RNA-seq 解析例

清水顕史

高速シーケンサー(NGS)を用いると、膨大な塩基配列（リードという）を一度に得ることができ、コストパフォーマンスも随分下がってきました。NGS の解析対象はゲノム DNA だけではなく、RNA 由来の二本鎖 cDNA も解読できます。NGS で解読される転写産物の配列は、RNA-seq と呼ぶことがあります。

RNA-seq は、ライブラリー作製の手間をかけずに cDNA クローン毎の EST 配列を大量に取得できるトランスクリプトーム解析技術です。ゲノムや完全長 cDNA などの既知情報(レファレンス）へのアライメント（リード・マッピングとも）を行うことにより転写産物別の遺伝子発現行列を得られるだけでなく、*de novo* アセンブルによって新規の転写産物を見つけられるため、非モデル生物でもトランスクリプトームが解析できます。

ここではサンプルデータによる RNA-seq のレファレンスへのマッピングと可視化を体験してみましょう。実際の NGS リードの解析を進めるめには、数十 G バイトの RAM メモリを搭載した 64 ビット Unix 系 OS が必要です。

1. テストデータのダウンロード

実際の NGS データは、NCBI の SRA からダウンロードできます。これらは１つのファイルが数 G バイト以上にもなる巨大なファイルのため、本講義では、公開されている NGS データ（Mizuno ら 2010 BMC Genomics 11:683 より）を 100 分の 1 以下に削減したイネの RiceRNAseq.fastq(zip ファイルを解凍)を用いることにします（～54K×36bp シングルリード)。

2. FASTQ フォーマット

ショート・リードは FASTQ フォーマットで記述されており、例えば下のように 1 つのリードを 4 行で表わします。まず 1 行目は'@'で始まり配列識別子とオプション記述、(FASTA 形式のタイトル行）で、2 行目は塩基配列の生データ、3 行目は'+'で始まる配列識別子（1 行目と同じ)で、4 行目は 2 行目の塩基配列にそれぞれ対応するクオリティを示す文字です。FASTQ フォーマットではこの 4 行単位で配列情報が繰返し記述されます。

```
リード1 ┐ @ILLUMINA-544730:1:6:1:996:10325#0/1
        │ NACGATTGTATCTCCTCGACTTCTGCTACTGTTGCTGGCGCAATGACTGCT
        │ +ILLUMINA-544730:1:6:1:996:10325#0/1
        ┘ BIJJJTTLHJ`````````U`````````````````````````VVOWWU````T
リード2 ┐ @ILLUMINA-544730:1:6:1:996:6896#0/1
        │ NCTGGATGTAAGATGATGCTTCCCCTCCCTCCAAGGCACCTACTTGGATCG
        │ +ILLUMINA-544730:1:6:1:996:6896#0/1
        ┘ BMMTMTTTQQV`````V`````````````VVV``OL`````````W``V```
   ⋮
```

FASTQファイルのクオリティの統計を診るのに便利なソフトウェアにFastQCがあります。

3 参照配列の準備

RNA-seqは転写産物のせいぜい数百bp程度の短い配列情報で、転写産物完全長(～数Kbp)の一部です。RNA-seqから遺伝子発現行列を得るには、ゲノムや完全長cDNAの既知情報を参照配列としてRNA-seqデータを由来する転写産物別にまとめる作業が必要になります。そのためのレファレンスは、ゲノム・データベースやUniGeneデータベースなどから得ることができます。

