UCSC hg19 における cDNA リファレンスデータの統合

佐藤智之[i]　知久季倫[i]　安田和基[ii]

An unification method of cDNA reference data in UCSC hg19

Toshiyuki Sato　Suenori Chiku　Kazuki Yasuda

UCSC genome browser[12] から提供されている RNA 転写物のアノテーション情報である knownGene.txt，refGene.txt，ensGene.txt，lincRNAsTranscripts.txt の統合を行い，網羅的な cDNA リファレンス配列を作成した．

(**キーワード**): RNA, cDNA, transcriptome, NGS

1　はじめに

Next generation sequencing (NGS) における whole transcriptome sequencing (WTS) では，TopHat[10] の様に cDNA 配列をゲノム配列にマッピングする方法と cDNA リファレンス配列にマッピングする方法がある．適切な cDNA リファレンス配列を用いれば後者の方が多くの read がマッピングされ，その後 single nucleotide variation (SNV) や indel 等の探索若しくは確認解析において有効である．また Cufflinks[9] による転写物の発現量解析においても，多くの遺伝子アノテーション情報は解析に有利であることが期待される．

そこで世界的に最も利用頻度の高いゲノムデータベースの一つである UCSC genome browser[12] の hg19 を基準に，Known Genes[6]，RefSeq[4]，Ensembl[11]，large intergenic noncoding RNA (lincRNA)[2] らの hg19 へのゲノムマッピング情報をマージすることにより，統合された cDNA リファレンス配列の作成を行った．また Cufflinks[9] で発現解析を行うにあたっては，gene_id, tss_id，p_id の情報が必要であり，これらについても統合を行った．

2　統合に用いたデータ

2.1　RNA annotation 情報

統合対象とした遺伝子情報は，hg19 の UCSC Known Genes[6]，及び hg19 ゲノムにマッピングされた RefSeq[4]，Ensembl[11]，lincRNA[2] である．Known Genes[6] とは，Swiss-Prot/TrEMB (UniProt) と関連する GenBank の mRNA データより自動で構築されているアノテーション情報である．RefSeq[4] は，reference sequence として相応しい遺伝子配列を，NCBI のスタッフが GenBank 等のデータベースから選別している RNA 配列である．Ensembl gene[11] は，EMBL-EBI と Welcome Trust Sanger Institute が行っている Ensembl プロジェクトにおいて，gene-build と呼ばれる手続きによって構築されている．lincRNA catalog[2] は，RNA-Seq データを Cufflinks[9] と Scripture[5] によりアセンブリした結果を filtering して構築しているデータである．

実際に使用したゲノム上の exon 位置情報の集合を表す GenePred 形式のファイルのバージョンと登録数を表 1 に纏めた．RefSeq の予測レコード (ID が XM, XR で始まる) は除いており，また hg19 にマッピングされたレコード数がオリジナルよりも多いのは複数箇所にマッピングされたレコードがあるためである．実際 refGene.txt に含まれる重複を除いた ID 数は 40,448 であった．文献 [2] で作成された lincRNA データは 4,662 遺伝子 14,353 転写物であったが，UCSC にこれよりも登録数が多いデータがあったためにこちらを用いることにした．

2.2　RNA annotation 情報の対応表

UCSC では Known Genes[6] に対応する RefSeq[4] 及び Ensembl gene[11] の転写物単位の対応表を提供している．本解析では表 2 に示した情報を使った．kgXref.txt には refGene.txt との対応も 57,173 レコード記載されているが，knownToRefSeq.txt の方が対応数が多かったためにこちらを用いることにした．

2.3　ゲノム配列

本プロジェクトでは hg19 のゲノム配列を用いているが，他の解析との関係のため 1000 genome[3] プロジェ

[i]サイエンスソリューション部 バイオエンジニアリングチーム シニアマネージャー
[ii]国立国際医療研究センター 糖尿病研究センター 代謝疾患研究部 部長

表 1 統合に用いた GenePred 形式の RNA データ

Name	#original	hg19 file name	date	hg19 #record
Known Genes[6]	80,922	knownGene.txt	2012/02/05	80,922
RefSeq[4]	40,765	refGene.txt	2012/11/11	43,801
Ensembl[11]	205,272	ensGene.txt	2012/10/21	191,891
lincRNA[2]	14,353	lincRNAsTranscripts.txt	2011/11/27	21,630

