP 遺伝子プロモーターを TSS 分布のタイプと TATA box の有無に基づいて分類した。non-RP 遺伝子プロモーターは、TSS 分布に基づいて “sharp”、“broad”、“other” の 3 タイプに分類された（図 20）。sharp-type プロモーターとは、転写が狭い領域で始まり、シャープな TSS 分布を持つプロモーターである。一方、broad-type プロモーターとは、転写が広い範囲で始まり、明確なピークをもたない TSS 分布を持つプロモーターである。other グループは “sharp” にも “broad” にも属さないプロモーターである。non-RP 遺伝子プロモーターはさらに、TATA box の有無に基づいて TATA-containing プロモーターと TATA-less プロモーターに分類された（表 9）。その結果、ほとんどの TATA-containing プロモーターはシャープな TSS 分布を示すことがわかった。また、ほとんど全ての broad-type プロモーターは TATA-less プロモーターであった。これらの結果は、TATA 結合タンパク質が転写開始の正確な位置に決定に関与しているという事実と一致するものである (Baumann et al., 2010)。しかしながら、シャープな TSS 分布をもつ TATA-less プロモーターも多数存在し、TATA box がないからといって必ずしもブロードな TSS 分布を示す訳ではなかった。本研究では、3 つの主要なプロモータークラス（TATA-containing sharp-type プロモーター、TATA-less sharp-type プロモーター、TATA-less broad-type プロモーター）の性質を調べた。

カタユウレイボヤのプロモーターにおいて、どの 2 塩基が TSS として使われているのかを調べた。その結果、どのプロモータークラスにおいても、3 つの PyPu 塩基（CA、TA、TG）がゲノム上での出現頻度と比較して有意に高頻度に TSS として使用されていることが分かった（$P < 0.01$、二項検定、図 21）。この性質は、カタユウレイボヤとヒト間で保存されていた。ただし、ヒトでは CG の使用頻度も高かった。また、4 つの PyPu 塩基の内、CA と TA はカタユウレイボヤにおいてより高頻度に使用されていたが、ヒトでは CA のみが優先的に使用されていることが分かった。さらに、TSS としての 2 塩基の使用率は、sharp-type と broad-type で差があることも分かった。CA の使用頻度は broad-type よりも sharp-type で高かったが、他の PyPu 塩基（TA、TG、CG）は sharp-type よりも broad-type での使用頻度が高かった。この差はカタユウレイボヤとヒト間で保存されているように見えた（図 22）。

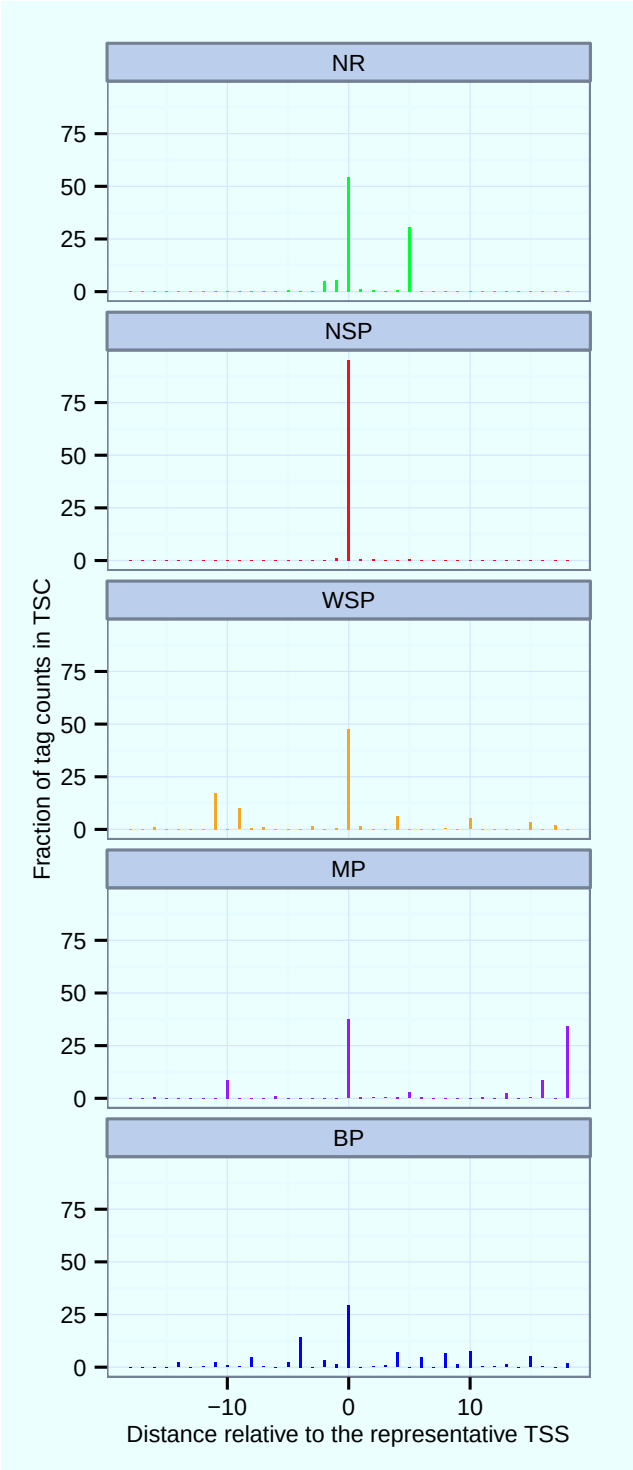

図 20: TSS 分布の例。プロモーターは、TSS 分布に基づいて 5 つのタイプ（NR、NSP、WSP、MP、BP）に分類された（付録 A.6 参照）。各パネルは TSS 分布の各タイプの例を示す。NSP プロモーターと BP プロモーターは、それぞれ “sharp-type” プロモーターと “broad-type” プロモーターと呼ぶ。その他プロモーターは “other” プロモーターとして統合された。

表 9: TATA box と TSS 分布の関係。non-RP 遺伝子プロモーターを TATA box の有無と TSS 分布のタイプに基づいて 6 つのクラスに分類した。TATA+ と TATA− はそれぞれ TATA-containing と TATA-less プロモーターを表す。

	C. intestinalis			Human		
	TATA+	TATA−	total	TATA+	TATA−	Total
Sharp	229	301	530	388	1015	1403
Broad	27	700	727	43	1933	1976
Other	32	555	587	61	1560	1621
Total	288	1556	1844	492	4508	5000

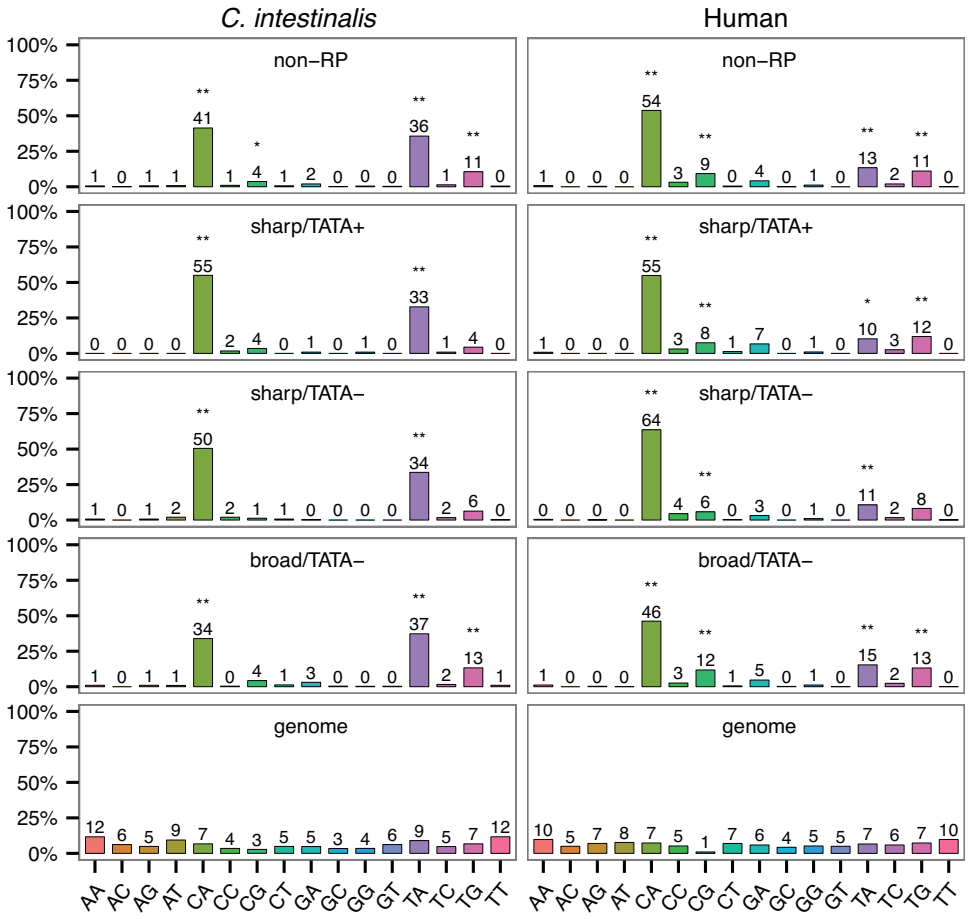

図 21: TSS として使われる 2 塩基の頻度。non-RP 遺伝子プロモーターにおいて、$-1, 0$ 位置にある 2 塩基の頻度を調べた。ここで 0 位置とは代表 TSS を表す。各バーば各 2 塩基の使用割合を示す。どのプロモータークラスでどの 2 塩基がよく使用されているのかを調べるため、各プロモータークラスでの割合とゲノム上での出現頻度の割合の差を二項検定で評価した。アスタリスク（*と**）は Bonferroni 補正された後の P 値（$P < 0.05$ と $P < 0.01$）を表す。

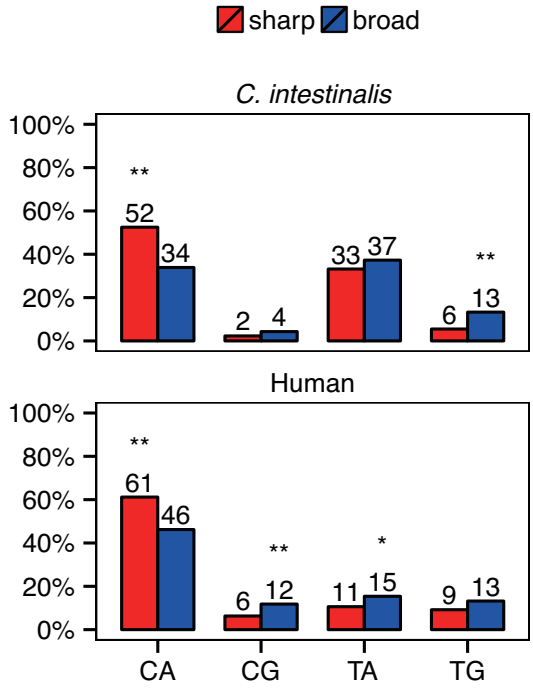

図 22: sharp-type と broad-type 間での TSS 使用率の差。4 つの PyPu 塩基の使用率を sharp-type と broad-type 間で比較した。赤と青のバーはそれぞれ、sharp-type と broad-type での各 PyPu の使用率を表す。差はフィッシャーの直接確率検定で評価された。アスタリスク（*と**）は Bonferroni 補正された後の P 値（$P < 0.05$ と $P < 0.01$）を表す。

次にプロモータークラスと CpG 塩基との関連を調べた。ヒトでは broad-type プロモーターと TATA-less プロモーターが CpG アイランドやユビキタスに発現する遺伝子と関連があるということが知られている (Carninci et al., 2006; Yang et al., 2007)。実際、TATA-less プロモーターは TATA-containing プロモーターよりも有意に高い CpG 含量と低発現特異性を示した（$P < 0.01$、マン・ホイットニーの U 検定、図 23)。一方、カタユウレイボヤでは CpG アイランドはないと言われている (Okamura et al., 2011)。そこで、各プロモータークラスにおいて CpG 含量に差があるかどうかを調べた。その結果、3 つのプロモータークラスにおいて CpG 含量の有意な差は観察されなかった。このことは、カタユウレイボヤにおいて、TATA box も TSS 分布も CpG には関連がないことを示唆している。それにもかかわらず、カタユウレイボヤ TATA-less プロモーターは、ヒト TATA-less プロモーターと同様に TATA-containing プロモーターよりも低い発現特異性を示した（$P < 0.01$、マン・ホイットニーの U 検定、図 23)。

CAGE 法を用いた最近のヒトプロモーターの解析によると、broad-type プロモーターは sharp-type プロモーターよりも正確なヌクレオソームポジショニングを持ち、+1 ヌクレオソームの位置において WW モチーフの分布が 10.5 bp 周期を示す (Forrest et al., 2014)。実際、TSS-seq のデータを用いた本研究でも（sequencing depth が低いため先行研究の結果よりは明確に観察される訳ではないが）ヒト broad-type プロモーターにおいて同様の WW モチーフの周期分布が観察された（図 24)。また、同様の分布はカタユウレイボヤの broad-type プロモーターでも観察され、特に +120 から +210 の領域の周期性は sharp-type プロモーターよりも明確であった（図 24)。この結果は、カタユウレイボヤにおいても broad-type プロモーターは sharp-type プロモーターよりも正確なヌクレオソームポジショニングを持つことを示唆している。

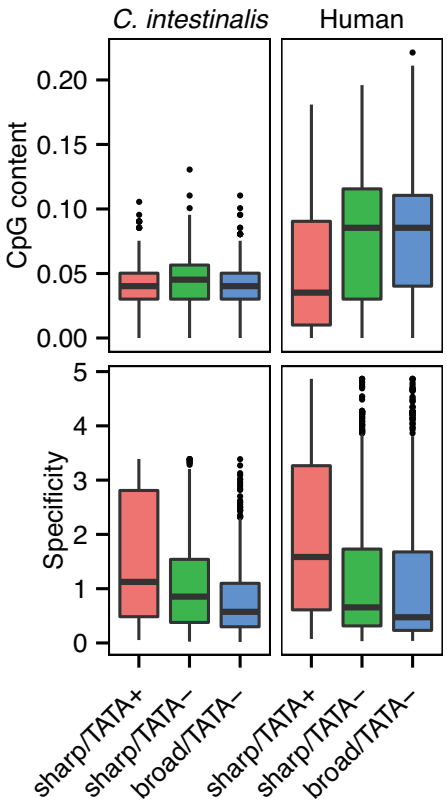

図 23: CpG 含量と発現特異性。各プロモータークラスの各コアプロモーター配列（-100 から +99 の 200 塩基）の CpG 含量（CpG の数 / $(200 - 1)$）を計算した。発現特異性は相対エントロピーによって評価された（付録 A.7 参照）。

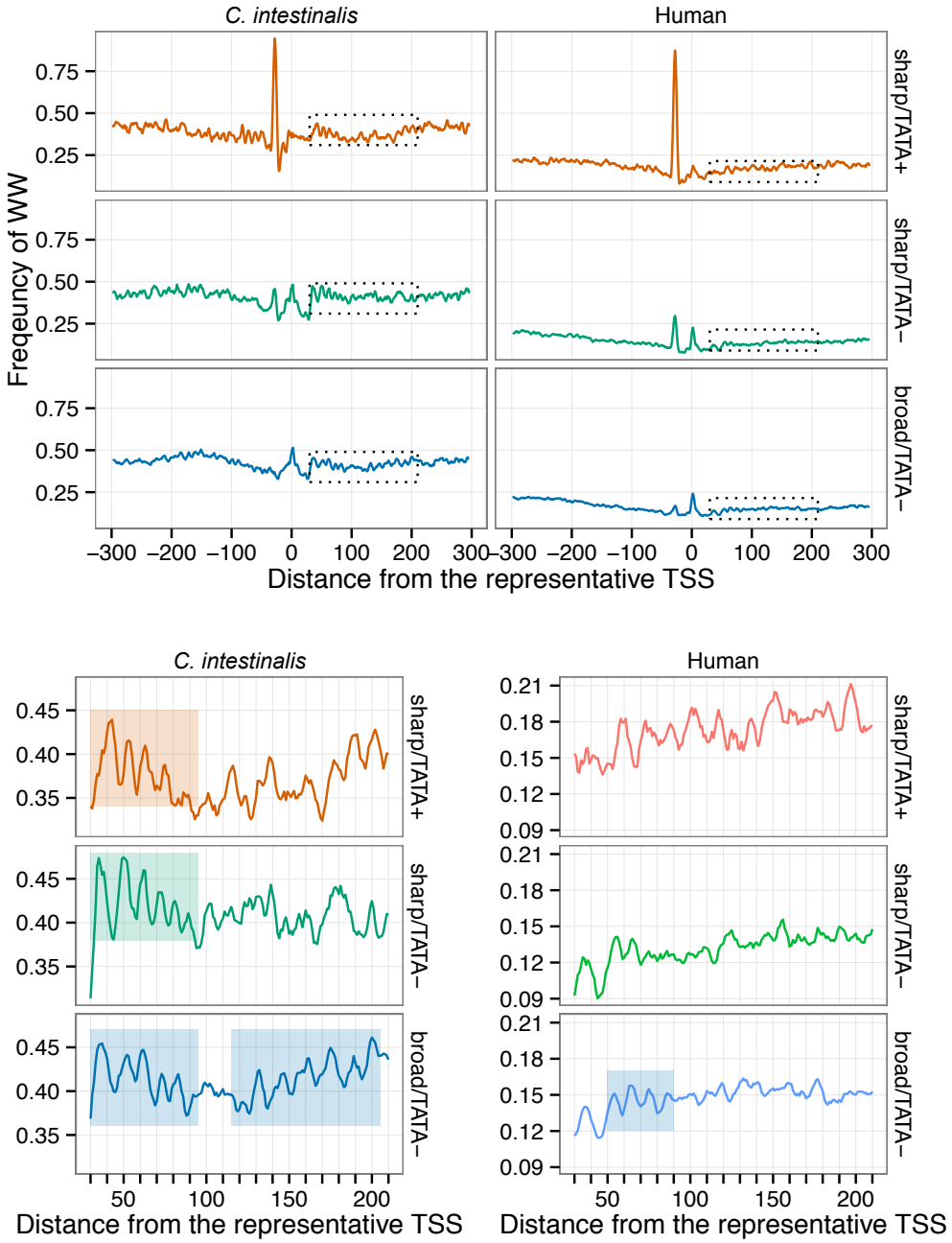

図 24: WW モチーフの頻度分布。上図は各プロモータークラス（sharp/TATA+、sharp/TATA−、broad/TATA−）における WW（W とは A もしくは T）モチーフの頻度を示す。各分布は 5 point のスライディングアベレージを用いて平滑化された。下図は +30 から +210 領域（上図において点線で囲まれた領域）の拡大図を表す。また、半透明な四角で覆われる部分は約 10-bp の周期性を明確に示す領域を表す。