テスト用の参照配列データをダウンロードしてみましょう。これはイネの発現遺伝子のうち選抜した344個の配列を、FASTA形式でまとめたものです。

```
>Os12t0274750-00 Hypothetical gene.
TCGCAGCCCGGGGGCTTGTAGGCGATGAAGCTGATGAGCTGCACCTGCCTGACGTTGTCG
…
>Os12t0115100-00 Nonspecific lipid-transfer protein 1 precursor (LTP 1) (PAPI).
ATCCATCCATCATCCATCTCATCATCAGCAACCAATTCGCACCGATCGATCGATCGATCC
…
>Os12t0114900-01 Non-protein coding transcript.
ACAAAGCGTGAGACGACAGCTGCATCTCGGCCGGCGCTGCATGTCATCCTCCGGTAGCGA
…
>Os12t0115300-01 Plant lipid transfer protein and hydrophobic protein, helical domain containing protein.
ATCCATCGCCGGCGACGAGGAGACGCTGACTGAACGAAGCTTACCTAGCTACTCGATCGA
…
>Os12t0188700-02 Similar to Thioredoxin (TRX).
ACAAAAAATTTTCTCCTCCTCTTTCTCTTCTCTCTGTGTCACTGCGTGTGAGGAGCGTAG
…
>Os12t0189400-01 Similar to Photosystem I reaction centre subunit N, chloroplast precursor (PSI- N).
GTCGTCGTCTCCAACCAACCAAATTCTTGCACACGCGCAGCTCGAGCCATGGCCGGAGTG
…
```

4 解析ソフトの準備

Bowtie ダウンロード　(リードをレファレンスにアライメントする)

Samtools ダウンロード　(アライメント結果の整理)

Artemis ダウンロード　(アライメント結果の表示用)

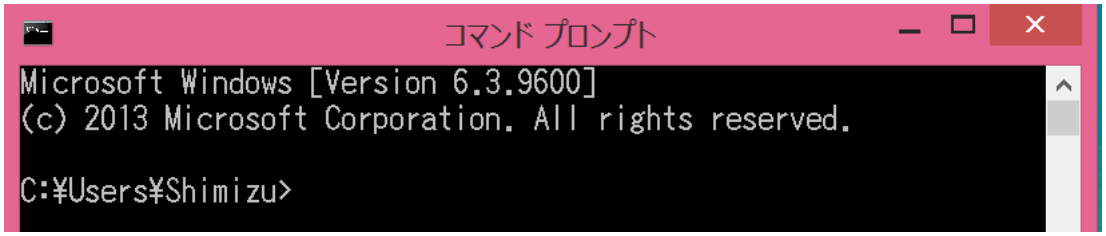
Bowtie と Samtools は圧縮ファイルを解凍します。

5 操作準備

Bowtie と Samtools は、コマンドプロンプトで操作します。その作業の便宜上、解凍してできたフォルダはそれぞれ bowtie および samtools と名前を変更し、C ドライブ直下に作成したRNAseqフォルダに移動しておきます。また、テストデータおよび参照配列データもRNAseqフォルダに移動しておきます。

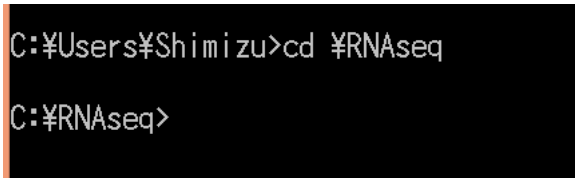
コマンドプロンプトは、コマンド入力によりPCを操作するツールです。スタートメニューなどから起動すると、下図のような黒背景に白文字のウィンドウが開きます。普段のWindowsの操作

と異なり、コマンドプロンプトでは全ての操作をコマンド入力で行います。>の右側の点滅しているアンダーバーは、カーソルといってキーボードから入力した文字情報はこの位置に出力されます。