表 2 RNA データベース間の対応テーブル

Name	date	#record	Contents
kgXref.txt	2012/02/05	80,922	Known Genes の gene symbol
knownToRefSeq.txt	2012/02/05	71,350	Known Genes と RefSeq の対応
knownToEnsembl.txt	2012/10/21	74,020	Known Genes と Ensembl の対応
ensemblToGeneName.txt	2012/10/21	191,891	Ensembl の gene symbol

クトに倣い，haplotype 情報 (ファイル名に_hap を含む) を除いた全てのゲノム配列とした．そのため_gl をファイル名に含む配列にマッピングされた情報も使っている．今後は染色体としてアセンブルされている chr1 〜chr22，chrX，chrY，chrM をアセンブルドゲノム，gl を配列名に含む配列を contig ゲノム，hap を含む配列を haplotype ゲノムと呼び，解析対象であるアセンブルドゲノムと contig ゲノムを合わせてターゲットゲノムと呼ぶことにする．

3 統合方法

3.1 統合手続きの概要

リファレンスゲノムを hg19 と定めたため，UCSC Known Genes に対して他の RNA mapping 情報 (表 1 参照) を統合することにした．この手順を，

1. 前処理
   (a) 各 genePred ファイルの clean up
   (b) 各種 ID の付与
   (c) 重複の削除
2. GenePred ファイルのマージ
   (a) refGene.txt の knownGene.txt へのマージによる knownRefGene.txt の作成
   (b) ensGene.txt の knownRefGene.txt へのマージによる knownRefEnsGene.txt の作成
   (c) lincRNAsTranscripts.txt の knownRefEnsGene.txt へのマージによる knownRefEnsLincGene.txt の作成
3. 後処理
   (a) RNA 配列が同一となるレコードの削除
   (b) Protein coding の UTR の両端の違いによるマージ
   (c) Noncoding の両端の違いによるマージ
   (d) gene_id の確認
   (e) tss_id の付与
   (f) p_id の付与

とした．次節で各項目の詳細を説明する．

3.2 各統合手続きの説明

ここで各項目の説明で使用する用語を定義する．

1. 転写領域のオーバラップ
   同じ strand で転写領域にオーバーラップがある場合で，必ずしも exon の位置に重なりがある必要はない．
2. Exon のオーバラップ
   同じ strand で exon にオーバーラップがある．
3. Exon の共有
   1 つ以上の exon の開始位置と終了位置が完全に一致する．

Exon のオーバラップは本統合手続きでは使用していないが，Ensembl において遺伝子を統合する際に用いている [11]．

gene_id とは Cufflinks[9] で用いられる ID で，同じ遺伝子であることを意味する．一方で gene symbol (gene_name)

はHUGO[7] 等に登録されている遺伝子名であるが，Ensembl ではこれらは必ずしも一致しない．本統合手続きでは gene_id の統合は行うが gene symbol については元のレコードに記載されたままの記号を残している．

1.(a) 各 genePred ファイルの clean up

(1) hg19 hap ゲノムへのマッピング情報の削除．knownGene.txt，refGene.txt，ensGene.txt，lincRNAsTranscripts.txt の内，マッピング先が haplotype ゲノムとなっているレコードを削除．

(2) 連続する exon の統合．knownGene.txt には長さ 0 の intron が 22 transcripts 存在 (exon がゲノム上で連続している) したため，これを一つの exon に繋げた．

(3) knownGene.txt に含まれる RefSeq で後に NCBI によって RefSeq から削られた 4 つの transcripts (uc011kuw.2, uc010qoz.1, uc010vei.1, uc002wvw.2) を削除した．

(4) ゲノム上の複数箇所にマップされているため同じ transcript_id を持つ複数の RefSeq レコードが合計 621 transcripts あり，これらにリビジョン番号を付加して区別した．