4.9 推定プロモーターの同定

前節のプロモーター解析で用いた既知 TSS 上に存在する TSC の他に、5′ UTR やイントロン、遺伝子間領域にも多くの TSC が存在した（表 7)。これらの TSC は既知 TSS と一致していなので、既知 TSS 上に存在する TSC よりは信頼性が低いと考えられる。これらの TSC のクオリティーを確認するために、イニシエーターモチーフを調べた。その結果、比較的保存された PyPu モチーフを持つことが分かった（図 25)。この結果はこれらの TSC の多くが真の TSC であることを示唆している。今後の解析のために可能な限り信頼性を高めるため、ピーク TSS に TA、CA、TG を持つ TSC だけを選択した。ここでピーク TSS とは、最も頻度の高い TSS の頻度の 2 分の 1 以上の頻度をもつ TSS のことである。選択された TSC は、現在の遺伝子モデル上にはアノテーションされていない新たに同定された推定プロモーターと考えられる。多くの推定プロモーターが 5′ UTR やイントロン、遺伝子間領域に発見された。幾つかの推定プロモーターは TAS とオーバーラップしていた（図 26)。今後の解析では、79 個の RP TSC を含む既知 TSS 上の TSC と選択された TSC のセットを最終 TSC セットと呼ぶ（表 10)。

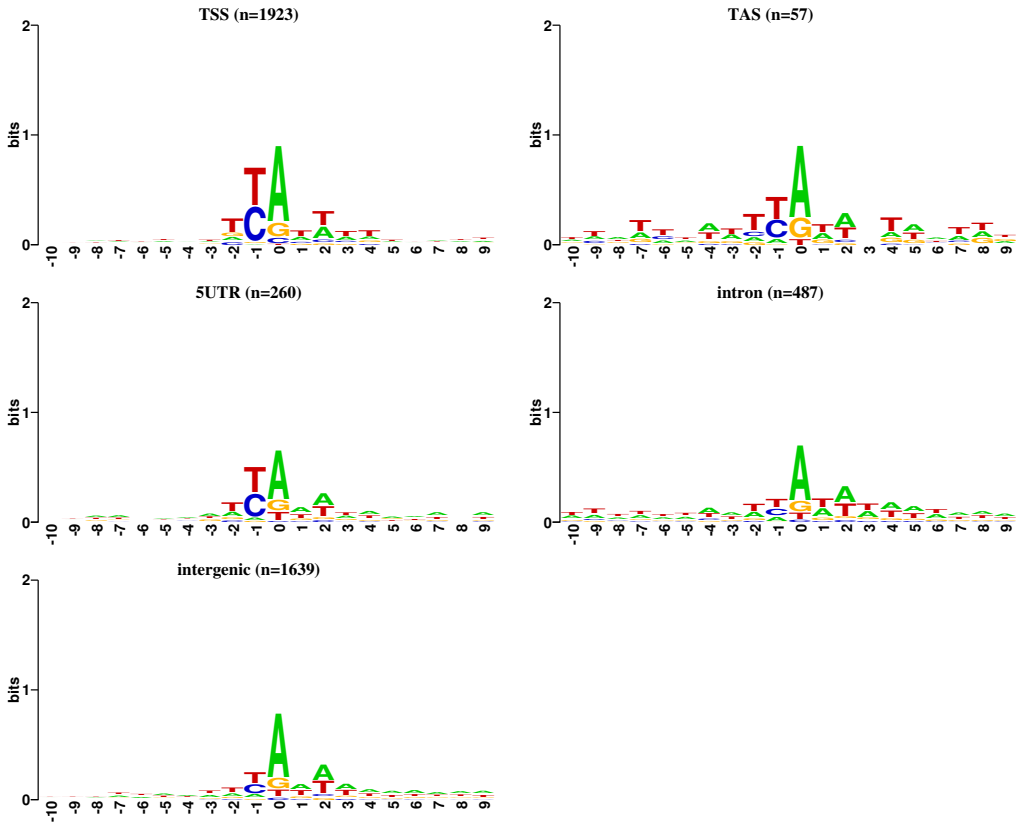

図 25: イニシエーターモチーフ。各位置に存在する TSC のイニシエーターモチーフを示した。x 軸は代表 TSS からの距離を表す。括弧内の数字は TSC の数を表す。

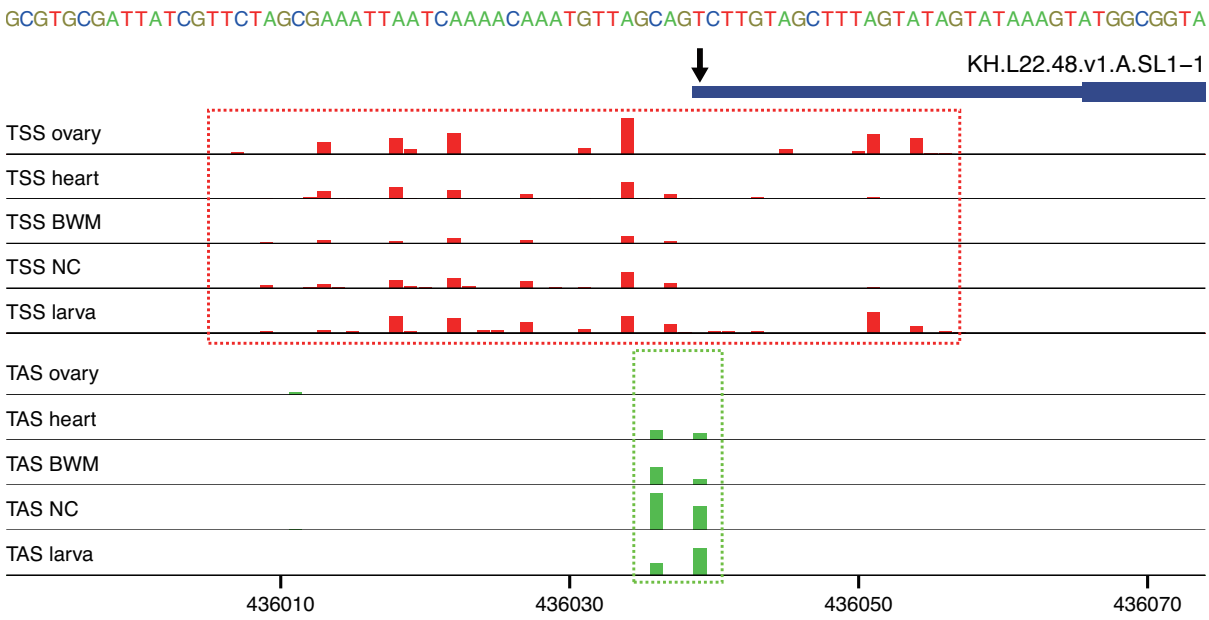

図 26: TAS とオーバーラップする TSC の例。図は piggyBac transposable element derived 4 をコードする転写産物（KH.L22.48.v1.A.SL1-1）のアノテーションされている TAS にオーバーラップする TAC と TSC を示す。赤と緑の四角はそれぞれ TSC と TAC を表す。また、赤と緑のバーはそれぞれ TSS と TAS の分布を表す。矢印はアノテーションされている TAS を示す。BWM と NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

表 10: カタユウレイボヤにおいて除去された TSC の数と最終 TSC セットの TSC の数。各列は右から順に「TSC の位置」,「初期 TSC セットの TSC 数」、「CTGG TSC の数」、「A+T-rich な幅 1-bp の TSC の数」、「 逆鎖のスプライスドナーサイト付近にある幅 1-bp の TSC の数」、「CDS もしくは 3′ UTR 上に存在する切断された TSC 由来の可能性がある TSC の数」、「PyPu (TA、CA、TG）モチーフを持たない TSC の数」、 「除去された TSC の総数」、「最終的に残った TSC の数（最終 TSC セット)」を表す。括弧の中の数値は、初期 TSC に対する各 TSC の割合を表す。

Location	Initial	CTGG	A+T-rich	Donor	CDS+3′ UTR	Non-PyPu	Removed	Final
TSS	2097 (100)	11 (0.5)	-	0 (0)	163 (7.8)	-	174 (8.3)	1923 (91.7)
TAS	122 (100)	0 (0)	-	0 (0)	65 (53.3)	25 (20.5)	90 (73.8)	32 (26.2)
5′ UTR	420 (100)	16 (3.8)	-	0 (0)	144 (34.3)	117 (27.9)	277 (66.0)	143 (34.0)
CDS	1623 (100)	18 (1.1)	-	0 (0)	1605 (98.9)	-	1623 (100)	0 (0.0)
3′ UTR	1459 (100)	25 (1.7)	-	0 (0)	1434 (98.3)	-	1459 (100)	0 (0.0)
intron	721 (100)	24 (3.3)	159 (22.1)	0 (0)	51 (7.1)	248 (34.4)	482 (66.9)	239 (33.1)
intergenic	3350 (100)	1339 (40.0)	328 (9.8)	34 (1.0)	10 (0.3)	770 (23.0)	2481 (74.1)	869 (25.9)
total	9792 (100)	1433 (14.6)	487 (5.0)	34 (0.3)	3472 (35.5)	1160 (11.8)	6586 (67.3)	3206 (32.7)

4.10 SL *trans*-spliced 遺伝子のプロモーター候補

SL *trans*-spliced 遺伝子のプロモーター候補を予測した。トランススプライシングを受ける遺伝子の中には、頻繁にトランススプライシングを受ける遺伝子とそうでない遺伝子がある。このことを考えると、頻繁にトランススプライシングを受けない遺伝子と頻繁にトランススプライシングを受けるが高発現している遺伝子に対しては、その 5′ 末端に TAC があるだけではなく、上流に TSC もあることが期待される。また、この場合、両方のクラスターは同じ遺伝子に由来するものなので、その発現特異性は同じである可能性がある。そこで、同じサンプルで有意に高発現しているクラスターのペア（TSC と TAC のペア）を探索することで、SL *trans*-spliced 遺伝子のプロモーター候補を予測した。発現の有意さは相対エントロピーと超幾何分布による検定によって評価した（付録 A.8 参照)。TAC と上流の TSC の距離、すなわちアウトロンの長さの閾値は 51 bp から 2000 bp とした。この下限値（51 bp）は、線虫において上流の 3′ スプライスサイトに結合させた 51 bp 以上の AU-rich な合成 RNA が効率的なトランススプライシングをもたらすという報告 (Conrad et al., 1995) に基づいている。また、上限値（2000 bp）は、最近線虫において同定されたアウトロンのほとんど（90% 以上）が 2000 bp 以内であったことに由来する (Kruesi et al., 2013)。本研究では、アウトロンとは “non-operon-type” の *trans*-spliced 遺伝子において取り除かれる 5′ 末端領域のことを意味する。したがって、“opeon-type” の *trans*-spliced 遺伝子に対応するクラスターペアは探索には含まれない。

探索の結果、同じ発現特異性をもつクラスターペアを 264 個発見した。これらのペアは 2 個の unannotated-operon-type のペアと 262 個の non-operon-type のペアに分類された（付録 A.9 参照)。unannotated-operon-type のペアは現在の遺伝子モデルではアノテーションされてないオペロンを示している可能性がある。non-operon-type ペアの TAC と上流の TSC 間の距離の平均、つまり、推定アウトロンの平均は 438 bp であった（図 27)。アウトロンの長さの分布のピークは 100 bp 付近にあり、多くアウトロンの長さは 500 bp 以内であった。これらのアウトロンの長さの性質は線虫と同様であった (Kruesi et al., 2013)。また、推定アウトロンの塩基組成を調べてみたところ、イントロンと同様の A、T、C 含量を示す一方で、イントロンよりも有意に高い G 含量を示した（$P < 0.01$、マン・ホイットニーの U 検定、図 28)。

262 個の non-operon-type ペアは、262 個の TAC と 233 個の TSC から成っていた。233 個の TSC は 72 個の既知プロモーターと 161 個の推定プロモーターに対応した。この 161 個の推定プロモーターは新規の SL *trans*-spliced 遺伝子のプロモーター候補と考えられる。図 29 は、*Heat shock protein beta-1* をコードする SL *trans*-spliced 遺伝子（KH.S455.4.v1.A.SL1-1）のプロモーター候補を示している。既知 TAS 上の TAC だけでなく、上流 489 bp に発現特異性が同じである TSC が存在しており、この TSC は heat shock protein beta-1 の転写開始点候補と考えられる。

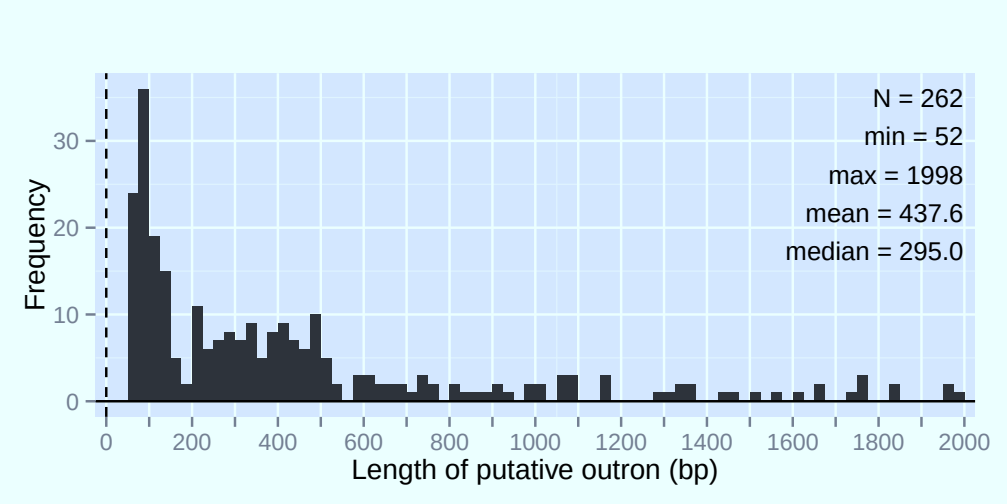

図 27: 推定アウトロンの長さの分布。

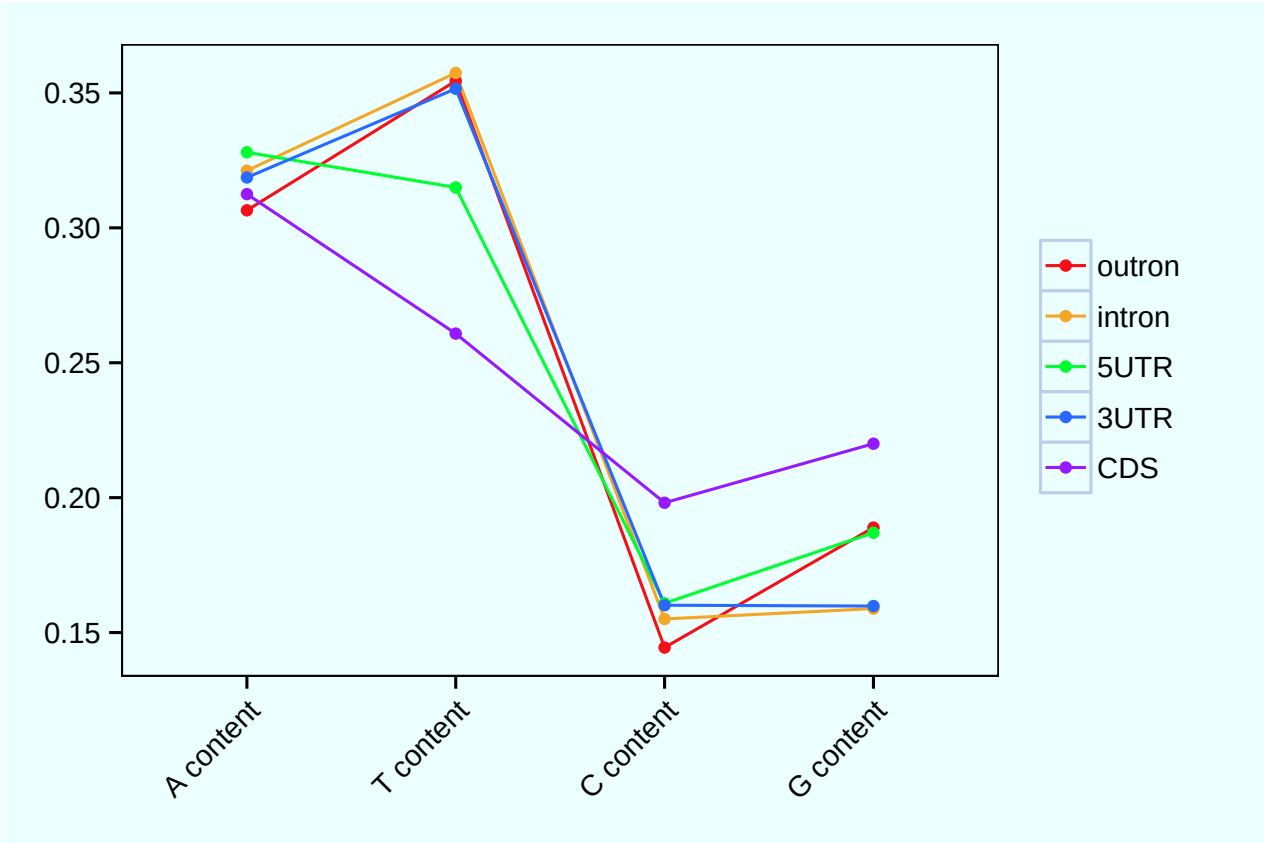

図 28: 推定アウトロンの A、C、G、T 含量。推定アウトロン、イントロン、5′ UTR、 CDS、3′ UTR の 5 つのクラスに対して A、C、G、T 含量を調べた。推定アウトロンの長さ（51 から 2000 nt）を考慮して、51 から 2000 nt の長さのイントロン、5′ UTR、 CDS、3′ UTR だけを用いた。

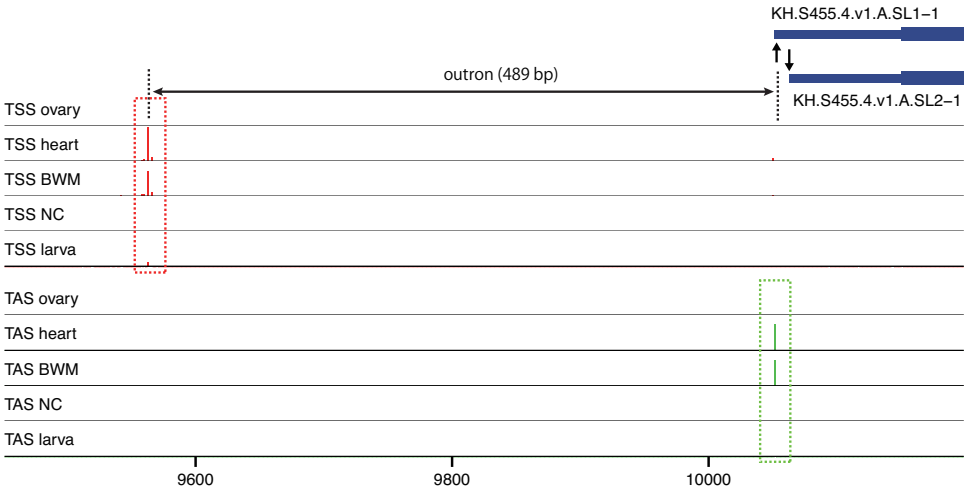

図 29: *Heat shock protein beta-1* 遺伝子のプロモーター候補。赤と青の四角はそれぞれ、TSC と TAC を表す。矢印はアノテーションされた TAS を示す。両矢印は推定されたアウトロンの領域を示す。BWM と NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