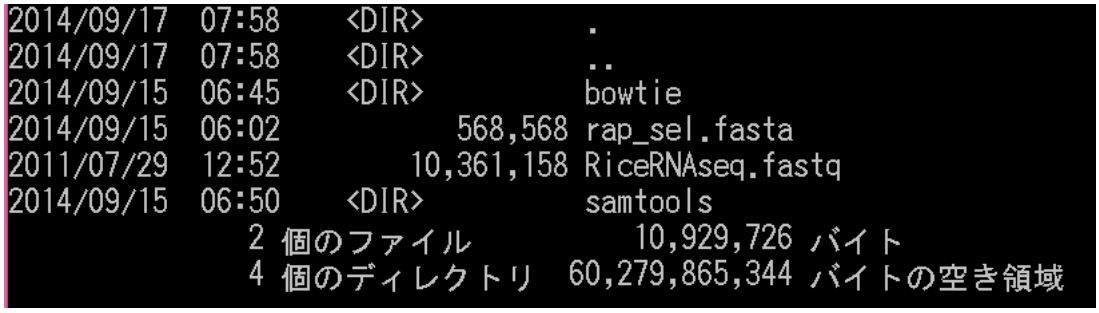


コマンドプロンプトでファイルを指定するとき、その位置情報（Path という）が重要です。上の図で、>の左側の文字列 C:¥Users¥Shimizu は現在のパスを表しており、C:はドライブ名を、¥はルートと呼びパスの区切りを表しており、¥Users¥Shimizu は Users フォルダの Shimizu フォルダにいることを表しています。コマンドプロンプトではフォルダをディレクトリと呼びます。

ファイルは C ドライブの RNAseq フォルダに置いているので、**cd ¥RNAseq** と入力して Enter キーを押して RNAseq フォルダに移動します。**cd** は change directory の意味です。



カーソルの左側の現在位置の表示が RNAseq に変わります。この状態で **dir** と入力して Enter キーを押してみると、このフォルダの中にあるファイルおよびフォルダが表示されます。



予め作成しておいた 2 個のファイルとおよび 2 個のフォルダが確認できます（ちなみに.はカレントディレクトリを、..はその 1 つ上の階層にある親ディレクトリを表します）。

コマンドプロンプトの操作は、Unix の操作に似ています(**dir** コマンドは Unix では **ls** コマンドになるなど、違いもある)。

6 解析手順

6.1 Bowtie によるマッピング

C:¥RNAseq>**cd .¥bowtie**

と入力し、カレントディレクトリ（ここでは C:¥RNAseq）の下にある bowtie フォルダに移動しましておきます。次に、

```
>bowtie-build ¥RNAseq¥rap_sel.fasta rap_sel
```

と入力し、参照配列を bowtie で使用できるようにフォーマットします。すると 6 つのファイル(～.1.ebwt, ～.2.ebwt, ～.3.ebwt, ～.4.ebwt, ～.rev1.ebwt, ～.rev2.ebwt)、が作成できていれば OK です。

例データを参照配列にアライメントするには、次のようにします。

```
>bowtie -S rap_sel ¥RNAseq¥RiceRNAseq.fastq RiceEST.sam
```

```
C:¥RNAseq¥bowtie>bowtie -S rap_sel ¥RNAseq¥RiceRNAseq.fastq RiceEST.sam
# reads processed: 54466
# reads with at least one reported alignment: 9538 (17.51%)
# reads that failed to align: 44928 (82.49%)
Reported 9538 alignments to 1 output stream(s)
```

6.2 Samtools によるファイル変換

アライメントの結果を閲覧するのに Samtools を使います。sam 形式のファイルはヒトが読めるようにまとめられていて便利ですが、ファイルサイズを圧縮した方が扱いやすいので bam 形式に変換しておきます。