1.(b) 各種 ID の付与

(1) gene_id の付与

i. knownGene
kgXref.txt の 5 番目のカラム (geneSymbol) を gene_name と gene_id に使用した．ただし 29 transcripts は symbol ではなく gene の説明文になっていたため，他と重複しない様に，20 文字を超え無い範囲で説明文のスペースを削除して繋げ-kgX を付与すことで置き換えた．Gene symbol が unknown となっている 2 transcripts には，最終カラムである description から説明文の場合と同様に独自 ID を付与した．

ii. RefSeq
refGene.txt の 13 番目のカラムを gene_name と gene_id として使用した．

iii. Ensembl
ensGene.txt の 13 番目のカラムにある ENSG を gene_id，ensemblToGeneName.txt に記載されている gene symbol を gene_name として使用した．

iv. lincRNA
独自 ID(linc+chr#.pos) を付与した．ただし Exon を共有している transcript 同士は同じ gene_id とした．

(2) 対応表の clean up

i. 転写領域のオーバーラップの確認
knownToRefSeq.txt にば転写領域にオーバーラップが無いペアが存在したため，これを削除した．

ii. 配列相同性の確認
knownToEnsembl.txt にはデフォルトパラメーターの BLAST+[1] 検索でヒットしないペアが存在したため，これを削除した．

(3) RNA 属性の付与 (protein coding 若しくは noncoding RNA)

i. knownGene

a) knownGene.txt に coding sequence (CDS) 情報が無い場合は noncoding RNA に変更した．

b) CDS 情報があっても，knownToRefSeq.txt 及び knownToEnsembl.txt で対応する transcript が全て noncoding であった 563 transcripts を noncoding RNA に変更した．

c) CDS 長が 70 アミノ酸未満であった 226 transcripts の内，knownToRefSeq.txt 及び knownToEnsembl.txt において他の RNA DB と対応が無かった 30 transcripts を noncoding RNA に変更した．

ii. RefSeq
NM で始まる転写物は protein coding，NR で始まる転写物は noncoding RNA．

iii. Ensembl
ensGene.txt に CDS 情報がある場合は protein coding，無い場合は noncoding RNA．

iv. lincRNA
全て noncoding RNA．

1.(c) 重複の削除

RefSeq には，protein coding と noncoding の区別が付いていない transcript が NM と NR で 2 重に登

録されていることがある. そこで knownGene.txt，refGene.txt，ensGene.txt，lincRNAsTranscripts.txt のそれぞれ内において，全く同じ位置に全ての exon がマップされているレコードを削除した. この時,

(1) Protein coding と noncoding RNA の場合は protein coding を残す

(2) Protein coding 同士，noncoding RNA 同士の場合で knownToRefSeq.txt や knownToEnsembl.txt に対応がある場合は他の DB と gene symbol が一致している方を残す

(3) Protein coding 同士で上記で決まらなかった場合，CDS の長い方を残す

のルールを採用した.knownGene.txt に 24 箇所，refGene.txt に 384 箇所，ensGene.txt に 698 箇所，lincRNAsTranscripts.txt に 2 箇所のゲノム位置に重複があった.

2.(a) refGene.txt の knownGene.txt へのマージによる knownRefGene.txt の作成

(1) knownGene.txt 及び refGene.txt を染色体とストランド別にソートした. RNA の転写方向に従い，plus ストランドは昇順に，minus ストランドは降順し，この順序で処理を行った.

(2) refGene.txt 中の knownGene.txt と全く同じマッピング情報の transcript を削除した.

(3) knownToRefSeq.txt で対応しているとされている transcript の gene_id をマージした. ただし転写領域にオーバラップが無い場合や RNA 属性が異る場合は除いた. この時 knownGene の gene_id が説明文から作成した ID の場合には RefSeq の gene_id に変更した. また一つの RefSeq が複数の knownGene と対応しているとされており，且つ knownGene の gene_id が異る場合は,

i. Gene symbol が一致する

ii. 転写開始点が 5 base 以内

iii. 最も相同性 (相同性検索により算出) が高い

の順に gene_id にマージした (複数 TSS が一致した場合はその中で iii. を適用).