4.11 Non-*trans*-spliced 遺伝子プロモーターと *trans*-spliced 遺伝子プロモーターの比較

Non-*trans*-spliced 遺伝子プロモーターと *trans*-spliced 遺伝子プロモーターの性質の差を調べるために、1844 個の non-RP プロモーターを (1) 予測された *trans*-spliced 遺伝子プロモーター、(2) non-*trans*-spliced 遺伝子プロモーター、(3) オペロン遺伝子プロモーター、(4) 予測されたオペロン遺伝子プロモーター、(5) その他のプロモーター (ND) の 5 つのクラスに分類した（表 11)。ここで、non-*trans*-spliced 遺伝子は TSC から翻訳開始点までの領域に TAC が存在しない遺伝子と定義した。また、アウトロンの長さによって差が生じるかを調べるために、*trans*-spliced 遺伝子プロモーターをさらに、短いアウトロン（$<$ 200 bp）のものと長いアウトロン（$\geq$ 200 bp）のものに分類した（表 11)。長さの閾値（200 bp）は、アウトロンの長さの分布に基づいて決定された。TSS 周辺の N_1+N_2 含量を調べたところ、*trans*-spliced 遺伝子プロモーターは +1 から +20 の領域において non-*trans*-spliced 遺伝子プロモーターに比べて高い G+T 含量を示すことが分かった（図 30)。そこで、G+T 含量に着目し、G+T 含量がアウトロンの長さに関わらず保存されているかどうかを調べた。その結果、*trans*-spliced 遺伝子プロモーターは、アウトロンの長さに関わらず TSS 下流 10 から 20 bp 付近で高い G + T 含量を示すことが分かった（図 31)。また、non-*trans*-spliced 遺伝子プロモーターは、その他のクラスに比べてプロモーター下流 (+1 to +20) において有意に低い G + T 含量を示した（FDR$<$ 0.05、マン・ホイットニーの U 検定)。また、同様の解析を 79 個の RP 遺伝子プロモーターに対して行ったところ、数が少ないため有意な差は得られなかったが、*trans*-spliced 遺伝子プロモーターは TSS 下流 30 bp 付近で non-*trans*-spliced 遺伝子プロモーターより高い G + T 含量を示した（図 31)。

表 11: 各クラスのプロモーター数

Class	non-RP	RP	total
ND	947	45	992
non-*trans*-spliced	525	13	538
annotated operon	305	14	319
predicted operon	2	0	2
trans-spliced($<$200bp)	42	4	46
trans-spliced($>$=200bp)	23	3	26
total	1844	79	1923

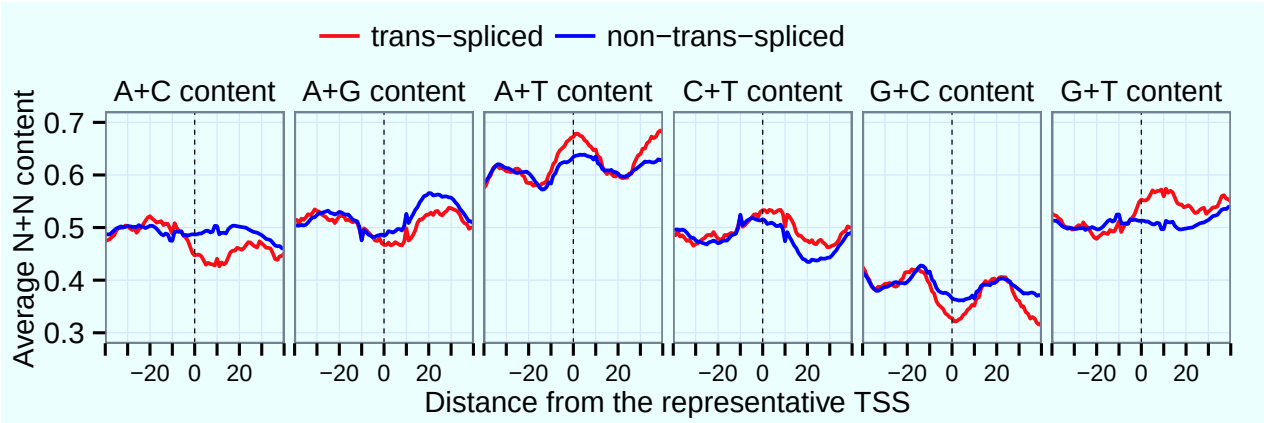

図 30: Non-RP *trans*-spliced 遺伝子プロモーターと non-RP non-*trans*-spliced 遺伝子プロモーターにおける N_1+N_2 含量の分布。各プロモータークラスにおいて平均 N_1+N_2 含量を 20-bp のスライディングウィンドウを用いて計算した。

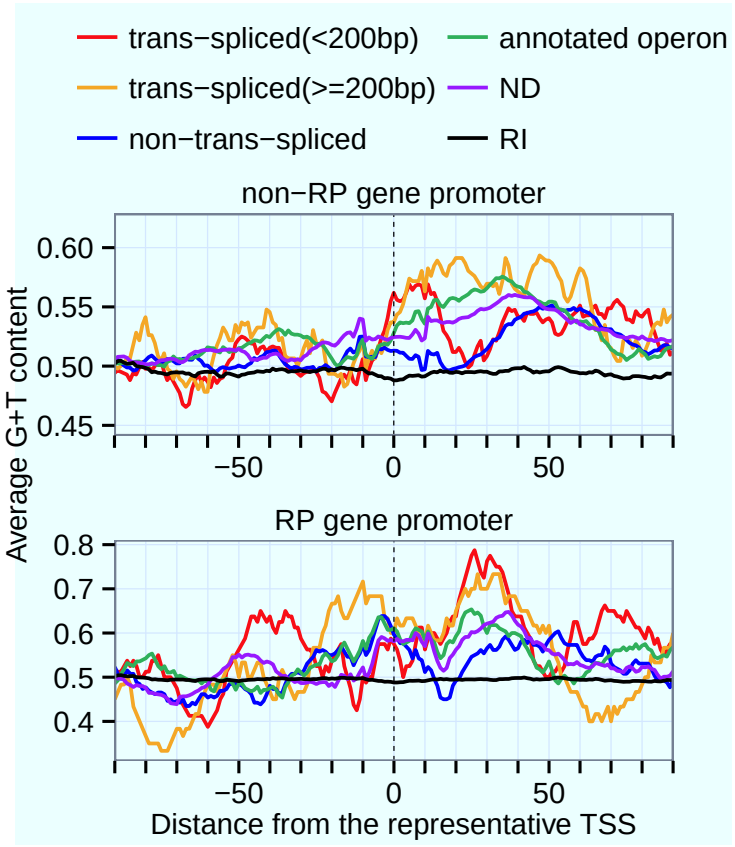

図 31: G+T 含量の分布。上図は non-RP 遺伝子プロモーターの G+T 含量の分布を示す。20-bp のスライディングウィンドウを用いて、各プロモータークラスの平均 G+T 含量を計算した。予測されたオペロン遺伝子プロモーターのクラスは数が少ない (N=2) ため示していない。RI はランダムに取得された遺伝子間領域を表す。下図は RP 遺伝子プロモーターの G+T 含量の分布を示す。

4.12 選択的プロモーターの同定

最終 TSC セットの TSC を次のように遺伝子に割り当てることで、カタユウレイボヤにおける選択的プロモーターを探索した。まず、遺伝子間領域に存在する TSC 以外の TSC をその TSC が位置する遺伝子に割り当てた。次に、もし最も近い下流に存在する遺伝子の 5′ 末端までの距離が 500 bp 以内ならば、遺伝子間領域の TSC をその遺伝子に割り当てた。さらに、遺伝子間領域の TSC の内、SL *trans*-spliced 遺伝子のプロモーター候補と予測された TSC を対応する遺伝子に割り当てた。

3206 個の TSC の内、2703 個が 2581 個の遺伝子に割り当てられた。2581 個の遺伝子の内、およそ 4.5% (115 / 2581) が 2 つ以上の選択的プロモーターを持っていた（表 12)。図 32 に、*Betagamma crystallin* 遺伝子 (Shimeld et al., 2005) の選択的プロモーターを示した。興味深いことに、この遺伝子は神経複合体と幼生で異なるプロモーターを使用していた。幼生特異的プロモーターは上流 30 bp に TATA box を持っていたが、神経複合体特異的プロモーターは持たなかった。

表 12: 1 遺伝子に対するの TSC の数。TSC の最終セットを用いて、選択的プロモーターを探索した。各 TSC は位置に従って遺伝子に割り当てられた。各 TSC は既知もしくは推定プロモーターを表す。

割り当てられた TSC の数	遺伝子数
1	2466
2	109
3	5
4	1
total	2581

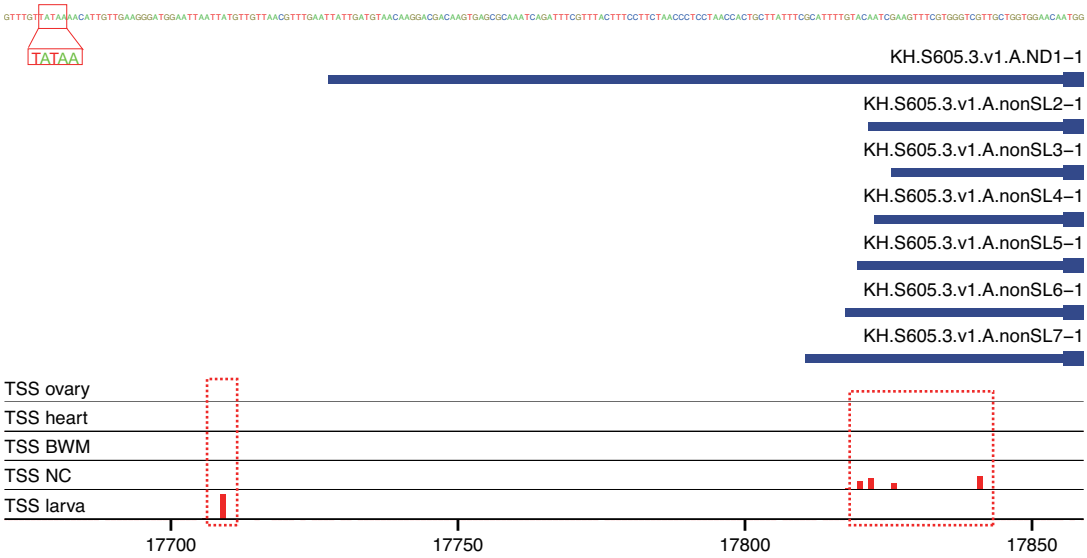

図 32: *Betagamma crystallin* 遺伝子の選択的プロモーター

第 5 章　考察

本研究ではカタユウレイボヤにおいて TSS-seq 法で得られた計 5 つのサンプルを用いてゲノムワイドに TSS を同定した。同定した TSS は同じプロモーター由来だと思われる TSS の集合体（TSC）にクラスタリングされた。その結果、既知 TSS 上だけでなく、CDS や 3′ UTR、遺伝子間領域などの領域にも多数の TSC を発見した。しかしながら、同定した TSC は必ずしもプロモーターを表すわけではないことが分かった。本研究において、3 つの異なるタイプの幅 1 bp の TSC（CTGG TSC、AT-rich TSC、逆鎖のスプラスドナー部位付近の TSC）を発見した。Zhao らは、CAGE のデータを用いた解析において、ultra-dense TSS 分布と呼ばれる異なるタイプの幅 1 bp の TSC を報告し (Zhao et al., 2011)、それらの幾つかは CAGE におけるエラーによって生じた可能性が高いことを示した。TSS の直前に CTGG モチーフを持つ CTGG TSC はおそらく TSS-seq の技術的なアーティファクトである。このアーティファクトは、TSS-seq プロトコル内での 5′ oligo 配列のミスハイブリダイゼーションによって生じる可能性がある。したがって、TSS-seq のデータを用いるときは、CTGG TSC を除去することが重要なステップとなる。他の 2 タイプの幅 1 bp の TSC がどのようなメカニズムで生成されたかは定かではないが、PyPu モチーフを持たないことから、実験的もしくは生物学的ノイズによるものかもしれない。AT-rich な領域における RNA ポリメラーゼ II の一時停止とその後に起こる RNA の切断 (Nechaev et al., 2010) が、もしかしたら AT-rich TSC の生成に関与しているかもしれない 。これらの幅 1 bp の TSC が新規のプロモーターを表しているかどうかを結論づけるためには、DNase-seq のデータや RNA ポリメラーゼ II やヒストン修飾の ChIP-seq データを使ったさらなる解析が必要である。

幅 1 bp の TSC に加えて、CDS 上に多くの TSC を発見した。TSC が発見されたほとんどの CDS において、半分以上の領域が TSC でカバーされていた。このことは、TSS-seq タグが CDS の広範囲に分布していることを示している。また、全体のイントロン領域は TSC でカバーされていないのに、全体の CDS が TSC でカバーされている多くの転写産物モデルを発見した。このことから、多くの転写産物モデルで、TSS-seq タグが CDS 領域に特異的に分布していることが分かった。これらの結果から、CDS 上の TSC は切断された mRNA の 5′ 末端由来である可能性が示唆された。また、エキソン上に存在する TSC の多くがスプライスアクセプター部位の近くにピークをもつことから、切断は成熟 mRNA 後に起こるのではなくスプライシング中に起こるのかもしれない。その場合、イントロン領域が TSC であまりカバーされていない理由は、スプライソソームによってイントロン領域が守られているからであろう。

本研究において、初めて正確な RP 遺伝子の主要 TSS 候補を同定した。現在の遺伝子モデル（KH モデル）では、各 RP 遺伝子に対して 5′ 末端の異なる転写産物モデルが複数存在しており、どの既知 TSS が主要な TSS なのかはモデルだけからは判断できないという問題がある。したがって、本研究の結果は RP 遺伝子プロモーターの TSS に関して重要かつ有用な情報であると言える。また、驚くべきことに 6 つの RP 遺伝子の主要 TSS はどの既知 TSS とも離れたところに存在しお

り、新規に同定された TSS であった。その内の一つである *Rpl21* 遺伝子は面白いことにオペロンの第二遺伝子であり、トランススプライシングを受けることが示唆された。しかしながら、RP 遺伝子はあまりトランススプライシングを受けない遺伝子とされており、その理由としては RP 遺伝子が 5′ 末端領域に翻訳効率に関わっているとされる TOP 配列を持っているからと考えられている (Satou et al., 2006; Matsumoto et al., 2010)。この結果は、カタユウレイボヤにおいても、ワカレオタマボヤで示唆されているようにトランススプライシングによって付加される SL 配列が翻訳効率に関わっているということを示唆しているかもしれない (Danks et al., 2015)。

カタユウレイボヤにおける RP 遺伝子プロモーターの性質を調べたところ、RP 遺伝子プロモーターはヒトと同様にシトシン塩基から転写が始まる高度に保存されたポリピリミジンイニシエーターモチーフを持っていることが分かった。特にカタユウレイボヤでは、ポリピリミジンイニシエーターモチーフの中心 6 塩基（−1 から +4）は非常に高度に保存されていた。この保存されたポリピリミジン配列はショウジョウバエの RP 遺伝子プロモーターでも観察されており、変異解析によって中心の 6 塩基が RP 遺伝子の転写に最も重要であることを示している (Parry et al., 2010)。このことから、カタユウレイボヤにおいてもシトシン塩基から転写が始まる −1 から +4 の中心 6 塩基のポリピリミジン配列が転写に重要な役割をもつと予想される。しかしながら、*Rpl22* 遺伝子では、転写はシトシン塩基から始まらず、2 nt 下流のチミン塩基から始まっていた。このことは、シトシン塩基が RP 遺伝子における絶対的な転写開始点でないことを示している。なぜ、*Rpl22* 遺伝子において転写はシトシン塩基から始まらず、2 nt 下流のチミン塩基から始まるのかは不明であるが、全ての RP 遺伝子のポリピリミジンイニシエーターモチーフにシトシン塩基が含まれていること考慮するとシトシン塩基の存在は重要であると考えられる。つまり、シトシン塩基を含まないチミン塩基のみから構成されるイニシエーターモチーフでは転写は起きないかもしれない。以上の結果を統合すると、カタユウレイボヤの RP 遺伝子ではポリピリミジン配列のシトシン塩基もしくはチミン塩基が主要 TSS のようである。しかしながら、低頻度ではあるがポリピリミジン配列のすぐ下流に存在する PyPu の Pu 塩基からも転写が始まることもあるようである（図 S9)。これらの TSS はマイナー TSS と考えられるが、マイナー TSS から転写されて生じた mRNA はその 5′ 末端に翻訳制御タンパク質の結合を阻害する配列と考えれている TOP 配列を持たないことになる。したがって、マイナー TSS から転写されて生じた mRNA は主要 TSS から転写され 5′ 末端に TOP 配列を持つ mRNA とは異なる転写後調節を受けるかもしれない。

カタユウレイボヤでもヒトでも RP 遺伝子プロモーターはよく保存されたポリピリミジンイニシエーターモチーフを持っていたが、ヒト RP 遺伝子プロモーターと違い、ほとんどのカタユウレイボヤの RP 遺伝子プロモーターは TATA box を持たないことが示唆された。RP 遺伝子プロモーターの構造の一般的特徴はヒト、ニワトリ、両生類、魚においてよく保存されているので (Perry, 2005)、カタユウレイボヤにおいて TATA box が存在しなかったことは、RP 遺伝子プロモーターが TATA box を獲得したのは脊椎動物の進化の初期段階であることを示唆している。そのため、ヤツメウナギなど原始的な脊椎動物において TATA box や TATA-like な配列があるのかどうかを調べるのは興味深いと思われる。最近の研究によって、ショウジョウバエの RP 遺伝子の転写には、TBP (TATA box-binding protein) は必要ではなく、TRF2 (TBP-related factor 2) が必要

であることがわかっている (Wang et al., 2014)。したがって、TATA box を持たないカタユウレイボヤの RP 遺伝子の転写も TRF2 依存的である可能性がある。RP 遺伝子プロモーターに存在するモチーフの同定はヒトやショウジョウバエ、酵母、線虫、 原始的な後生動物など多くの種において行われているが (Tanay et al., 2005; Roepcke et al., 2006; Ma et al., 2009; Perina et al., 2011)、本研究ではイニシエーターと最もよく解析されているモチーフである TATA box にしか焦点を当てなかった。カタユウレイボヤの RP 遺伝子の転写メカニズムの解明や脊索動物の RP 遺伝子プロモーターの進化の理解のためには、さらなる詳細な解析が必要である。

カタユウレイボヤプロモーターにおいて、どの 2 塩基が TSS としてよく用いられるのかを調べた。16 個の 2 塩基の内、3 つの PyPu 塩基（CA、TA、TG）が最も頻繁に TSS として使用されていた。これらの PyPu 塩基はヒトにおいても TSS としてよく使用されていたが、その使用率はカタユウレイボヤとヒト間で明確に異なっていた。ヒトでは、CA だけが最も優先的に使用されていたが、カタユウレイボヤでは CA と TA が最も頻繁に使用されていた。この差は、おそらく AT-rich なカタユウレイボヤゲノムに起因していると思われる。実際、カタユウレイボヤにおいては、TSS 付近において T 含量がその他の塩基の含量より高かった（図 S10）。この高 T 含量がカタユウレイボヤにおける TA の高い使用率の原因だと思われる。一方、CG の使用率はカタユウレイボヤよりもヒトで高かった。これはヒトプロモーターの高い GC 含量を反映していると考えられる。