一つ上の階層にある samtools フォルダに移動するため、

C:¥RNAseq¥bowtie>**cd ..¥samtools**

と入力します。その後、

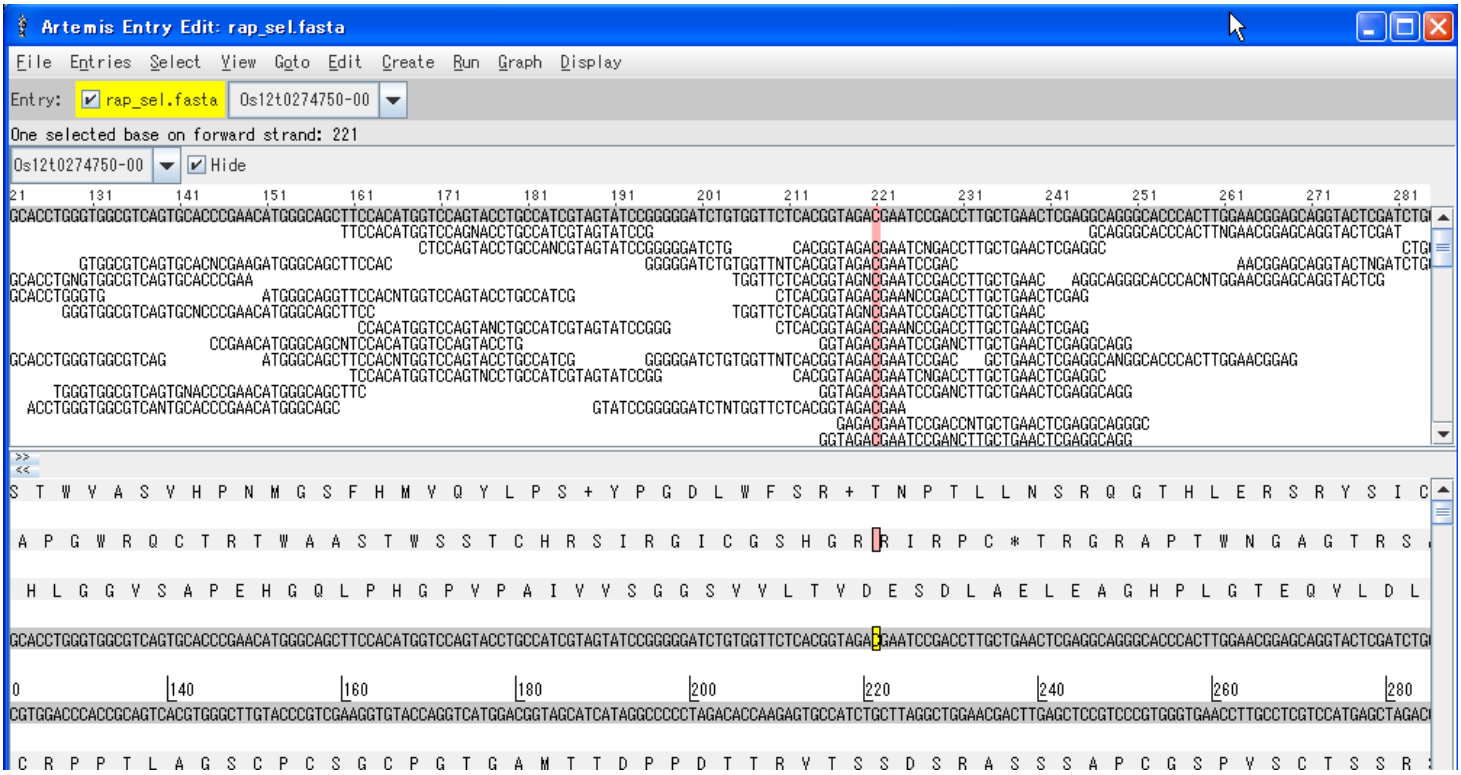
```
>samtools faidx ¥RNAseq¥rap_sel.fasta
>samtools import ¥RNAseq¥rap_sel.fasta.fai ¥RNAseq¥bowtie¥RiceEST.sam RiceEST.bam
>samtools sort RiceEST.bam RiceEST_sorted
>samtools index RiceEST_sorted.bam
```

と入力し、出力される RiceEST_sorted.bam が閲覧可能なアライメント結果になります。

6.3 ソート済み bam ファイルの閲覧

ソート済みの bam ファイルを閲覧できるソフトウェアは複数ありますが、ここでは手軽な JAVA ソフト Artemis を使ってみます。

artemis.jar をクリックしソフトウェアを起動します。File メニューから[Open…]を選び、レファレンス配列（FASTA 形式）を指定します。エントリー編集画面に切り替わりレファレンス配列が表示されたら、File メニューから[Read BAM/VCF…]を選び、ソート済みの BAM ファイルを指定します。