(4) knownGene.txt の exon と同じ exon を持っているレコードには同じ gene_id を付与した. ただし knownGene の gene_id が説明文から作成した ID の場合には RefSeq の gene_id に変更した. また RNA の属性が異る場合は gene_id の統合から除外した. この時マージされた refGene を新たに exon 共有の探索対象とすることはしなかった. knownGene.txt 中の複数の gene_id とマッチした場合は,

i. Gene symbol が一致する

ii. 転写開始点が 5 base 以内

iii. 異る gene_id で転写開始点が 5 base 以内で一致した場合は，最も共通 exon 長の長い gene_id にマージし，上記以外はマージしない.

とした.

2.(b) ensGene.txt の knownRefGene.txt へのマージによる knownRefEnsGene.txt の作成

(1) knownRefGene.txt 及び ensGene.txt を染色体とストランド別にソートした. RNA の転写方向に従い，plus ストランドは昇順に，minus ストランドは降順し，この順序で処理を行った.

(2) ensGene.txt 中の knownRefGene.txt と全く同じマッピング情報の transcript を削除した.

(3) knownToEnsembl.txt で対応しているとされている transcript の gene_id をマージした. ただし転写領域にオーバラップが無い場合や RNA 属性が異る場合は除いた. この時 knownGene の gene_id が説明文から作成した ID の場合には Ensembl の gene_id に変更した. また一つの Ensembl が複数の knownGene と対応しているとされており，且つ knownGene の gene_id が異る場合は,

i. Gene symbol が一致する

ii. 転写開始点が 5 base 以内

iii. 最も相同性 (相同性検索により算出) が高い

の順に gene_id にマージした (複数 TSS が一致した場合はその中で iii. を適用).

(4) knownRefGene.txt の exon と同じ exon を持っているレコードに同じ gene_id を付与した. ただし knownGene の gene_id が説明文から作成した ID の場合には ensGene の gene_id に変更した. また RNA の属性が異る場合は gene_id の統合から除外した. この時マージされた ensGene を新たに exon

共有の探索対象とすることはしなかった. knownRefGene.txt 中の複数の gene_id とマッチした場合は,

i. Gene symbol が一致する

ii. 転写開始点が 5 base 以内

iii. 異る gene_id で転写開始点が 5 base 以内で一致した場合は, 最も共通 exon 長の長い gene_id にマージした. 最長共通 exon 長の候補が複数残った 8 ケースについては, 最も短い transcript を選んだが, この中にやや主観的であるが機械的な gene symbol で無い候補があった場合にはこれを優先した.

とした.

2.(c) lincRNAsTranscripts.txt (lRT) の knownRefEnsGene.txt へのマージによる knownRefEnsLincGene.txt の作成

(1) knownRefEnsGene.txt 及び lRT を染色体とストランド別にソートした.RNA の転写方向に従い, plus ストランドは昇順に, minus ストランドは降順し, この順序で処理を行った.

(2) lRT 中の knownRefEnsGene.txt と全く同じマッピング情報のレコードを削除した.

(3) knownRefEnsGene.txt の exon と同じ exon を持っている noncoding 属性の transcript に同じ gene_id を付与した.knownRefEnsGene.txt 中の複数の gene_id とマッチした場合は,

i. 転写開始点が 5 base 以内

ii. 異る gene_id で転写開始点が 5 base 以内で一致した場合は, 最も共通 exon 長の長い gene_id にマージし, 上記以外はマージしない.

とした.

3.(a) RNA 配列が同一となるレコードの削除

Protein coding, アセンブルドゲノム上のレコードを優先し, 更に knownGene.txt で独自 ID を与えた transcript は可能な限り選択しない様にした上で knownGene.txt, refGene.txt, ensGene.txt, lincRNAsTranscripts.txt 順に優先して残した. 加えて CDS 長が長い順, 染色体番号が小さい順, プラスストランド優先, 転写領域の染色体位置が小さい順 (ストランドを考慮せず) とした.