ヒトでは、TATA-less プロモーターは TATA-containig プロモーターよりも高い CpG 含量と低い発現特異性を示した。このことは、TATA-less プロモーターが CpG アイランドやハウスキーピング遺伝子と関連があることを示唆している (Carninci et al., 2006; Yang et al., 2007)。一方、カタユウレイボヤでは、TATA-less プロモーターは高い CpG 含量を示さず、この結果はカタユウレイボヤが CpG アイランドを持たないという仮説を支持している (Okamura et al., 2011)。それにも関わらず、カタユウレイボヤの TATA-less プロモーターはヒトと同様に TATA-containing プロモーターよりも低い発現特異性を示した。果たしてカタユウレイボヤのプロモーターはハウスキーピング遺伝子に関連する CpG アイランドに代わるものを持っているのだろうか？現在のところ、どのようなエレメントが TATA-less プロモーターに RNA ポリメラーゼをリクルートするのに重要な役割をもつのかわかっていない。ヒトの TATA-less プロモーターでは、TBP をリクルートできる SP1 の結合部位である GC box が多数 CpG islands 上に存在する (Butler and Kadonaga, 2002; Deaton and Bird, 2011)。カタユウレイボヤの TATA-less プロモーターの未知エレメントも、もしかしたら CpG islands 上の GC box のように多数プロモーター上に存在し、基本転写因子をリクルートする未知のタンパク質の結合部位として機能するのかもしれない。このようなエレメントが存在するかどうかを知るには、カタユウレイボヤの TATA-less プロモーターに対するより詳細な解析が必要である。少なくとも、JASPAR (Sandelin et al., 2004) に登録されている既知コアプロモーターエレメントおよびショウジョウバエのプロモーターモチーフは sharp-type TATA-less プロモーターと broad-type TATA-less プロモーターのどちらにおいても関連性を示さなかった（図 S11）。

本研究では TSS-seq データを用いることによって既知プロモーターと推定プロモーターを同定

した。多くの推定プロモーターは遺伝子間領域に存在しており、KH モデルにアノテーションされていない転写単位が多数存在することが示唆された。カタユウレイボヤにおいては、マイクロ RNA を含む多くのノンコーディング RNA 候補が同定されており (Sasakura et al., 2012)、これらのアノテーションされていない転写産物のいくつかはノンコーディング RNA をコードするかもしれない。また、推定プロモーターは 5′ UTR やイントロンにも多く発見された。これらの推定プロモーターは選択的プロモーターを表していると考えられる。面白いことに、幾つかの推定プロモーター（TSC）は TAS 上に存在していた。この結果は、TAS 付近の領域が SL *trans*-spliced 遺伝子の選択的プロモーターとして機能することを示唆してるかもしれない。また、これらの *trans*-spliced 遺伝子は発現するのにトランススプライシングを必要としない可能性がある。トランススプライシングの機能の一つとして、mRNA の輸送や翻訳効率に対して害となる 5′ UTR 上に存在するエレメントの除去が提唱されているが (Hastings, 2005)、TAS 付近の選択的プロモーターからの転写は成熟 mRNA の 5′ UTR 上に有害なエレメントが存在しないようにするもう一つの方法なのかもしれない。

本研究において同定した TSC は、non-*trans*-spliced 遺伝子のプロモーターだけでなく、*trans*-spliced 遺伝子のプロモーターも表している可能性がある。実際、Khare らは 2 つの異なる方法（その内 1 つは、理論的に TSS-seq と同じ方法）を用いて、*Troponin I* 遺伝子の TSS を同定・検証している (Khare et al., 2011)。また、本研究でもタグ数は少ないが *Troponin I* 遺伝子の TSS 上に TSS-seq タグを確認できた（図 S12)。この結果は、同定した TSC の中に、*trans*-spliced 遺伝子のプロモーターを含む可能性を示している。そこで、本研究では *trans*-spliced 遺伝子のプロモーターを表している可能性が高い TSC を探索した。その結果、*trans*-spliced 遺伝子のプロモーターである可能性が高い TSC 群を発見した。この TSC 群の中には推定プロモーターが含まれており、これらは *trans*-spliced 遺伝子の新規プロモーターを表しているかもしれない。同定した *trans*-spliced 遺伝子プロモーター候補と non-*trans*-spliced 遺伝子プロモーターを比較したところ、non-*trans*-spliced 遺伝子プロモーターの下流領域は、他のプロモーターに比べて G+T 含量が低いことがわかった。一方、*trans*-spliced 遺伝子プロモーター候補の下流領域は、高い G+T 含量を示す傾向にあった。また、この傾向は、RP 遺伝子プロモーターにおいても同様に観察された。この高 G 含量がどのような役割を担っているのかは定かではないが、トランススプライシングのしやすさに影響があるのかもしれない。本研究において予測された TSC が真に SL *trans*-spliced 遺伝子であるかを判断するには実験的な検証が必要になるが、これらのデータは今後の *trans*-spliced 遺伝子のシス調節領域の解析に役立つと考えられる。

謝辞

本研究を行うにあたり、多くの方々にご支援・ご協力を頂きました。ここに感謝の意を表したいと思います。東京大学医科学研究所ヒトゲノム解析センターの中井謙太教授には、研究テーマや研究方針を始めとして、多くの有益なご指導を頂きました。甲南大学理工学部生物学科の日下部岳広教授には、カタユウレイボヤのサンプルの採取および研究に関する助言をして頂きました。東京大学新領域創成科学研究科の鈴木穣教授には、サンプルのシークエンシングをして頂きました。最後に、研究および研究生活において大変お世話になった中井研究室の先輩方に、心より御礼申し上げます。

参考文献

Agabian, N. (1990). Trans splicing of nuclear pre-mRNAs. *Cell* **61**:1157–60.

Baumann, M., Pontiller, J., and Ernst, W. (2010). Structure and Basal Transcription Complex of RNA Polymerase II Core Promoters in the Mammalian Genome: An Overview. *Molecular Biotechnology* **45**:241–247.

Butler, J. E. and Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev* **16**:2583–92.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* **309**:1559–1563.

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

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engstrom, P. G., Frith, M. C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**:626–635.

Conrad, R., Lea, K., and Blumenthal, T. (1995). SL1 trans-splicing specified by AU-rich synthetic RNA inserted at the 5' end of Caenorhabditis elegans pre-mRNA. *RNA* **1**:164–70.

Corbo, J. C., Levine, M., and Zeller, R. W. (1997). Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Development* **124**:589–602.

Dabney, A., Storey, J. D., and with assistance from Gregory R. Warnes (2013). *qvalue: Q-value estimation for false discovery rate control*.

Danks, G. B., Raasholm, M., Campsteijn, C., Long, A. M., Manak, J. R., Lenhard, B., and Thompson, E. M. (2015). Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol Biol Evol* **32**:585–99.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev* **25**:1010–22.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., et al. (2002). The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins. *Science* **298**:2157–2167.

Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**:965–8.

Di Gregorio, A. and Levine, M. (2002). Analyzing gene regulation in ascidian embryos: new tools for new perspectives. *Differentiation* **70**:132–9.

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V., Lassman, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., et al. (2014). A promoter-level mammalian expression atlas. *Nature* **507**:462–70.

Ganot, P., KallesÃÿe, T., Reinhardt, R., Chourrout, D., and Thompson, E. M. (2004). Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome. *Mol Cell Biol* **24**:7795–805.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**:1017–8.

Harafuji, N., Keys, D. N., and Levine, M. (2002). Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. *Proc Natl Acad Sci U S A* **99**:6802–5.

Hastings, K. E. (2005). SL trans-splicing: easy come or easy go? *Trends Genet* **21**:240–7.

Hoskins, R. A., Landolin, J. M., Brown, J. B., Sandler, J. E., Takahashi, H., Lassmann, T., Yu, C., Booth, B. W., Zhang, D., Wan, K. H., et al. (2011). Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research* **21**:182–192.

Irvine, S. Q. (2013). Study of Cis-regulatory Elements in the Ascidian Ciona intestinalis. *Curr Genomics* **14**:56–67.

Johnson, D. S., Davidson, B., Brown, C. D., Smith, W. C., and Sidow, A. (2004). Non-coding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Res* **14**:2448–56.

Johnson, D. S., Zhou, Q., Yagi, K., Satoh, N., Wong, W., and Sidow, A. (2005). De novo discovery of a tissue-specific gene regulatory module in a chordate. *Genome Res* **15**:1315–24.

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. M., and Kadonaga, J. T. (2008). The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology* **20**:253–259.

Kawaji, H., Frith, M. C., Katayama, S., Sandelin, A., Kai, C., Kawai, J., Carninci, P., and Hayashizaki, Y. (2006). Dynamic usage of transcription start sites within core promoters. *Genome Biology* **7**.

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH (TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Research* **31**:3576–3579.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res* **12**:656–64.

Khare, P., Mortimer, S. I., Cleto, C. L., Okamura, K., Suzuki, Y., Kusakabe, T., Nakai, K., Meedel, T. H., and Hastings, K. E. (2011). Cross-validated methods for promoter/transcription start site mapping in SL trans-spliced genes, established using the Ciona intestinalis troponin I gene. *Nucleic Acids Res* **39**:2638–48.

Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T., and Meyer, B. J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *Elife* **2**:e00808.

Kusakabe, T. (2005). Decoding cis-regulatory systems in ascidians. *Zoolog Sci* **22**:129–46.

Kusakabe, T., Yoshida, R., Ikeda, Y., and Tsuda, M. (2004). Computational discovery of DNA motifs associated with cell type-specific gene expression in Ciona. *Dev Biol* **276**:563–80.

Ma, X., Zhang, K., and Li, X. (2009). Evolution of Drosophila ribosomal protein gene core promoters. *Gene* **432**:54–9.

Maruyama, K. and Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**:171–4.

Matsumoto, J., Dewar, K., Wasserscheid, J., Wiley, G. B., Macmil, S. L., Roe, B. A., Zeller, R. W., Satou, Y., and Hastings, K. E. (2010). High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. *Genome Res* **20**:636–45.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**:D108–10.

Nakao, A., Yoshihama, M., and Kenmochi, N. (2004). RPG: the Ribosomal Protein Gene database. *Nucleic Acids Res* **32**:D168–70.

Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**:335–8.

Ni, T., Corcoran, D. L., Rach, E. A., Song, S., Spana, E. P., Gao, Y., Ohler, U., and Zhu, J. (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods* **7**:521–U57.

Nilsen, T. W. (1993). Trans-splicing of nematode premessenger RNA. *Annu Rev Microbiol* **47**:413–40.

Okamura, K., Yamashita, R., Takimoto, N., Nishitsuji, K., Suzuki, Y., Kusakabe, T., and Nakai, K. (2011). Profiling ascidian promoters as the primordial type of vertebrate promoter. *BMC Genomics* **12(Suppl 3)**.

Parry, T. J., Theisen, J. W., Hsu, J. Y., Wang, Y. L., Corcoran, D. L., Eustice, M., Ohler, U., and Kadonaga, J. T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24**:2013–8.

Perina, D., Korolija, M., Roller, M., Harcet, M., JeliÄĐiÄĞ, B., MikoÄĐ, A., and CetkoviÄĞ, H. (2011). Over-represented localized sequence motifs in ribosomal protein gene promoters of basal metazoans. *Genomics* **98**:56–63.

Perry, R. P. (2005). The architecture of mammalian ribosomal protein promoters. *BMC Evol Biol* **5**:15.

Ponjavic, J., Lenhard, B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Sandelin, A. (2006). Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol* **7**:R78.

Putnam, N. H., Butts, T., Ferrier, D. E., Furlong, R. F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J. K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**:1064–71.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and GlÃűckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**:D590–6.

Rach, E. A., Yuan, H.-Y., Majoros, W. H., Tomancak, P., and Ohler, U. (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology* **10**.

Roepcke, S., Zhi, D., Vingron, M., and Arndt, P. F. (2006). Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters. *Gene* **365**:48–56.

Sandelin, A., Alkema, W., EngstrÃűm, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**:D91–4.

Sasakura, Y., Sierro, N., Nakai, K., Inaba, K., and Kusakabe, T. (2012). *Genome Structure, Functional Genomics, and Proteomics in Ascidians*, volume 4 of *Genome Mapping and Genomics in Animals*, book section 4, pages 87–132. Springer Berlin Heidelberg.

Satou, Y., Hamaguchi, M., Takeuchi, K., Hastings, K. E., and Satoh, N. (2006). Genomic overview of mRNA 5'-leader trans-splicing in the ascidian Ciona intestinalis. *Nucleic Acids Res* **34**:3378–88.

Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A., and Satoh, N. (2005). An integrated database of the ascidian, Ciona intestinalis: Towards functional genomics. *Zoological Science* **22**:837–843.

Satou, Y., Mineta, K., Ogasawara, M., Sasakura, Y., Shoguchi, E., Ueno, K., Yamada, L., Matsumoto, J., Wasserscheid, J., Dewar, K., et al. (2008). Improved genome assembly and evidence-based global gene model set for the chordate Ciona intestinalis: new insight into intron and operon populations. *Genome Biology* **9**.

Shimeld, S. M., Purkiss, A. G., Dirks, R. P., Bateman, O. A., Slingsby, C., and Lubsen, N. H. (2005). Urochordate betagamma-crystallin and the evolutionary origin of the vertebrate eye lens. *Curr Biol* **15**:1684–9.

Stolfi, A. and Christiaen, L. (2012). Genetic and genomic toolbox of the chordate Ciona intestinalis. *Genetics* **192**:55–66.

Suzuki, Y., Tsunoda, T., Sese, J., Taira, H., Mizushima-Sugano, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Nakamura, Y., et al. (2001). Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Research* **11**:677–684.

Suzuki, Y., Yoshitomo-Nakagawa, K., Maruyama, K., Suyama, A., and Sugano, S. (1997). Construction and characterization of a full length-enriched and a 5'-end-enriched cDNA library. *Gene* **200**:149–56.

Takahashi, H., Mitani, Y., Satoh, G., and Satoh, N. (1999). Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Development* **126**:3725–34.

Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* **102**:7203–8.

van Heeringen, S. J., Akhtar, W., Jacobi, U. G., Akkers, R. C., Suzuki, Y., and Veenstra, G. J. C. (2011). Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Research* **21**:410–421.

Vandenberghe, A. E., Meedel, T. H., and Hastings, K. E. M. (2001). mRNA 5 '-leader trans-splicing in the chordates. *Genes & Development* **15**:294–303.

Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**:e178.

Wang, Y. L., Duttke, S. H., Chen, K., Johnston, J., Kassavetis, G. A., Zeitlinger, J., and Kadonaga, J. T. (2014). TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes Dev* **28**:1550–5.

Yamamoto, Y. Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J. (2009). Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant Journal* **60**:350–362.

Yamashita, R., Sathira, N. P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S., Sugano, S., Nakai, K., and Suzuki, Y. (2011). Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* **21**:775–89.

Yamashita, R., Sugano, S., Suzuki, Y., and Nakai, K. (2012). DBTSS: DataBase of Transcriptional Start Sites progress report in 2012. *Nucleic Acids Res* **40**:D150–4.

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., and Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**:52–65.

Yoshihama, M., Uechi, T., Asakawa, S., Kawasaki, K., Kato, S., Higa, S., Maeda, N., Minoshima, S., Tanaka, T., Shimizu, N., et al. (2002). The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res* **12**:379–90.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning

DNA sequences. *J Comput Biol* **7**:203–14.

Zhao, X., Valen, E., Parker, B. J., and Sandelin, A. (2011). Systematic Clustering of Transcription Start Site Landscapes. *Plos One* **6**.

付録 A　Supplemental Methods

A.1　CTGG TSC の除去

TSS-seq では 5′ oligo がミスハイブリダイズすることによりアーティファクトが生成される可能性がある。ミスハイブリダイゼーションによって生じたリードは偽の TSS を生み出し、そのすぐ上流 4 塩基は CTGG となる。したがって、エキソンのアンチセンス鎖上に存在し、上流 4 塩基が CTGG もしくはそのバリアント（表 S2）である TSC を除去した。また、エキソンのアンチセンス鎖上に存在しない場合でも、上流 4 塩基が CTGG の TSC は除去した。

A.2　幅が 1 bp の TSC の除去

カタユウレイボヤとヒトそれぞれにおいて、AT 含量が 0.8 以上、0.66 以上の AT-rich な幅が 1 bp の TSC を除去した。また、逆鎖のスプライスドナー部位から 3 bp 以内に存在する幅が 1 bp の TSC も除去した。

A.3　切断された RNA 由来の TSC の除去

切断された RNA に由来する偽の TSC を見つけるために、TSS-seq リードが広範囲に分布しているエキソン領域を探索した。各エキソン領域に対して、TSC にカバーされる領域の割合を計算し、半分以上の領域がカバーされていた場合、そのエキソンを疑わしいエキソンと定義した。そして、疑わしいエキソンを少なくとも 1 つ持つ転写産物モデルのエキソンにオーバーラップしている TSC 全てを疑わしい TSC と定義した。ある転写産物の疑わしい TSC は、以下のステップで除去された。まず、その転写産物モデルの TSS 上に TSC が存在しているかを調べた。これは、転写された RNA の全てではなく一部が切断された場合、エキソン上には偽の TSC、TSS 上には真の TSC が存在するはずであるからである。この場合、TSS 上にある真の TSC のピークの高さはエキソン上の偽の TSC のピークより高くなると予想される。そこで、TSS 上に TSC があった場合、疑わしい TSC のピークの高さが、全サンプルにおいて TSS 上の TSC のピークの高さの 5 分の 1 より低いならば、疑わしい TSC を偽の TSC と見なした。次に、TSS 上に TSC はないが、他の疑わしい TSC が同じエキソン上か前後のエキソン上に存在した場合、もし全サンプルにおいて疑わしい TSC のピークの高さが、他の疑わしい TSC のピークの 5 倍以上高くないならば、偽の TSC と見なした。ただし、転写産物モデルが一つしかエキソンを持たないときは、エキソンの半分の領域をカバーしているならば、疑わしい TSC を偽の TSC と見なし除去した。さらに、アクセプター部位の 5 bp 以内に存在し、右歪曲な TSS 分布を示す疑わしい TSC も除去した（付録 A.4 参照）。上記のステップによって除去された TSC は偽の TSC と見なされた。最後に、エキソン上にある TSC の内、同じ転写産物モデル上にある偽の TSC のピークの高さの 5 倍以上のピークを全

サンプルにおいて持たない TSC を除去した。

A.4 右歪曲な TSC

右歪曲な TSC は、次の 3 つ条件の内 2 つの条件を満たす TSS 分布をもつ TSC と定義した。(1) 5 パーセンタイルが 95 パーセンタイルと等しくなく、最頻値が 5 パーセンタイルと等しいか、それより左側にある。(2) ピーク TSS が最頻値より右側にない。(3) 歪度が 0 より大きい。歪度は、以下の式によって定義される。

$$y = \sqrt{n}\frac{\sum(x_i - \overline{x})^3}{\{\sum(x_i - \overline{x})^2\}^{3/2}}$$

ここで、n、x、$\overline{x}$ はそれぞれ、 TSC の総タグ数、タグの位置、タグの平均位置を表す。

A.5 TATA box の探索

TRANSFAC の MATCH プログラム (Kel et al., 2003) を用いて TATA box を探索した。モチーフとして、TRANSFAC データーベース (Matys et al., 2006) の V$TATA_C と V$TATA_01 を、カットオフ値として minSUM.prf を用いた。本研究では、モチーフのコア配列の最初の位置を TATA box の位置として用いた。