6.4 sam ファイルの中身

Bowtie が出力する sam ファイルを Excel で開いてみると、下の様になります。

まず、A 列の 2 行目から参照配列のリストが表示され（@SQ で始まる行、例データの場合 344 個ある）、

@HD	VN:1.0	SO:unsorted
@SQ	SN:Os12t0274750-00	LN:618
@SQ	SN:Os12t0115100-00	LN:803
@SQ	SN:Os12t0114900-01	LN:718
@SQ	SN:Os12t0115300-01	LN:853
@SQ	SN:Os12t0188700-02	LN:938
@SQ	SN:Os12t0189400-01	LN:656
@SQ	SN:Os12t0190000-01	LN:2086
@SQ	SN:Os12t0230100-01	LN:3089
⋮	⋮	⋮

その後の@PG 行以下に、各リード毎のアライメントの結果が出力されています。例として用いたリードは、シングル・リードなので、結果は以下のようになります。

リード名	0…マップされた 4…マップされなかった 16… 逆向き相補鎖にマップ	マップされた参照配列名	位置	マップスコア max 255	マッチング状況 36Mは36bpのリードが完全一致				リードの配列	リードのスコア			
DRR000349.28565	16	Os01t0978100-01	1333	255	36M	*	0	0	GTGGAATTGGAACAGNTGGAACAATATCTGGTGCTG	B;BBA?0=>B@=>B=!?BBB@BB;@@38;@BCBAAB	XA:i:1	MD:Z:15G20	NM:i:1
DRR000349.28566 HWI-EAS421_90325:1:1:340:123 length=36	4	*	0	0	*	*	0	0	ACAAACTCCACTTTTCTCGGNAGTAAGAAGGCACAA	BBBBCBBCB@BBBBBBCB?=!<B@BA?A>>BBCBB?	XM:i:0		
DRR000349.28567 HWI-EAS421_90325:1:1:340:1095 length=36	4	*	0	0	*	*	0	0	CGCTGAAGCTTGCGAAGCCCNTGTTCCACAAGGCAA	*7>5:83?##########	XM:i:0		
DRR000349.28568	0	Os07t0148900-01	535	255	36M	*	0	0	CGATGTAATGATCATGTAAANTGTGCGTGGAATCAG	BABCCACCBBACCCB?@BA>!?AAB9@A7@BBBB?	XA:i:1	MD:Z:20T15	NM:i:1
DRR000349.28569 HWI-EAS421_90325:1:1:340:691 length=36	4	*	0	0	*	*	0	0	CCGCCTAGCGTTGAACGACGNCCCCAGCACCTCCGG	##########	XM:i:0		
DRR000349.28570 HWI-EAS421_90325:1:1:340:1859 length=36	4	*	0	0	*	*	0	0	GCCGCATCGCCGCCGCCGCCNCCTACGCGCGTAACG	97-BB@9@=45>BAB@<!<=>4(5/83#####	XM:i:0		
DRR000349.28571 HWI-EAS421_90325:1:1:340:814 length=36	4	*	0	0	*	*	0	0	CAAGAATCTGACTCGGGCAGNATTTCACGTACAGGC	BAB@CCB6C+.?BA6<@.<!=56@1;>*7C?4A<	XM:i:0		
DRR000349.28572 HWI-EAS421_90325:1:1:340:773 length=36	4	*	0	0	*	*	0	0	GGAGAAAAAGGTGCCTGACGNTCATATGGAGAAGGC	)@3*<)0>17</>7)@<7(<!=8@@;B#########	XM:i:0		
DRR000349.28573 HWI-EAS421_90325:1:1:340:528 length=36	4	*	0	0	*	*	0	0	CTACAGCAGGCCAGCGGCGTNGAGCAGAATGCTCCA	@.552*;7=:;########	XM:i:0		