3.(b) Protein coding の UTR の両端の違いによるマージ

Protein coding RNA 内で最初と最後の exon 以外が全て共通で且つ CDS が同じ場合に, どのペアにおいても 5' UTR 及び 3' UTR が 20 塩基以内で一致するクラスターが生じた場合 (図 1 上図参照), 最も長い transcript だけを残した. この時 2 つのクラスターのどちらにも入ってしまう transcript があった 2 ケース (IFT20 の uc002hav.1, NM_001267774, ENST00000585089 のトリオと SLC12A9 の uc003uwp.3, NM_020246, ENST00000354161 のトリオ) については, 2 つのクラスターを統合して最も長い transcript だけを残した.

3.(c) Noncoding の両端の違いによるマージ

Noncoding RNA 内で最初と最後の exon 以外が全て共通の場合に, どのペアにおいても 5' 末端と 3' 末端の違いの和が 5 塩基以内で一致するクラスターが生じた場合 (図 1 中央図参照), 最も長い transcript だけを残した. この時 2 つのクラスターのどちらにも入ってしまう transcript があった 2 ケース (uc002ads.3, ENST00000562300, TCONS_l2_00009507 のトリオと uc002fco.2, ENST00000555721, TCONS_l2_00010422 のトリオ) については, 2 つのクラスターを統合して最も長い transcript だけを残した.

3.(d) gene_id の確認

gene_id の転写領域のオーバーラップの有無を調べ, オーバーラップが無い場合には gene_id にリビジョンを追加し, 異る gene_id にした.

3.(e) tss_id の付与

knownRefEnsLincGene.txt において, どのペアにおいても transcript start が 5 塩基以内で一致するクラスター (図 1 下図参照) に同じ tss_id を付与した. この時 2 つのクラスターのどちらにも入ってしまう transcript がある場合には, 2 つのクラスターを統合して同じ tss_id を付与した. この結果転写開始点が最大 15 塩基離れている同一 tss_id が 3 箇所 (TMEM234, MPDU1, TMEM199 と CTB-96E2.3) で発生した.

3.(f) p_id の付与

CDS からアミノ酸配列 (stop コドンも含む) に直し, 同じ (stop コドンを含む) アミノ酸配列には異る gene_id でも同じ p_id を付与した. ここで refGene.txt 及び

ensGene.txt には CDS 長が 3 の倍数でない transcripts がそれぞれ 162，16,315 transcripts 存在していたため，stop コドンが最も少なくなるフレームを選択した．

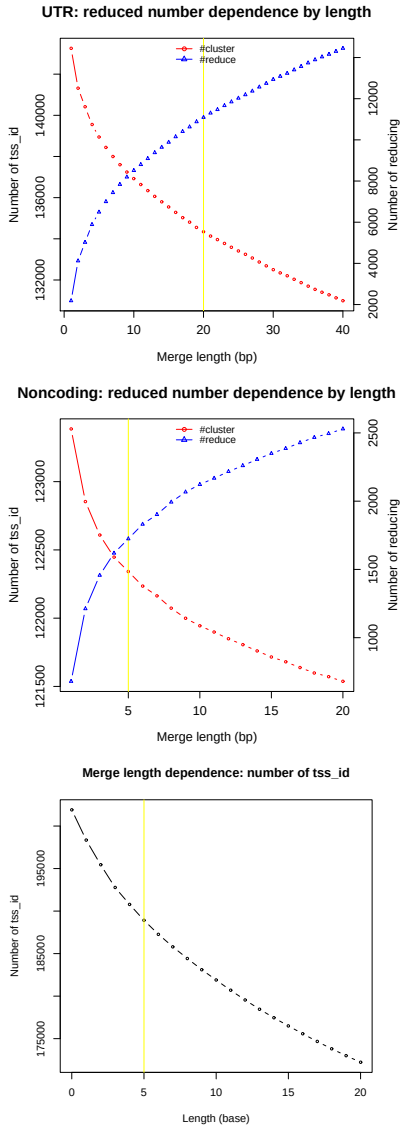

図 1　上図:クラスタリングする UTR 長の差 (横軸) と作成されるクラスター数 (左軸)，統合される転写物数 (右軸) の分布. 中央図:クラスタリングする両末端長の和 (横軸) と作成されるクラスター数 (左軸)，統合される転写物数 (右軸) の分布. 下図:統合する 5' 長と tss_id の数の分布.