A.6 TSS 分布のタイプ

プロモーターを、TSS 分布に従い 3 つのタイプ（sharp、broad、other）に分類した。まず最初に、タグの 90% 以上が 10 bp 以内に存在するプロモーターを narrow (NR) プロモーターと定義した。この定義は、ほとんどのタグが非常に狭い（narrow）領域に存在することに由来する。次に、NR プロモーターに対してピーク TSS を探索した。ここで、ピーク TSS とは、TSC の中で最も高頻度な TSS の頻度より 2 分の 1 以上の頻度をもつ TSS のことである。もし最初のピーク TSS と最後のピーク TSS の距離が 5 bp 未満ならば、そのプロモーターを narrow and sharp (NSP) プロモーターと定義した。次に、残りのプロモーターをシャープピークの数に従って 3 つのタイプに分類した。まず、10-bp スライディングウィンドウを用いてピーク TSS をクラスタリングすることでプロモーターのピークを同定した。次に、同定したピークがシャープなのかブロードなのかを評価するために、ピーク内の TSS-seq タグの分布の標準偏差を計算した。もし、標準偏差が閾値未満ならば、ピークはシャープと見なされた。NSP プロモーターのビークの標準偏差の分布の 90 パーセンタイルを閾値として用いた。もしプロモーターがシャープピークを 1 つもつ場合、そのプロモーターを wide and sharp peak (WSP) プロモーターと定義した。また、もしプロモーターが複数のピークをもちそれらが全てシャープピークならば、multiple peak (MP) プロモーターと定義された。最後に、NR、WSP、MP のどれにも属さないプロモーターを broad (BR) プロモーターと定義した。NSP プロモーターと BR プロモーターをそれぞれ“shar” プロモーターと“broad” プロモーターと呼ぶことにする。その他のタイプ（NR、WSP、MP）は曖昧なタイプとみ

なし、“other” に統合した。

A.7 相対エントロピー

発現特異性は、カルバック・ライブラー情報量（相対エントロピーとも呼ばれる）を用いて評価された (Zhao et al., 2011; Ponjavic et al., 2006) 。あるクラスター（TSC もしくは TAC）のカルバック・ライブラー情報量（KLD）は、次の式によって計算できる。

$$\mathrm{KLD} = \sum_{i=1}^{N} p_i log \frac{p_i}{q_i}$$

ここで、i はサンプル、N はカタユウレイボヤの場合 5、ヒトの場合 15 である。また、p と q は、それぞれクラスターのタグの離散確率分布、全クラスターを合計したタグの離散確率分布を表す。p_i と q_i は、それぞれサンプル i における確率を表す。

A.8 超幾何分布による検定

各サンプルにおけるクラスターの高発現度を統計的に評価するために超幾何分布を用いた。まず、各サンプルにおいて、各クラスターの ppm と呼ばれる標準化発現量を計算した。標準化発現量は、クラスターのタグ数 / サンプルの総タグ数 × 1,000,000 によって計算される (Yamashita et al., 2011)。超幾何分布による検定では、カウントデータしか用いることができないので、ppm 値を最も近い整数値に変換した。ただし、ppm 値が 0 以上 0.5 未満のときは 1 とした。そして、各クラスターに対して、超幾何分布による検定を用いてどのサンプルで統計的に有意に高発現しているかを調べた。クラスターのあるサンプルにおける p-value は、以下の式によって計算される。

$$p = \sum_{i=x}^{min(m,k)} \frac{\binom{m}{i}\binom{N-m}{k-i}}{\binom{N}{k}}$$

ここで、N, m, k, x はそれぞれ、全サンプルの全クラスターの総発現量、そのサンプルにおける全クラスターの総発現量、そのクラスターの全サンプルにおける総発現量、そのクラスターのそのサンプルにおける発現量を表す。

あるクラスターがどのサンプルで有意に高発現しているかを決定するために、超幾何分布による検定で求めた p-value と KLD によって評価された発現特異性を用いた。まず、KLD が 0.7 未満もクラスターは non-specific なクラスターと見なした。この閾値は、RP に関わるクラスターの発現特性に基づいている。そしてあるサンプルにおける p-value が 0.01 未満の場合、そのクラスターはそのサンプルで高発現していると見なした。また、クラスターが単一のサンプルで高発現している場合、それを組織特異的クラスターと見なした。

A.9 クラスターペアの分類

同じ組織特異性を示した TAC とその上流にある TSC のペア（クラスターペア）を次のように 2 つのクラス（unannotated-operon-type と non-operon-type）に分類した。もし、ペアの TAC と TSC が、異なる遺伝子座の転写産物モデルの 5′ 末端上に位置しており、その 2 つの遺伝子がオペロンを構成していない場合、そのペアは “unannotated-operon-type” に分類された。これらの遺伝子はまだアノテーションされていないオペロンを構成しているかもしれない。その他のペアは “non-operon-type” に分類された。non-operon-type の TSC と TAC 間の領域は推定アウトロンと考えられた。

付録 B　Supplemental Figures

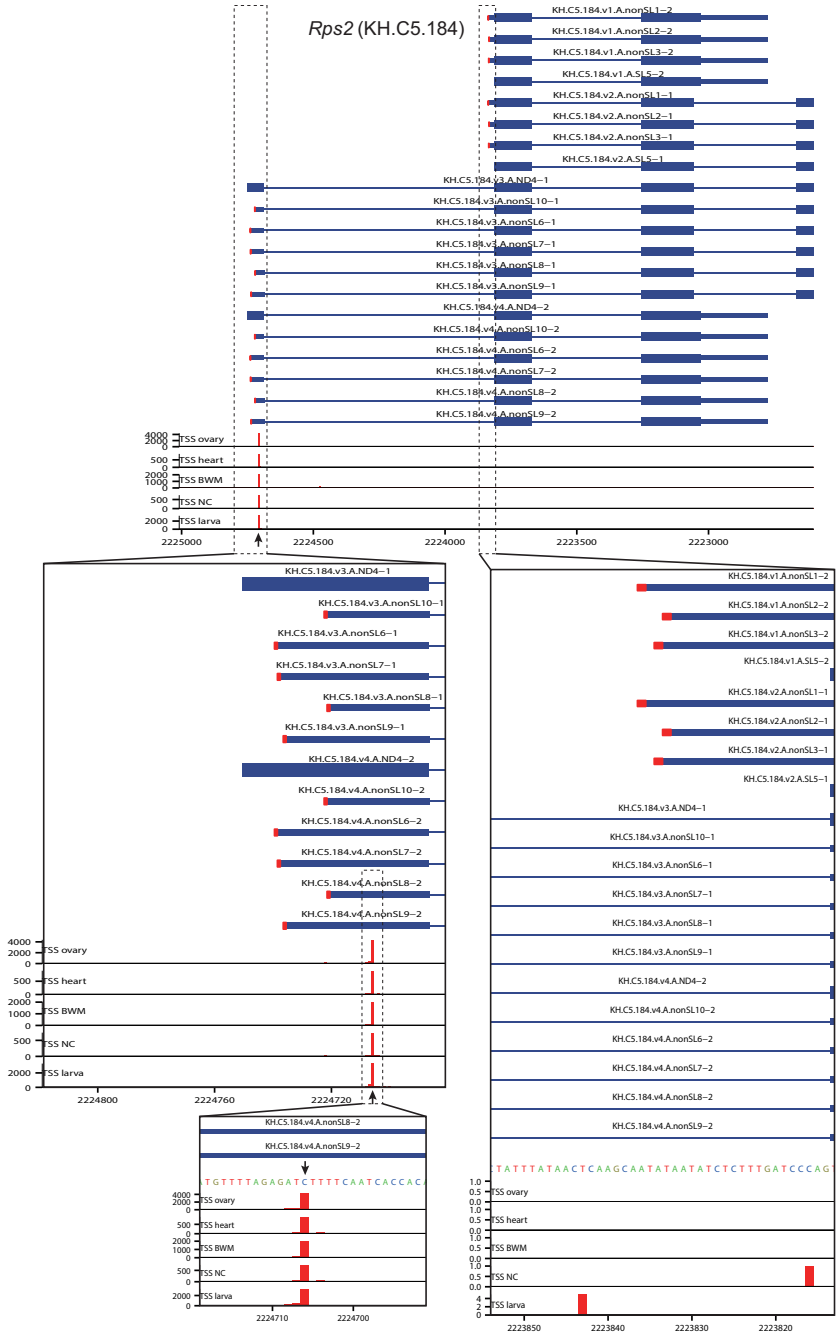

図 S1: *Rps2* 遺伝子（KH.C5.184）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

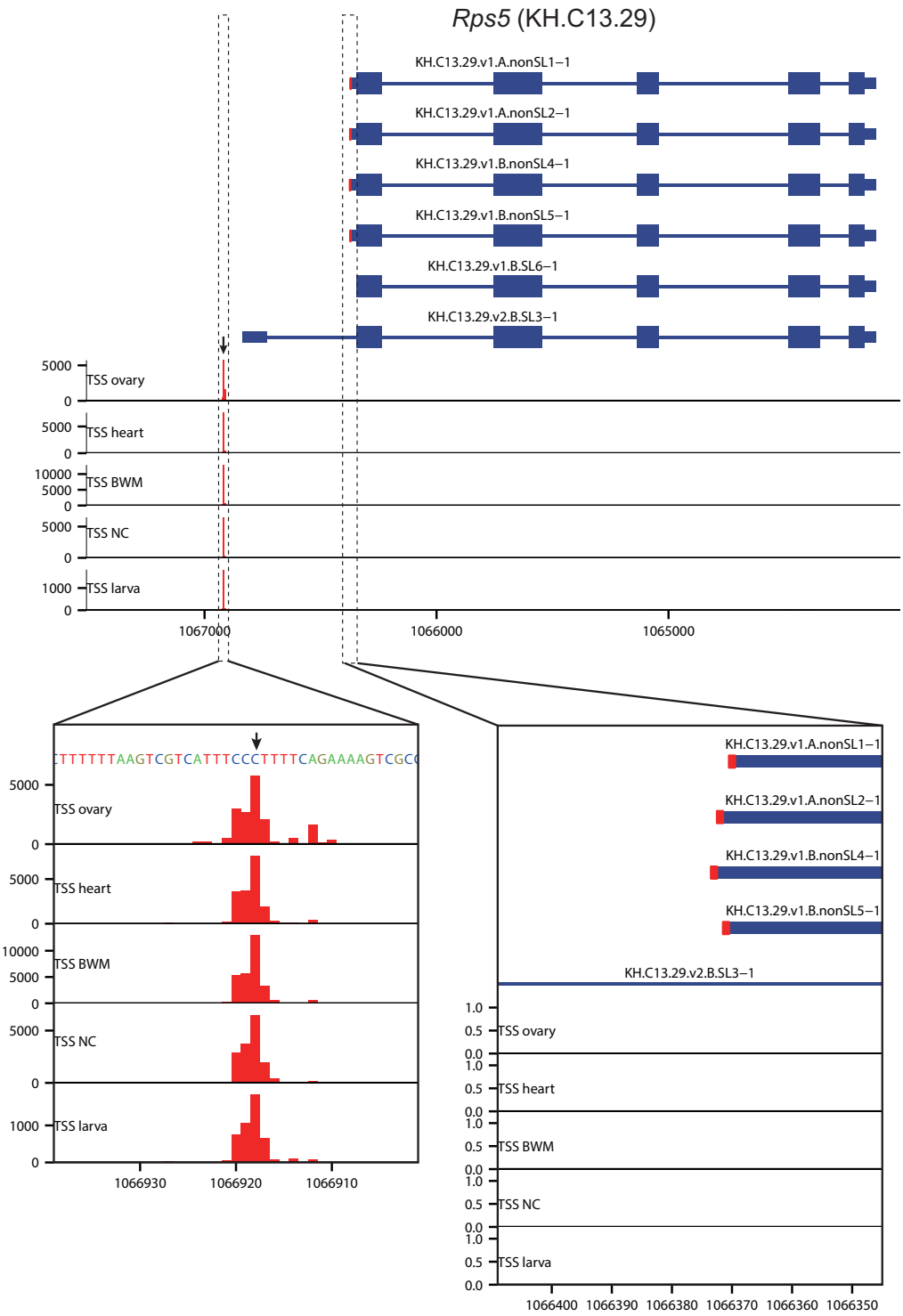

図 S2: *Rps5* 遺伝子（KH.C13.29）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

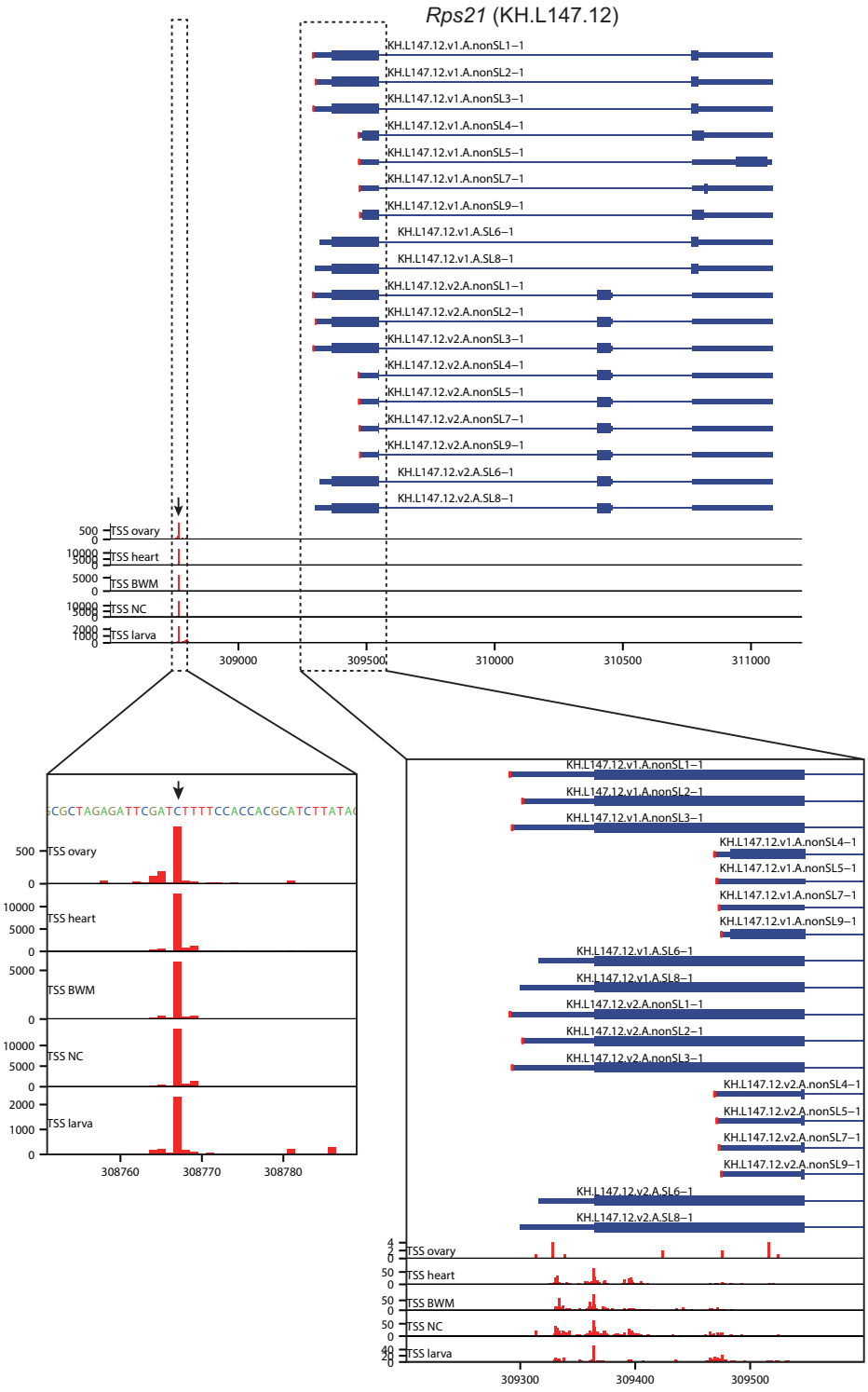

図 S3: *Rps21* 遺伝子（KH.L147.12）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

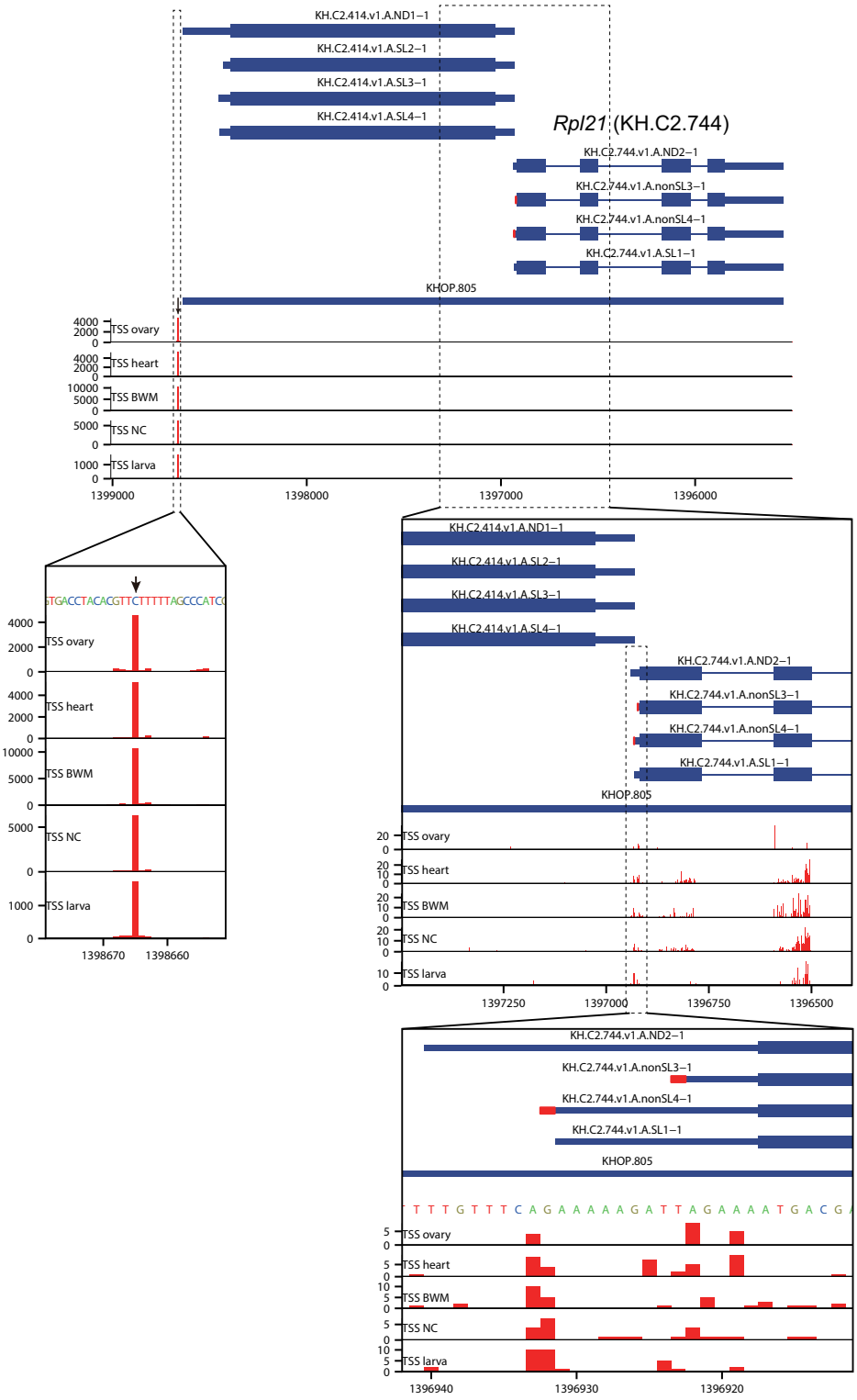

図 S4: *Rpl21* 遺伝子（KH.C2.744）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