4　統合結果

4.1　データ数

統合作業における transcripts 数の変遷を表 3 及び図 2 に，gene_id 数の変遷を表 4 に示した．表 4 の「3.(a) 同一 RNA 配列の削除」において RefSeq 由来の gene_id 数が増えているのは，Known Genes の transcripts が削除された結果，RefSeq 由来の gene_id のみが残ることがあるためである. 同様のことは Ensembl 由来 gene_id でも起っており，実際に最終的な 256,684 転写物に ENSG と付く gene_id は 49,438 ではなく 49,367 と 71 少なくなっている．Known Genes の構築では RefSeq データを参照しているため RefSeq 由来のデータ数は余り残らなくなる．ただし RefSeq は 2 週間に一度更新されるため，統合に使った Known Genes で参照した RefSeq とは異るバージョンの RefSeq データをマージしており，ある程度の数は残っている.

表 3 及び表 4 から，Ensembl には Known Genes 及び RefSeq と異る遺伝子，転写物が多く含まれていることが分かる. ただし Known Genes 及び RefSeq では gene symbol を gene_id としたが，Ensembl では ENSG 番号を使っており，対応表や exon 共有等の処理でしか減らないことに注意されたい (ただし ensemblToGeneName.txt に含まれる gene symbol は 50,770 種類であり，gene_id 数と大きくは変わらない).

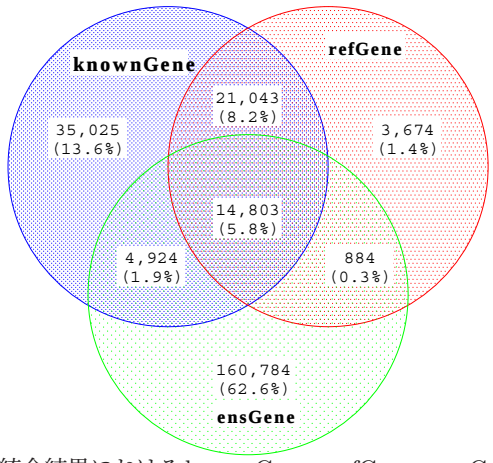

図 2　統合結果における knownGene，refGene，ensGene 由来の transcript の重なり

最終的な 89,981 遺伝子，256,684 転写物における gene_id，tss_id，p_id 毎の転写物数の分布を図 3 に示した. これらは Cufflinks [9] を用いた解析で使用される. また各種 ID 毎に転写物の数が多かった遺伝子を表 5 に示した.

4.2　Mapping への影響

Ensembl GRCh37.60 と今回統合した RNA 参照配列における BWA[8] による 63 サンプルのマッピング割合を図 4 に示した. マッピング割合は，UR/UR が 67.4±3.58 から 70.9 ± 3.58 になり，UR/M が 2.80 ± 0.954 から 1.23 ± 0.319 になった．Splicing variant や noncoding RNA 等が増えたため UR/UR でマッピングされた割

表 3　GenePred ファイル毎の transcript 数の変遷

Procedure	Known Genes	RefSeq	Ensembl	lincRNA	Total
Initial	80,922	43,801	191,891	21,630	338,244
1.(a) genePred の clean up	77,095	41,576	182,938	21,556	323,165
1.(c) 重複の削除	77,071	40,820	182,237	21,554	321,682
2. GenePred ファイルのマージ	77,071	5,845	172,502	16,296	271,714
3.(a) 同一 RNA 配列の削除	75,803	5,617	171,818	16,264	269,502
3.(b) Coding UTR による削除	75,796	4,622	161,726	16,264	258,408
3.(c) Noncoding 両端による削除	75,795	4,558	160,784	15,547	256,684
最終産物における割合	0.295	0.0178	0.626	0.0606	