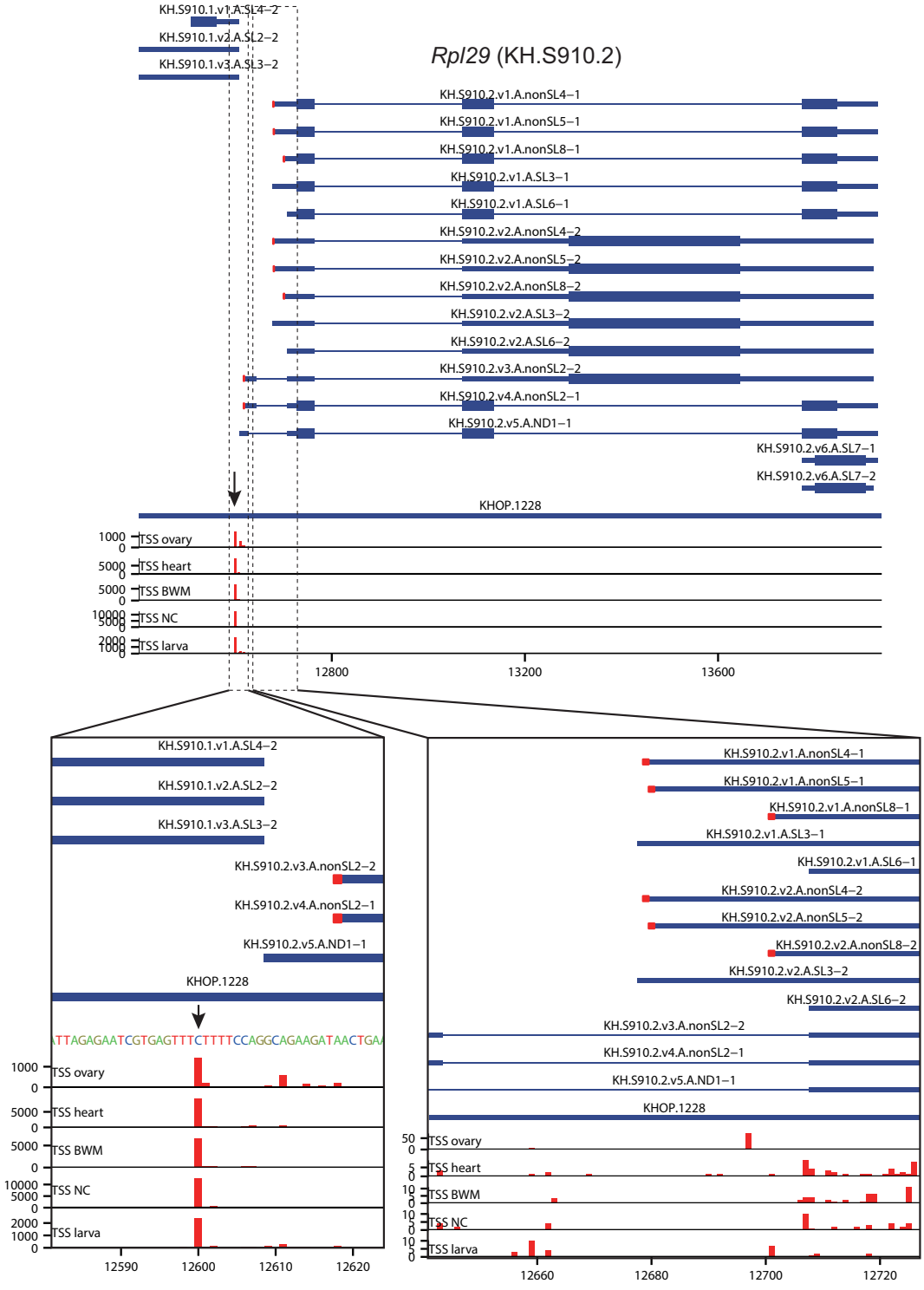

図 S5: *Rpl29* 遺伝子（KH.S910.2）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

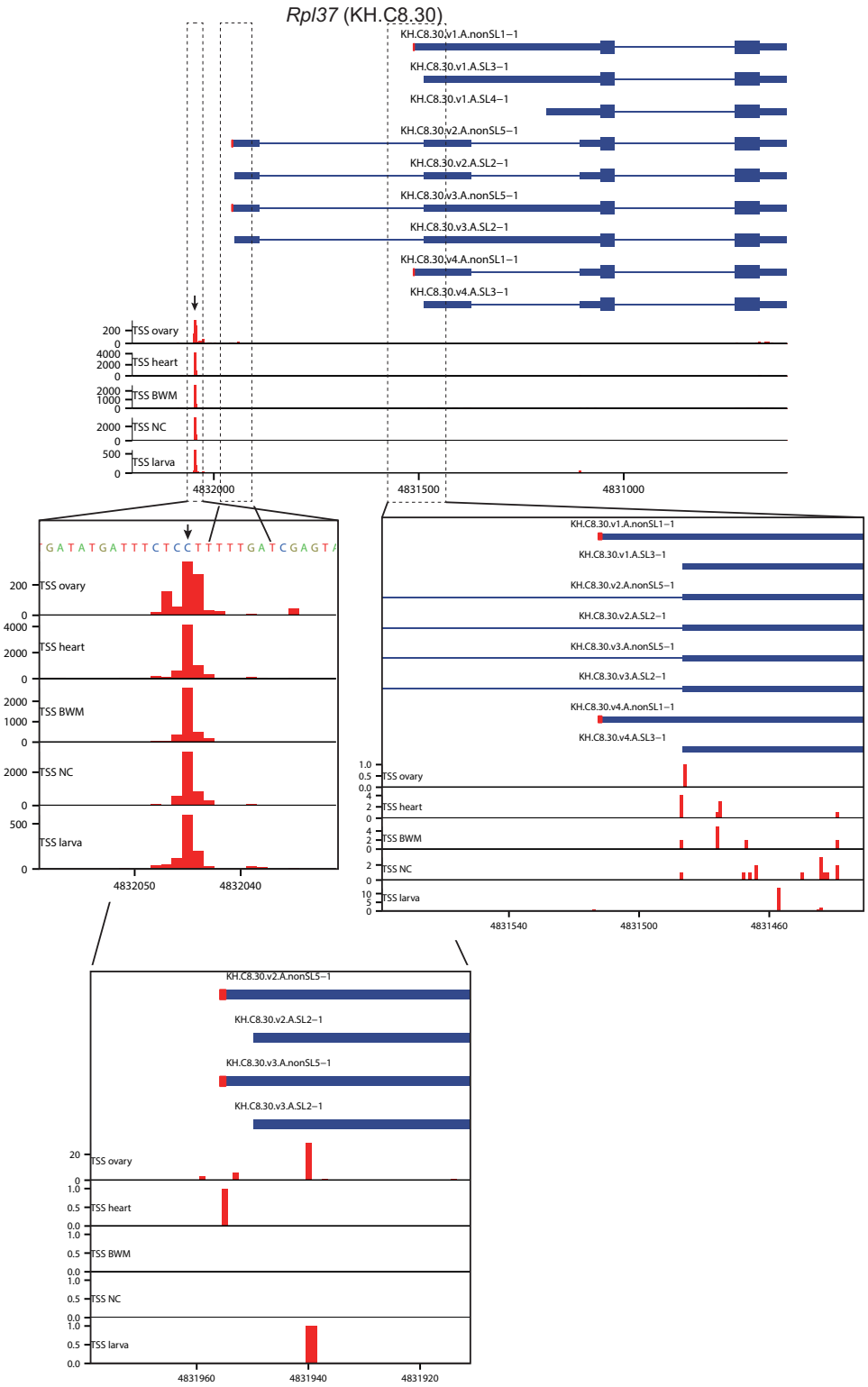

図 S6: *Rpl37* 遺伝子（KH.C8.30）の代表 TSS および既知 TSS 付近の TSS-seq タグの分布。代表 TSS および既知 TSS をそれぞれ黒矢印と赤の四角で示す。赤の棒はマップされた TSS-seq の分布を示す。BWM および NC はそれぞれ body wall muscle（体壁筋）と neural complex（神経複合体）を表す。

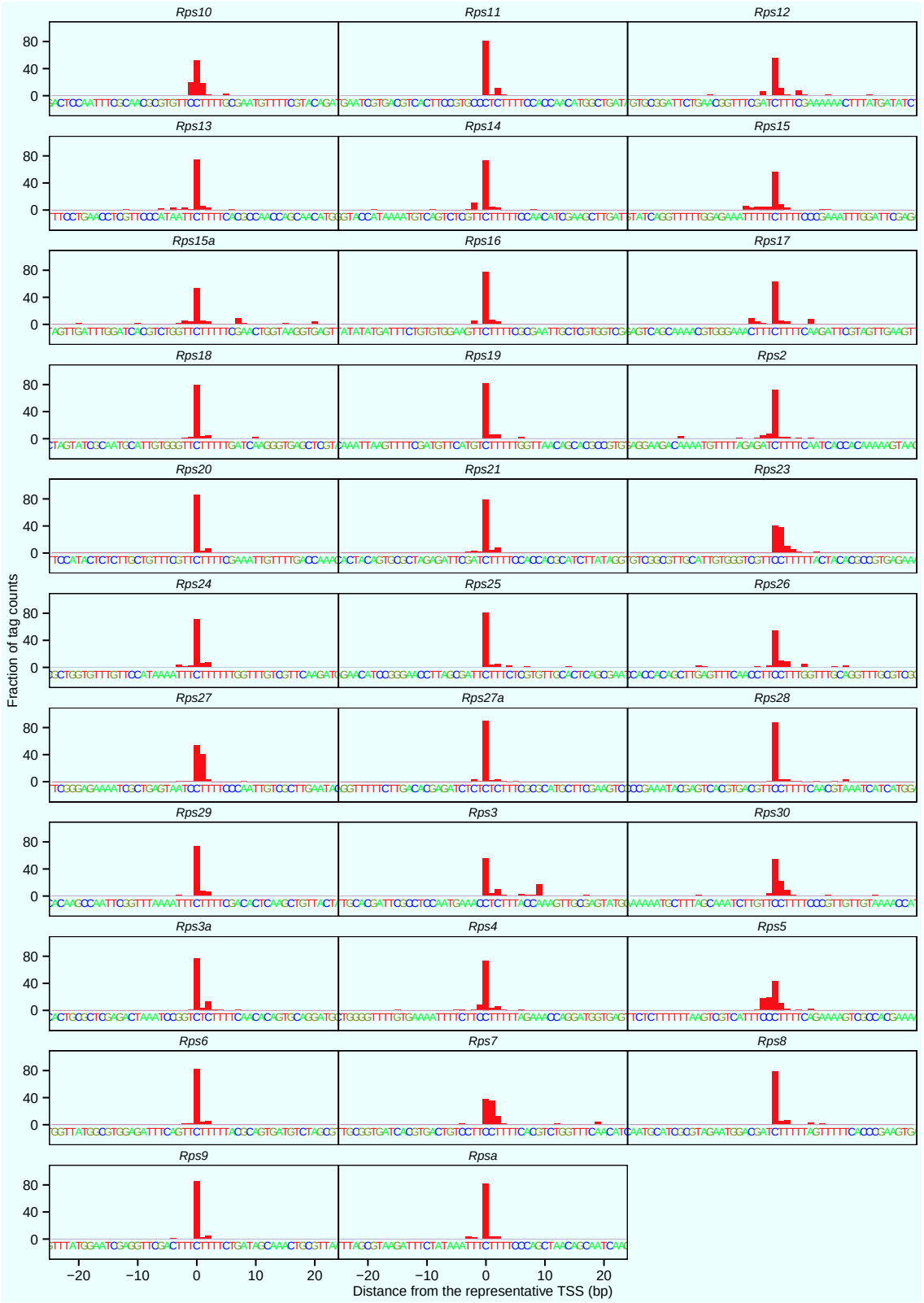

図 S7: リボソームタンパク質の小サブユニットをコードする遺伝子の TSS 分布。

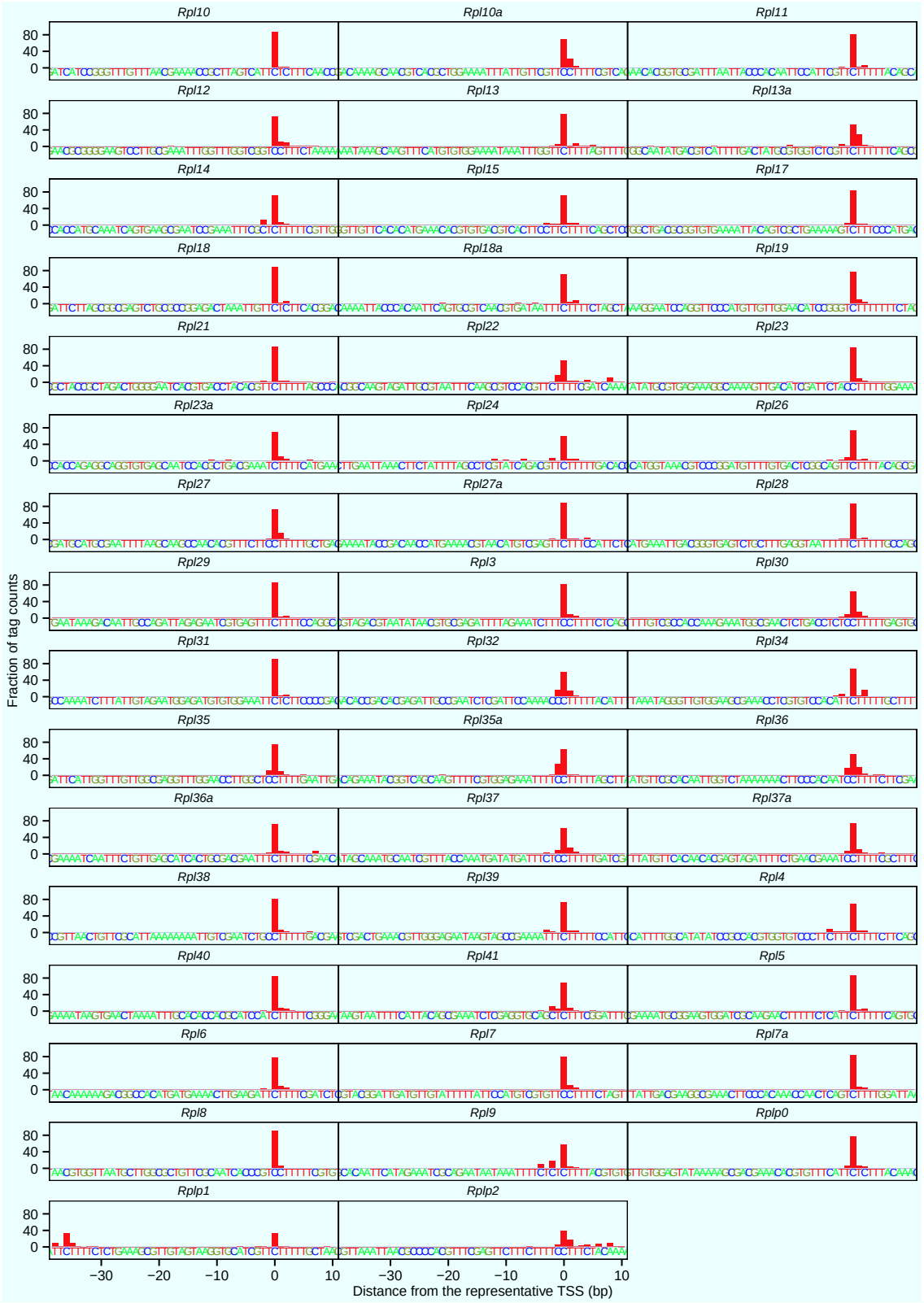

図 S8: リボソームタンパク質の大サブユニットをコードする遺伝子の TSS 分布。

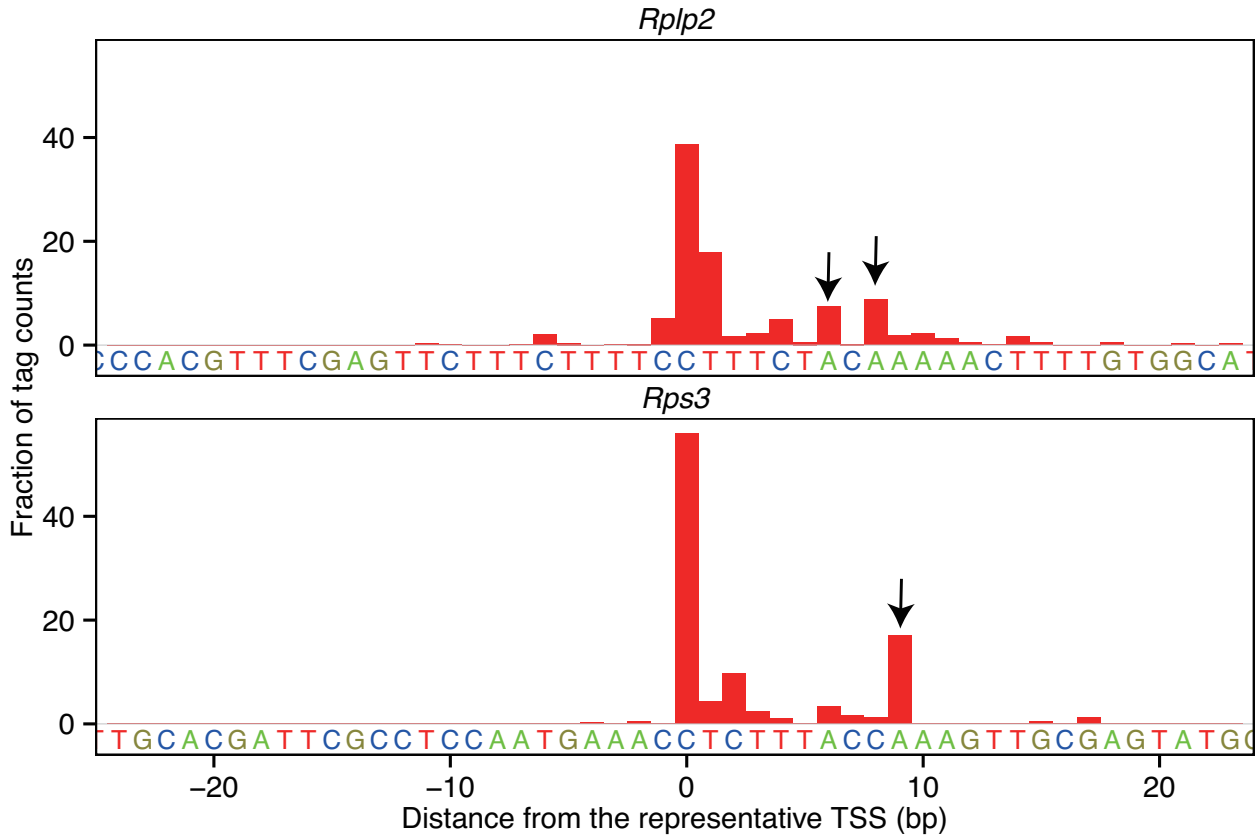

図 S9: RP 遺伝子のマイナー TSS の例。図は *Rplp2* と *Rps3* 遺伝子の TSS 分布を示す。矢印は PyPu の Pu から始まるマイナー TSS を表す。