表 4　GenePred ファイル毎の由来 gene_id 数 (遺伝子数) の変遷

Procedure	Known Genes	RefSeq	Ensembl	lincRNA	Total
Initial	27,994	23,681	57,605	15,242	-
1.(a) genePred の clean up	27,973	23,679	55,202	15,168	-
1.(c) 重複の削除	27,973	23,530	55,176	15,168	-
2.(2) 同一レコードの削除	27,973	580	51,543	11,622	91,718
2.(3) 対応表によるマージ	27,974	550	47,539	11,622	87,685
2.(4) Exon 共有によるマージ	27,974	464	43,589	9,956	81,983
3.(a) 同一 RNA 配列の削除	27,887	465	43,242	9,943	81,537
3.(b) Coding UTR による削除	27,887	464	43,235	9,943	81,529
3.(c) Noncoding 両端による削除	27,887	461	43,208	9,936	81,492
3.(d) gene_id の確認	30,100	492	49,438	9,951	89,981
最終産物における割合	0.335	0.00547	0.549	0.111	

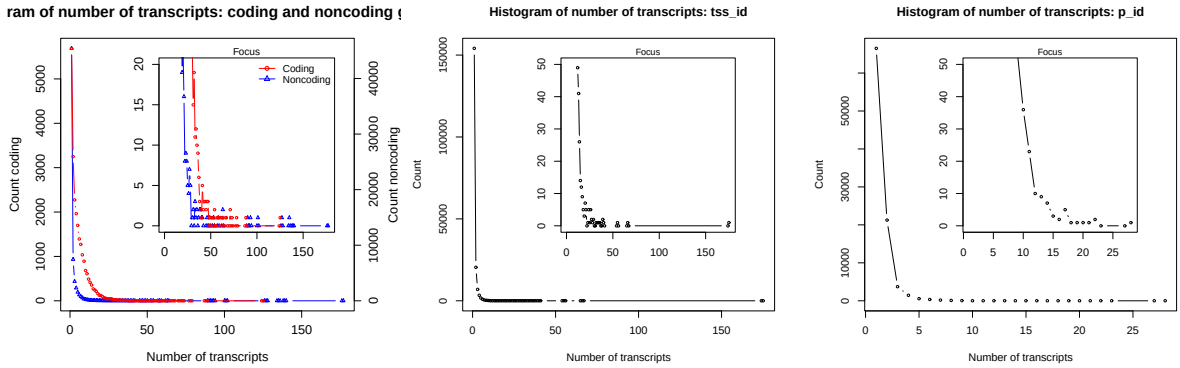

図 3　各種 ID 毎の transcripts 数の分布. 左図:gene_id 毎 (左軸:protein coding，右軸:noncoding)，中央図:tss_id 毎，右図:p_id 毎.

合が増え，soft-clip を伴うことが多い UR/M の割合は減ったと考えている.

謝辞：本稿は，独立行政法人 医薬基盤研究所の先駆的医薬品・医療機器研究発掘支援事業における「多層的疾患オミックス解析における，トランスクリプトーム情報に基づく創薬標的の網羅的探索を目指した研究」において，国立国際医療研究センターで行った「次世代シーケンサーデータ解析パイプラインツールの構築とデータ解析業務」の一部である. また国立がん研究センターの河野隆志分野長には，WTS データのマッピング割合のデータを使用させて頂き，感謝を申し上げる.

表 5 各種 ID 毎に転写物数が多かった上位 11 遺伝子 (gene_id 及び gene name は代表表記)

Coding gene_id			Nonccoding gene_id			tss_id			p_id		
#tran	gene_id	chr	#tran	gene_id	chr	#tran	gene_id	chr	#tran	gene name	chr
125	UTY	Y	177	U6	1	175	UTY	Y	28	TRAC	14
88	CACNA1G	17	139	5S_rRNA	1	66	ATXN3	14	22	MORF4L2	X
78	GPR56	16	135	Metazoa_SRP	1	55	MUC1	1	22	HIST2H4A	1
73	CREM	10	127	7SK	1	40	PCDH15	10	21	ANAPC11	17
71	NDRG2	14	101	TRNA_Pseudo	1	39	IKZF3	17	20	FAM156B	X
71	MUC1	1	93	Mir_548	1	39	DISC1	1	19	MRPL55	1
67	PTPN20A	10	90	Y_RNA	1	36	MYB	6	18	PLAGL1	6
66	CTNND1	11	63	UTY	Y	35	C17orf76-AS1	17	17	ZNF331	19
62	ATXN3	14	63	C17orf76-AS1	17	34	CACNA1G	17	17	CD59	11
61	TCF4	18	58	MEG3	14	33	ATXN3	14	17	BDNF	11
59	SORBS2	4	53	Mir_584	1	32	MEG3	14	17	NREP	5