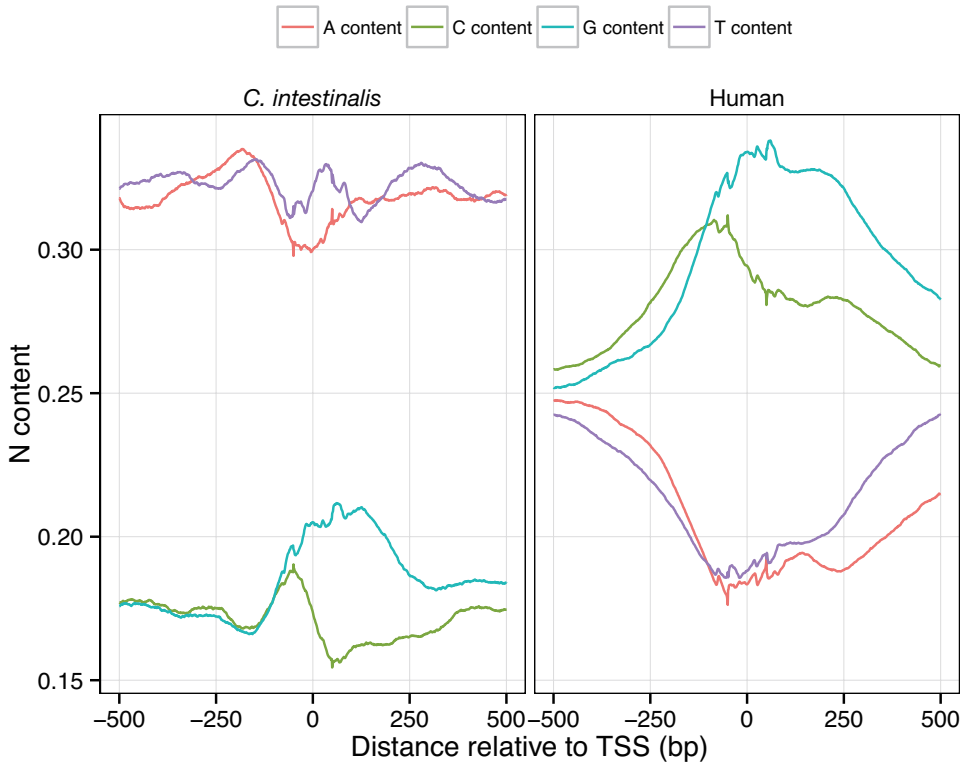

図 S10: A、C、G、T 含量の分布。Non-RP 遺伝子プロモーターにおける A、C、G、T 含量の分布を 100-bp のスライディングウィンドウを用いて調べた。N 含量はウィンドウ中の N の数 / 100 と定義された。ここで、N は A、C、G、T のいずれかである。x 軸と y 軸はそれぞれ代表 TSS からの距離と平均 N 含量を表す。

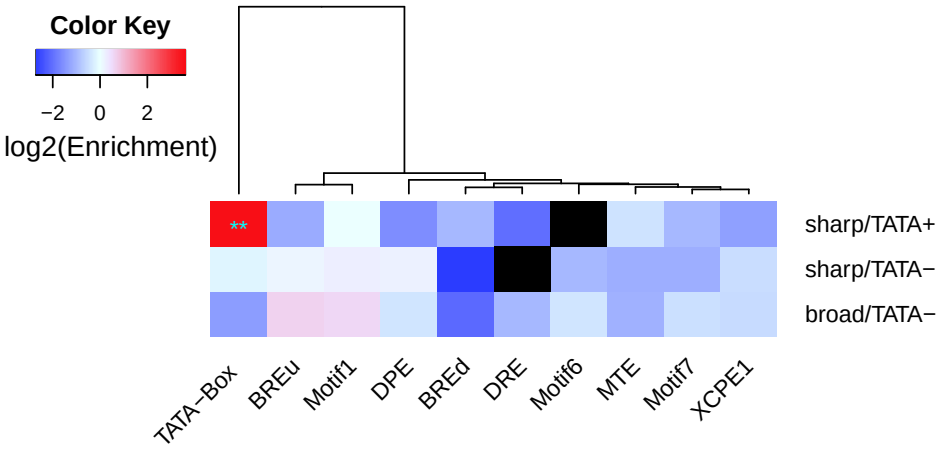

図 S11: コアプロモーターモチーフの Enrichment。各プロモータークラスにおいて、既知コアプロモーターモチーフ（TATA-box、BRE^{d}、BRE^{u}、DPE、MTE、XCPE1）と DRE、motif 1、6、7 のコアプロモーター領域（−60 to +49）における Enrichment を調べた。コアプロモーター領域内の各モチーフの存在は FIMO (Grant et al., 2011) を用いて予測した（デフォルトパラメータを使用)。各モチーフの Enrichment は、そのモチーフを持つコアプロモーター領域の数/全コアプロモーター配列の数と定義された。各モチーフのバックグラウンドレベルを評価するために、コアプロモーター配列と同数で同じ長さのランダム遺伝子間配列を 10 セット作成し、モチーフを探索した。ランダム遺伝子間配列の各セットにおいて各モチーフの Enrichment を計算し、10 個の Enrichment の最大値をそのモチーフのバックグラウンドレベルとした。図のヒートマップは各プロモータークラスにおける各モチーフのバックグラウンドレベルに対する Enrichment の強さを示す。二項検定を用いて、コアプロモーター領域とバックグラウンド間の Enrichment の統計的有意差を調べた。アスタリスク（*と**）は、Bonferroni 補正された後の P 値（$P < 0.05$ と $P < 0.01$）を表す。黒色のセルはコアプロモーター領域にモチーフが存在していなかったことを示す。

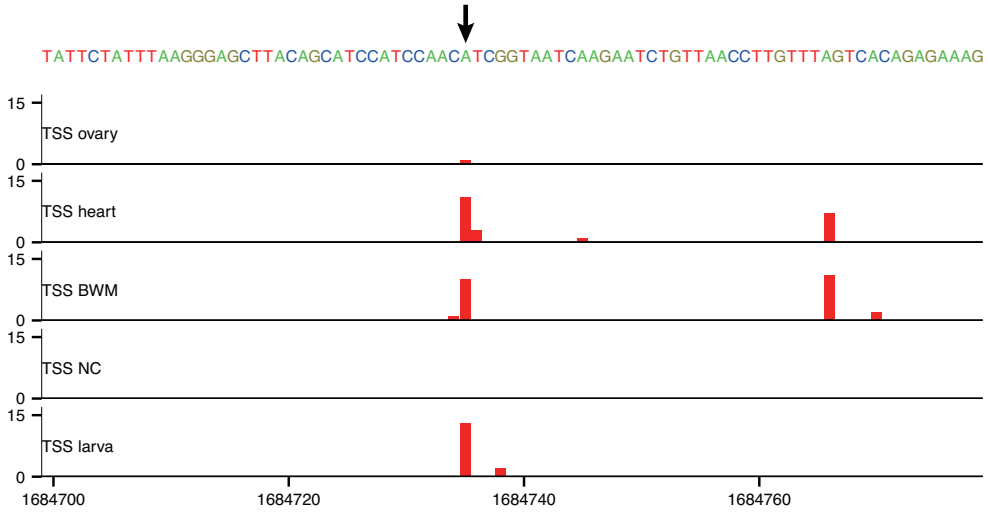

図 S12: *Troponin I* 遺伝子の TSS。矢印は *Troponin I* 遺伝子の TSS を示す。赤のバーは本研究で同定された TSS を表す。y 軸はタグ数を表す。BWM, body wall muscle （体壁筋）; NC, neural complex （神経複合体)。

付録 C　Supplemental Tables

表 S1: ヒトにおいて除去した TSC の数。各列は右から順に「TSC の位置」,「初期 TSC の数」、「CTGG TSC の数」、「A+T-rich な幅 1-bp の TSC の数」、「 逆鎖のスプライスドナーサイト付近にある幅 1-bp の TSC の数」、「CDS もしくは 3′ UTR 上に存在する切断された TSC 由来の可能性がある TSC の数」、「除去された TSC の総数」、「最終的に残った TSC の数」を表す。括弧の中の数値は、初期 TSC に対する各 TSC の割合を表す。

Location	initial	CTGG	A+T-rich	donor	CDS+3′ UTR	removed	final
TSS	5207 (100)	29 (0.6)	-	0 (0)	105 (2.0)	134 (2.6)	5073 (97.4)
5′ UTR	1452 (100)	4 (0.3)	-	0 (0)	112 (7.7)	116 (8.0)	1336 (92.0)
CDS	1622 (100)	28 (1.7)	-	0 (0)	1594 (98.3)	1622 (100)	0 (0.0)
3′ UTR	1312 (100)	17 (1.3)	-	0 (0)	1295 (98.7)	1312 (100)	0 (0.0)
exon(ncRNA)	273 (100)	3 (1.1)	-	0 (0)	26 (9.5)	29 (10.6)	244 (89.4)
intron	1650 (100)	29 (1.8)	38 (2.3)	0 (0)	34 (2.1)	101 (6.1)	1549 (93.9)
intergenic	3982 (100)	648 (16.3)	85 (2.1)	2 (0.1)	0 (0)	735 (18.5)	3247 (81.5)
total	15498 (100)	758 (4.9)	123 (0.8)	2 (0.0)	3166 (20.4)	4049 (26.1)	11449 (73.9)

表 S2: 高頻度に出現する CTGG の変種。CTGG の変種が既知 TSS 上の TSC に比べて、エキソンのアンチセンス鎖上の TSC で高頻度に出現するかどうかをフィッシャーの直接確率検定で調べた。ここで、CTGG の変種とは、CTGG 配列と比較して最大で 2 塩基のミスマッチをもつ 4 塩基の配列である。表はカタユウレイボヤにおいて、少なくとも 1 つのエキソンのアンチセンス鎖上の TSC に出現する CTGG の変種を示す。アスタリスク（*と**）は、CTGG の変種がエキソンのアンチセンス鎖上の TSC で有意に高頻度に出現することを意味する（q-value<0.05 と q-value<0.01）。q-value は R パッケージである “qvalue” (Dabney et al., 2013) を用いて計算された。

CTGG variant	# (on antisense exon)	%	# (in TSS)	%	q-value
TTGG**	90	9.5	0	0	1.77E-30
CTGA**	81	8.5	0	0	1.34E-27
CCGG**	71	7.5	0	0	2.88E-24
TCGG**	46	4.8	0	0	9.63E-16
CAGG**	45	4.7	0	0	1.70E-15
ATGG**	37	3.9	0	0	7.61E-13
GTGG**	27	2.8	0	0	1.60E-09
CTGC**	24	2.5	8	0.7	1.51E-03
CTCG**	23	2.4	0	0	3.13E-08
CCAG**	17	1.8	0	0	2.89E-06
CGGG**	13	1.4	2	0.2	1.82E-03
CTTG**	13	1.4	2	0.2	1.82E-03
CTAG**	11	1.2	0	0	2.20E-04
CCGA**	10	1.1	0	0	4.07E-04
ACGG**	10	1.1	0	0	4.07E-04
AAGG**	10	1.1	1	0.1	2.53E-03
CAGA**	8	0.8	0	0	1.65E-03
TGGG**	7	0.7	0	0	2.85E-03
CTTT	6	0.6	36	3.3	1.51E-05
ATGC	6	0.6	8	0.7	0.24
CTGT	6	0.6	2	0.2	0.07
GAGG*	6	0.6	1	0.1	0.03
TTGC	4	0.4	11	1	0.08
CTTC	4	0.4	30	2.8	1.91E-05
TAGG	3	0.3	1	0.1	0.11

CTGG variant	# (on antisense exon)	%	# (in TSS)	%	q-value
GCGG	3	0.3	1	0.1	0.11
GTGT	3	0.3	7	0.6	0.11
CGAG	3	0.3	1	0.1	0.11
TTGA	3	0.3	0	0	0.05
GTGA	3	0.3	0	0	0.05
TTGT	3	0.3	4	0.4	0.24
CAGT	2	0.2	7	0.6	0.08
CAAG	2	0.2	1	0.1	0.16
CTCC	2	0.2	3	0.3	0.24
TTAG	2	0.2	0	0	0.08
GGGG	2	0.2	0	0	0.08
ATGA	2	0.2	0	0	0.08
CTAA	2	0.2	0	0	0.08
GTAG	2	0.2	0	0	0.08
TTTG	2	0.2	0	0	0.08
CTCT	1	0.1	5	0.5	0.08
CCGC	1	0.1	6	0.6	0.06
CGGT	1	0.1	7	0.6	0.04
CATG	1	0.1	0	0	0.13
CTAT	1	0.1	2	0.2	0.24
CGTG	1	0.1	0	0	0.13
ATGT	1	0.1	3	0.3	0.17
CACG	1	0.1	0	0	0.13
CGGA	1	0.1	0	0	0.13

表 S3: RP 遺伝子の転写開始点。表は、カタユウレイボヤにおいて同定した RP 遺伝子の正確な転写開始点の位置を示す。Distance は、最も近い annotated TSS までの距離を表す。TATA は、予測された TATA box の有無を表す。TSSD は、転写開始点分布のタイプを表す。

RP	Gene	Str	TSS	Distance	TATA	TSSD
RPSA	KH.C11.148	+	4605968	0	(-)	sharp
RPS2	KH.C5.184	-	2224706	15	(-)	sharp
RPS3	KH.C2.108	+	2417046	0	(-)	sharp
RPS3A	KH.C2.693	+	2684394	0	(-)	sharp
RPS4	KH.C12.188	+	398620	0	(-)	sharp
RPS5	KH.C13.29	-	1066918	-545	(-)	sharp
RPS6	KH.L40.23	+	74912	0	(-)	sharp
RPS7	KH.S761.1	-	17725	0	(-)	sharp
RPS8	KH.L133.9	+	163	0	(-)	sharp
RPS9	KH.C9.358	+	955038	0	(-)	sharp
RPS10	KH.L170.101	+	357395	0	(-)	sharp
RPS11	KH.C10.57	-	364848	0	(-)	sharp
RPS12	KH.C7.264	+	2464977	0	(-)	sharp
RPS13	KH.C11.268	+	977235	0	(-)	sharp
RPS14	KH.C14.231	+	1117739	0	(-)	sharp
RPS15	KH.S852.1	-	17301	0	(-)	sharp
RPS15A	KH.L10.10	-	245674	0	(-)	sharp
RPS16	KH.S406.5	+	19717	0	(+)	sharp
RPS17	KH.C8.322	-	3013236	0	(-)	sharp
RPS18	KH.C12.698	+	1592036	0	(-)	sharp
RPS19	KH.C10.101	+	1087698	0	(-)	sharp
RPS20	KH.C8.308	+	5021607	0	(-)	sharp
RPS21	KH.L147.12	+	308767	-524	(-)	sharp
RPS23	KH.L59.2	-	55120	0	(-)	sharp
RPS24	KH.C14.120	-	1965871	0	(-)	sharp
RPS25	KH.L95.1	-	54286	0	(-)	sharp
RPS26	KH.C2.257	+	513278	0	(-)	sharp
RPS27	KH.C3.248	-	5020968	0	(-)	sharp
RPS27A	KH.C10.239	+	828362	0	(-)	sharp
RPS28	KH.C8.209	-	2359911	0	(-)	sharp

RP	Gene	Str	TSS	Distance	TATA	TSSD
RPS29	KH.C1.115	-	2975443	0	(-)	sharp
RPS30	KH.C10.10	-	828091	0	(-)	sharp
RPL3	KH.C2.198	+	7077677	0	(-)	sharp
RPL4	KH.C1.263	+	6435388	0	(-)	sharp
RPL5	KH.C9.386	+	4526102	0	(-)	sharp
RPL6	KH.L41.49	-	115469	0	(-)	sharp
RPL7	KH.C2.77	+	6659849	0	(-)	sharp
RPL7A	KH.C3.237	-	809952	0	(-)	sharp
RPL8	KH.L123.1	+	66846	0	(-)	sharp
RPL9	KH.C14.209	+	1675066	0	(-)	sharp
RPL10	KH.C12.107	-	4957040	0	(-)	sharp
RPL10A	KH.C2.28	+	7439737	0	(-)	sharp
RPL11	KH.L22.58	+	445445	0	(-)	sharp
RPL12	KH.C9.73	-	5150572	0	(-)	sharp
RPL13	KH.S793.1	-	17774	0	(-)	sharp
RPL13A	KH.C6.46	-	586197	0	(-)	sharp
RPL14	KH.L39.8	-	18620	0	(-)	sharp
RPL15	KH.S595.6	-	26554	0	(-)	sharp
RPL17	KH.C2.141	-	1833902	0	(-)	sharp
RPL18	KH.C9.355	-	3738118	0	(-)	sharp
RPL18A	KH.C8.109	+	5232761	0	(-)	sharp
RPL19	KH.C1.333	+	6706573	0	(-)	sharp
RPL21	KH.C2.744	-	1398665	-1733	(-)	sharp
RPL22	KH.C1.296	-	2810711	0	(-)	sharp
RPL23	KH.L24.2	+	64332	0	(-)	sharp
RPL23A	KH.L152.7	-	5730	0	(-)	sharp
RPL24	KH.C6.158	-	1915621	0	(-)	sharp
RPL26	KH.C7.70	-	572729	0	(-)	sharp
RPL27	KH.L141.29	-	397818	0	(-)	sharp
RPL27A	KH.C10.212	+	752607	0	(-)	sharp
RPL28	KH.C9.87	+	4460129	0	(-)	sharp
RPL29	KH.S910.2	+	12600	-18	(-)	sharp
RPL30	KH.C2.631	-	6802560	0	(-)	sharp

RP	Gene	Str	TSS	Distance	TATA	TSSD
RPL31	KH.C2.551	+	2411222	0	(-)	sharp
RPL32	KH.C5.166	+	3392501	0	(-)	sharp
RPL34	KH.S1840.1	+	1179	0	(-)	sharp
RPL35	KH.C9.404	+	5128873	0	(-)	sharp
RPL35A	KH.C7.10	-	3302392	0	(-)	sharp
RPL36	KH.C7.196	-	1499243	0	(-)	sharp
RPL36A	KH.C14.360	+	2927170	0	(-)	sharp
RPL37	KH.C8.30	-	4832045	-90	(-)	sharp
RPL37A	KH.C9.517	+	3600439	0	(-)	sharp
RPL38	KH.C8.20	+	188030	0	(-)	sharp
RPL39	KH.L20.54	+	4292	0	(-)	sharp
RPL40	KH.C4.189	-	3575352	0	(-)	sharp
RPL41	KH.C9.469	-	1204471	n/a	(-)	sharp
RPLP0	KH.C11.21	+	4986946	0	(+)	sharp
RPLP1	KH.C2.216	-	227537	0	(-)	other
RPLP2	KH.C11.339	+	5012447	0	(-)	sharp

表 S4: TAC と TSC のペア。表は、同じ組織特異性を示した TAC とその上流に存在する TSC のペアの座標を示す。TSC が既知の転写産物の TSS 上にあるときは、その転写産物の ID を示す。putative は、TSC が既知 TSS 上になく、その TSC が推定プロモーターであることを表す。Dist は、TAC と TSC の距離、すなわちアウトロンの長さを表す（non-operon-type の場合)。non-op および unann-op は、それぞれ non-operon-type と unannotated-operon-type を表す。

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
1	KH.C6.81.v1.A.nonSL4-1	non-op	KhC6	+	1167747	1167694	53
2	KH.C1.384.v1.A.nonSL6-1	non-op	KhC1	+	8948441	8948384	57
3	KH.C1.1026.v1.A.ND1-1	non-op	KhC1	+	6676118	6676058	60
4	KH.L7.2.v1.A.nonSL7-1	non-op	KhL7	+	112700	112640	60
5	KH.C12.284.v1.A.nonSL5-1	non-op	KhC12	-	1041980	1042041	61
6	KH.C12.308.v1.A.nonSL4-1	non-op	KhC12	-	640033	640094	61
7	KH.L34.6.v1.C.nonSL20-1	non-op	KhL34	+	93707	93643	64
8	KH.C4.550.v1.A.nonSL3-1	non-op	KhC4	-	2915560	2915627	67
9	KH.C7.164.v1.C.ND4-1	non-op	KhC7	-	4802686	4802756	70
10	KH.C3.108.v1.A.nonSL4-1	non-op	KhC3	+	1390781	1390708	73
11	KH.C14.279.v1.A.nonSL7-1	non-op	KhC14	+	4032289	4032212	77
12	KH.S521.3.v1.A.nonSL2-1	non-op	KhS521	+	24716	24639	77
13	KH.C8.506.v1.A.nonSL18-1	non-op	KhC8	+	5414381	5414302	79
14	KH.C3.268.v1.A.nonSL3-1	non-op	KhC3	-	2034253	2034333	80
15	KH.C1.171.v1.A.nonSL13-1	non-op	KhC1	-	7244116	7244198	82
16	KH.C3.87.v1.A.nonSL17-1	non-op	KhC3	-	6539226	6539308	82
17	KH.C12.617.v1.A.ND1-1	non-op	KhC12	+	5307877	5307793	84
18	KH.C2.61.v1.A.nonSL11-1	non-op	KhC2	+	1641389	1641305	84
19	KH.C2.82.v1.A.ND4-1	non-op	KhC2	+	547344	547260	84
20	KH.L41.66.v1.C.nonSL12-1	non-op	KhL41	-	316886	316971	85
21	KH.C8.402.v1.A.ND2-2	non-op	KhC8	-	398750	398836	86
22	KH.C10.66.v1.A.ND3-1	non-op	KhC10	-	3955779	3955870	91
23	KH.C2.257.v2.A.nonSL1-1	non-op	KhC2	+	513369	513278	91
24	KH.C7.149.v1.A.ND2-1	non-op	KhC7	-	880979	881071	92
25	KH.L170.18.v1.A.nonSL7-1	non-op	KhL170	-	136221	136313	92
26	KH.C5.629.v1.A.nonSL16-1	non-op	KhC5	-	1164355	1164449	94
27	KH.C13.33.v1.A.nonSL5-1	non-op	KhC13	-	1772827	1772923	96
28	KH.C8.180.v1.A.ND2-1	non-op	KhC8	+	2991188	2991091	97

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
29	KH.C5.227.v1.A.nonSL5-1	non-op	KhC5	+	3033409	3033311	98
30	KH.C12.308.v1.A.nonSL4-1	non-op	KhC12	-	639995	640094	99
31	KH.C9.593.v1.A.nonSL4-1	non-op	KhC9	+	2389127	2389026	101
32	KH.L18.86.v1.A.ND2-1	non-op	KhL18	-	755114	755223	109
33	KH.L116.44.v1.C.ND1-1	non-op	KhL116	+	510880	510768	112
34	KH.L41.79.v1.A.nonSL19-1	non-op	KhL41	-	87025	87137	112
35	KH.C3.88.v1.A.nonSL2-1	non-op	KhC3	-	673218	673331	113
36	KH.C2.141.v1.A.nonSL2-1	non-op	KhC2	-	1833785	1833902	117
37	KH.C7.79.v1.B.ND1-1	non-op	KhC7	-	2329369	2329487	118
38	KH.C3.207.v1.A.ND2-1	non-op	KhC3	-	3500725	3500851	126
39	KH.L51.6.v1.A.nonSL2-1	non-op	KhL51	-	17262	17390	128
40	KH.C9.212.v1.A.ND4-1	non-op	KhC9	+	5929375	5929246	129
41	KH.C12.107.v1.A.nonSL1-1	non-op	KhC12	-	4956902	4957040	138
42	KH.L8.1.v2.A.ND5-1	non-op	KhL8	+	195414	195268	146
43	KH.C9.358.v1.A.nonSL3-1	non-op	KhC9	+	955200	955038	162
44	KH.C1.1048.v1.A.ND1-1	non-op	KhC1	+	8071994	8071792	202
45	KH.S882.2.v1.A.nonSL4-1	non-op	KhS882	+	10430	10223	207
46	KH.C2.10.v1.A.nonSL5-2	non-op	KhC2	+	7121397	7121184	213
47	KH.L84.11.v2.A.nonSL5-1	non-op	KhL84	-	35290	35503	213
48	KH.C7.10.v1.A.nonSL4-1	non-op	KhC7	-	3302152	3302392	240
49	KH.C11.22.v3.A.nonSL33-1	non-op	KhC11	-	5139838	5140088	250
50	KH.L101.1.v3.A.nonSL6-1	non-op	KhL101	-	16450	16703	253
51	KH.C7.70.v2.C.nonSL1-1	non-op	KhC7	-	572474	572729	255
52	KH.C14.52.v1.A.nonSL1-1	non-op	KhC14	-	786365	786629	264
53	KH.C11.674.v1.A.nonSL6-1	non-op	KhC11	-	3753777	3754099	322
54	KH.S390.2.v1.A.nonSL9-1	non-op	KhS390	+	60478	60154	324
55	KH.C2.714.v1.A.nonSL6-1	non-op	KhC2	+	2629844	2629509	335
56	KH.C14.397.v1.A.nonSL5-1	non-op	KhC14	+	2024360	2024020	340
57	KH.L18.44.v4.A.nonSL6-2	non-op	KhL18	+	920760	920412	348
58	KH.C2.412.v1.A.nonSL8-1	non-op	KhC2	-	5559742	5560136	394
59	KH.L81.2.v1.A.nonSL7-1	non-op	KhL81	-	15798	16254	456
60	KH.C14.191.v3.A.nonSL10-1	non-op	KhC14	+	3263233	3262628	605
61	KH.C14.261.v1.B.nonSL2-1	non-op	KhC14	-	4136447	4137085	638

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
62	KH.L107.2.v2.A.nonSL17-1	non-op	KhL107	-	9162	9875	713
63	KH.L65.8.v1.A.nonSL2-1	non-op	KhL65	-	89621	90502	881
64	KH.C8.433.v2.B.nonSL2-1	non-op	KhC8	-	606846	608134	1288
65	KH.L10.10.v2.A.nonSL11-1	non-op	KhL10	-	243901	245674	1773
66	KH.L172.30.v2.A.ND2-1	non-op	KhL172	-	103947	105780	1833
67	putative	non-op	KhC1	-	2084190	2084245	55
68	putative	non-op	KhC4	+	4634060	4634005	55
69	putative	non-op	KhC13	+	1453841	1453782	59
70	putative	non-op	KhC9	+	5942580	5942521	59
71	putative	non-op	KhC12	+	1249594	1249528	66
72	putative	non-op	KhC5	-	4178088	4178154	66
73	putative	non-op	KhC9	-	3396556	3396625	69
74	putative	non-op	KhC10	-	1429147	1429218	71
75	putative	non-op	KhC12	+	1249603	1249528	75
76	putative	non-op	KhC11	-	2664941	2665017	76
77	putative	non-op	KhC11	+	1181525	1181447	78
78	putative	non-op	KhS215	+	63867	63787	80
79	putative	non-op	KhC3	-	3242071	3242158	87
80	putative	non-op	KhC7	+	5816699	5816609	90
81	putative	non-op	KhS681	+	7688	7598	90
82	putative	non-op	KhC1	-	2906408	2906500	92
83	putative	non-op	KhC12	+	913131	913038	93
84	putative	non-op	KhS437	+	14979	14886	93
85	putative	non-op	KhL3	+	152872	152778	94
86	putative	non-op	KhC12	+	913137	913038	99
87	putative	non-op	KhS966	+	3584	3485	99
88	putative	non-op	KhC4	-	2019941	2020047	106
89	putative	non-op	KhC9	-	5997804	5997912	108
90	putative	non-op	KhC3	+	5037382	5037271	111
91	putative	non-op	KhC12	+	3170846	3170734	112
92	putative	non-op	KhC2	-	7050280	7050396	116
93	putative	non-op	KhC3	-	3329490	3329606	116
94	putative	non-op	KhL141	-	513224	513341	117

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
95	putative	non-op	KhC2	+	1233415	1233296	119
96	putative	non-op	KhL44	-	18784	18903	119
97	putative	non-op	KhC3	+	5939450	5939328	122
98	putative	non-op	KhC4	+	4634127	4634005	122
99	putative	non-op	KhC10	+	3758639	3758516	123
100	putative	non-op	KhL89	-	13146	13275	129
101	putative	non-op	KhC12	-	2293623	2293753	130
102	putative	non-op	KhC8	+	4374625	4374495	130
103	putative	non-op	KhC10	+	4187267	4187135	132
104	putative	non-op	KhC14	-	3002680	3002813	133
105	putative	non-op	KhL18	-	321817	321950	133
106	putative	non-op	KhL41	-	326833	326966	133
107	putative	non-op	KhS606	+	32044	31905	139
108	putative	non-op	KhC3	+	5037414	5037271	143
109	putative	non-op	KhC11	-	4079597	4079749	152
110	putative	non-op	KhL18	-	321794	321950	156
111	putative	non-op	KhC1	+	6664990	6664830	160
112	putative	non-op	KhC9	-	4467514	4467698	184
113	putative	non-op	KhC2	-	1690022	1690228	206
114	putative	non-op	KhC4	-	4923213	4923425	212
115	putative	non-op	KhL9	-	51648	51862	214
116	putative	non-op	KhC5	-	3066275	3066490	215
117	putative	non-op	KhC6	-	1311700	1311920	220
118	putative	non-op	KhC3	+	3044662	3044428	234
119	putative	non-op	KhL28	+	38303	38062	241
120	putative	non-op	KhL131	-	59768	60013	245
121	putative	non-op	KhL157	-	19594	19842	248
122	putative	non-op	KhC11	+	3855628	3855377	251
123	putative	non-op	KhL142	-	44814	45065	251
124	putative	non-op	KhC3	+	3044697	3044428	269
125	putative	non-op	KhC5	-	831161	831431	270
126	putative	non-op	KhC2	-	1689952	1690228	276
127	putative	non-op	KhL84	+	180005	179723	282

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
128	putative	non-op	KhC1	-	6345762	6346049	287
129	putative	non-op	KhC2	-	1465666	1465958	292
130	putative	non-op	KhL37	-	31908	32206	298
131	putative	non-op	KhC2	-	7438953	7439261	308
132	putative	non-op	KhL154	-	172732	173043	311
133	putative	non-op	KhC7	+	3162061	3161747	314
134	putative	non-op	KhC4	+	2775253	2774931	322
135	putative	non-op	KhC11	-	4789814	4790149	335
136	putative	non-op	KhS815	-	19942	20278	336
137	putative	non-op	KhC6	+	1576998	1576657	341
138	putative	non-op	KhC3	-	4157363	4157710	347
139	putative	non-op	KhC1	+	2461736	2461387	349
140	putative	non-op	KhC3	+	2146664	2146314	350
141	putative	non-op	KhC3	-	6348818	6349174	356
142	putative	non-op	KhC3	-	5032096	5032456	360
143	putative	non-op	KhC3	+	2257607	2257237	370
144	putative	non-op	KhC12	+	4259708	4259330	378
145	putative	non-op	KhC1	+	7712496	7712117	379
146	putative	non-op	KhC3	+	2146694	2146314	380
147	putative	non-op	KhC2	-	5467795	5468181	386
148	putative	non-op	KhC3	+	6136786	6136388	398
149	putative	non-op	KhC4	-	4931894	4932293	399
150	putative	non-op	KhC1	-	9313249	9313652	403
151	putative	non-op	KhS815	-	19872	20278	406
152	putative	non-op	KhC12	-	4677364	4677772	408
153	putative	non-op	KhS597	+	21246	20838	408
154	putative	non-op	KhC10	+	219052	218642	410
155	putative	non-op	KhC7	+	5696247	5695832	415
156	putative	non-op	KhL94	+	12124	11703	421
157	putative	non-op	KhC4	-	2436258	2436688	430
158	putative	non-op	KhC10	+	4458685	4458254	431
159	putative	non-op	KhC1	+	4879829	4879397	432
160	putative	non-op	KhL76	-	183843	184275	432

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
161	putative	non-op	KhC4	-	2859915	2860352	437
162	putative	non-op	KhL76	-	183833	184275	442
163	putative	non-op	KhS815	-	19825	20278	453
164	putative	non-op	KhL108	+	489804	489348	456
165	putative	non-op	KhC7	-	4511697	4512168	471
166	putative	non-op	KhC2	+	629519	629046	473
167	putative	non-op	KhC12	-	3380101	3380576	475
168	putative	non-op	KhC1	+	8729379	8728899	480
169	putative	non-op	KhC7	-	2110713	2111195	482
170	putative	non-op	KhC3	+	3297067	3296584	483
171	putative	non-op	KhC14	+	2683893	2683409	484
172	putative	non-op	KhL134	+	181257	180769	488
173	putative	non-op	KhS455	+	10052	9563	489
174	putative	non-op	KhC3	+	1804147	1803653	494
175	putative	non-op	KhC7	+	2394918	2394424	494
176	putative	non-op	KhL18	+	626664	626169	495
177	putative	non-op	KhC4	-	4931780	4932293	513
178	putative	non-op	KhC11	+	3682153	3681634	519
179	putative	non-op	KhC7	+	1333215	1332690	525
180	putative	non-op	KhC3	-	6377215	6377747	532
181	putative	non-op	KhS256	+	47334	46792	542
182	putative	non-op	KhL37	-	60629	61216	587
183	putative	non-op	KhC14	+	4027850	4027258	592
184	putative	non-op	KhC11	+	1474401	1473741	660
185	putative	non-op	KhC2	+	4197040	4196378	662
186	putative	non-op	KhS854	+	696	13	683
187	putative	non-op	KhS256	+	47489	46792	697
188	putative	non-op	KhC4	+	5136141	5135408	733
189	putative	non-op	KhS391	-	38391	39129	738
190	putative	non-op	KhC3	-	5932958	5933708	750
191	putative	non-op	KhC1	+	3170168	3169409	759
192	putative	non-op	KhC1	-	7501470	7502241	771
193	putative	non-op	KhL11	-	12615	13426	811

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
194	putative	non-op	KhL144	+	144125	143311	814
195	putative	non-op	KhC5	+	1939057	1938207	850
196	putative	non-op	KhL23	-	57970	58885	915
197	putative	non-op	KhC4	+	4655471	4654541	930
198	putative	non-op	KhL23	-	57883	58885	1002
199	putative	non-op	KhC12	+	3171752	3170734	1018
200	putative	non-op	KhC8	+	721780	720723	1057
201	putative	non-op	KhC3	+	2258296	2257237	1059
202	putative	non-op	KhC1	-	6787051	6788116	1065
203	putative	non-op	KhC8	+	5286188	5285105	1083
204	putative	non-op	KhC1	-	3770189	3771279	1090
205	putative	non-op	KhC9	+	3413238	3412065	1173
206	putative	non-op	KhC11	+	3803116	3801941	1175
207	putative	non-op	KhC11	+	1362854	1361548	1306
208	putative	non-op	KhC10	+	2772993	2771653	1340
209	putative	non-op	KhC12	+	3326051	3324703	1348
210	putative	non-op	KhC11	-	1556577	1557943	1366
211	putative	non-op	KhC5	+	1797085	1795713	1372
212	putative	non-op	KhC3	+	4326823	4325389	1434
213	putative	non-op	KhL168	+	35457	33995	1462
214	putative	non-op	KhC1	-	8392381	8393895	1514
215	putative	non-op	KhC3	+	3046035	3044428	1607
216	putative	non-op	KhL18	+	64111	62458	1653
217	putative	non-op	KhC1	+	7308428	7306770	1658
218	putative	non-op	KhC11	+	508064	506309	1755
219	putative	non-op	KhL4	-	352173	353937	1764
220	putative	non-op	KhC3	+	1805492	1803653	1839
221	putative	non-op	KhL132	-	129949	131903	1954
222	putative	non-op	KhC12	+	156323	154352	1971
223	KH.S933.2.v1.A.nonSL2-1	non-op	KhS933	-	14961	15025	64
224	KH.C5.442.v1.A.ND1-1	non-op	KhC5	+	3632001	3631905	96
225	KH.C2.257.v2.A.nonSL1-1	non-op	KhC2	+	513421	513278	143
226	KH.C1.509.v1.A.nonSL4-1	non-op	KhC1	-	5299622	5299781	159

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
227	KH.C13.127.v1.A.nonSL8-1	non-op	KhC13	+	351125	350927	198
228	KH.C9.223.v1.A.ND1-1	non-op	KhC9	+	4164832	4164542	290
229	KH.L37.43.v1.A.nonSL3-1	non-op	KhL37	+	81746	81347	399
230	KH.C14.191.v3.A.nonSL10-1	non-op	KhC14	+	3263238	3262628	610
231	KH.L5.20.v1.A.nonSL1-1	non-op	KhL5	-	114553	115472	919
232	KH.C13.127.v1.A.nonSL8-1	non-op	KhC13	+	352003	350927	1076
233	putative	non-op	KhC8	-	3396211	3396263	52
234	putative	non-op	KhS2334	-	1744	1798	54
235	putative	non-op	KhC11	-	4420738	4420794	56
236	putative	non-op	KhC5	-	4522588	4522653	65
237	putative	non-op	KhL3	+	203360	203278	82
238	putative	non-op	KhC11	-	2664929	2665017	88
239	putative	non-op	KhL89	-	13186	13275	89
240	putative	non-op	KhC8	+	5308254	5308039	215
241	putative	non-op	KhC1	+	7897636	7897414	222
242	putative	non-op	KhC1	+	710315	710038	277
243	putative	non-op	KhC1	-	628739	629038	299
244	putative	non-op	KhC11	-	2725981	2726303	322
245	putative	non-op	KhC11	-	2725950	2726303	353
246	putative	non-op	KhC4	+	418812	418450	362
247	putative	non-op	KhC11	-	2725893	2726303	410
248	putative	non-op	KhC12	+	3171151	3170734	417
249	putative	non-op	KhC10	-	1831453	1831900	447
250	putative	non-op	KhC1	-	502466	502955	489
251	putative	non-op	KhC5	+	2965549	2965048	501
252	putative	non-op	KhC1	-	502445	502955	510
253	putative	non-op	KhL50	+	38713	38129	584
254	putative	non-op	KhC11	+	2268869	2268267	602
255	putative	non-op	KhC11	+	4421827	4421196	631
256	putative	non-op	KhS2008	-	739	1591	852
257	putative	non-op	KhS534	+	39587	38610	977
258	putative	non-op	KhC5	-	1641490	1642470	980
259	putative	non-op	KhC7	+	1697101	1695941	1160

No.	Transcript	Type	Chrom	Str	TAC	TSC	Dist
260	putative	non-op	KhC9	-	1002387	1003957	1570
261	putative	non-op	KhC1	+	3196379	3194643	1736
262	putative	non-op	KhC1	-	627040	629038	1998
263	KH.C13.23.v1.C.ND1-1	unann-op	KhC13	+	1072571	1071093	1478
264	KH.C11.509.v1.A.ND1-1	unann-op	KhC11	-	2826500	2828216	1716