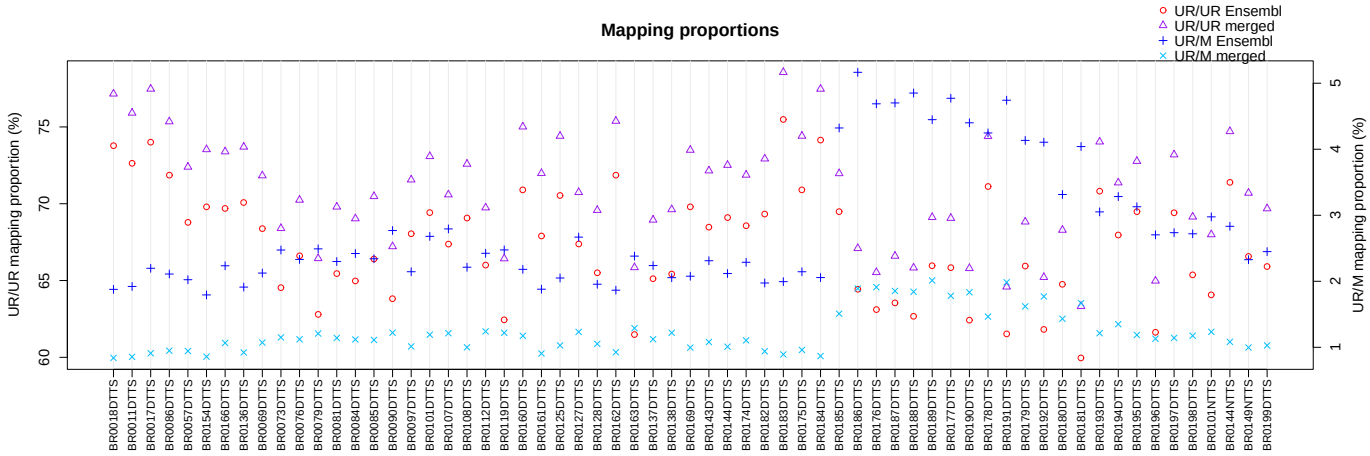

図 4 hg19 の Ensembl GRCh37.60 と統合した cDNA 参照配列における BWA のマッピング割合

引 用 文 献

1) Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., and Madden T. L. Blast+: architecture and applications. *BMC Bioinformatics*, Vol. 10, p. 421, Dec 2009.

2) Moran N. Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and Jhon L. Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes Dev.*, Vol. 25, No. (18), pp. 1915–1927, Sep 2011.

3) The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, Vol. 467, pp. 1061–1073, Oct 2010.

4) NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, Vol. 41, pp. D8–D20, Jan 2013.

5) Mitchell Guttman and *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat. Biotechnol*, Vol. 28, No. (5), pp. 503–510, May 2010.

6) Fan Hsu, W. James Kent, Hiram Clawson, Robert M. Kuhn, Mark Diekhans, and David Haussler. The ucsc known genes. *Bioinformatics*, Vol. 22, No. (9), pp. 1036–1046, May 2006.

7) Seal R. L., Gordon S. M., Lush M. J., Wright M. W., and Bruford E. A. genenames.org: the hgnc resources in 2011. *Nucleic Acids Res.*, Vol. 39, pp. D514–D519, Jan 2011.

8) Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, Vol. 25, No. (14), pp. 1754–60, Jul 2009.

9) Cole Trapnell and *et al.* Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*, Vol. 28, No. (5), pp. 511–515, May 2010.

10) Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, Vol. 25, No. (9), pp. 1105–11, May 2009.

11) Curwen V., Eyras E., Andrews T. D., Clarke L., Mongin E., Searle S. M., and Clamp M. The ensembl automatic gene annotation system. *Genome Res.*, Vol. 14, No. (5), pp. 942–950, May 2004.

12) Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D. The human genome browser at ucsc. *Genome Res.*, Vol. 12, No. 6, pp. 996–1006, Jun 2002.