DRR000349.28574 HWI-EAS421_90325:1:1:340:687 length=36	4	*	0	0	*	*	0	0	GAACAATTGAACACTTACTTNAAGGCCACAACATAA	B>=@3-<(()>6(=’<=-*A>!=B<73(+@(@6=07#	XM:i:0		
DRR000349.28575	16	Os12t0291100-01	593	255	36M	*	0	0	TAAGCTAAGATCAAGNATCGCGCTGGTGGATGGCTG	@@BB889@A@B6(=!39;A=;5@?B@?A;75@BCB	XA:i:2	MD:Z:15C15T4	NM:i:2
DRR000349.28576 HWI-EAS421_90325:1:1:340:1812 length=36	4	*	0	0	*	*	0	0	GGGCGGCCCATCCGAGGCGGNCCGCGCGACAACAAG	<(:@;@BB;7<(.;>@<!+5:9=>)}44.:57#	XM:i:0		
DRR000349.28577	0	Os08t0126300-01	402	255	36M	*	0	0	GGCTGCTGCTCACTTCAAGGNTGGTGCCAAGGAGGT	>7@##########	XA:i:2	MD:Z:15G4G10A	NM:i:3
DRR000349.28578 HWI-EAS421_90325:1:1:340:193 length=36	4	*	0	0	*	*	0	0	GTTTTGACCTAGGGCCAAAANTGTCAAGTTCGGCAA	BBBBBBBBBAAA)=BBB?@?!?A<BBB;=@C7;A@>	XM:i:0		
DRR000349.28579 HWI-EAS421_90325:1:1:340:342 length=36	4	*	0	0	*	*	0	0	GGAATGGTGCACTTTTATGTNACAAGAGCAGGTAGA	*<##########	XM:i:0		
DRR000349.28580	0	Os12t0420400-01	80	255	36M	*	0	0	CATCTAGCTAGGGAGAGAAANGAAGAGACGAGAGGA	=8;(;’(6(?@@@>B@9*?<!(;;)6’7?8@C>969	XA:i:1	MD:Z:20G15	NM:i:1
DRR000349.28581	0	Os03t0323200-02	2290	255	36M	*	0	0	CGGGCACCTTCCTCATACCTNGCAGATACAACCACA	>;<9A@@@AC9@;’=@;’5?!=@ABC>007?=>@BB	XA:i:2	MD:Z:18A1A15	NM:i:2
DRR000349.28582 HWI-EAS421_90325:1:1:340:1751 length=36	4	*	0	0	*	*	0	0	GGTGACTCCATCTGGATGTTTNTAACCAGCAAGTGTC	?~’3@CBA’8=A########	XM:i:0		
DRR000349.28583 HWI-EAS421_90325:1:1:340:1171 length=36	4	*	0	0	*	*	0	0	CGGAGGCGCTGGAGGCGTGCNCCGTGCAGTGGGGCA	B3@B=>@BBB9A>;AA=@!5>@>?=5=9A9A<AB	XM:i:0		
DRR000349.28584 HWI-EAS421_90325:1:1:340:12 length=36	4	*	0	0	*	*	0	0	ANNNNNNNNNNNNNNNNNNNNNNNANNNNNNNNNNNNNNNN	#!!!!!!!!!!!!!!!!!!!!#!!!!!!!!!!!!	XM:i:0		
DRR000349.28585 HWI-EAS421_90325:1:1:340:711 length=36	4	*	0	0	*	*	0	0	GCTGTACAGGACTCAAACCTNGCCTTTGAACAAAAC	ABCB<BBBCAC@BCAC6BC!8BBCCCCC>CC==>B	XM:i:0		
DRR000349.28586	16	Os12t0292400-01	530	255	36M	*	0	0	TGCAGTTAATCAGCTNCATCGCGTATAAGCCACCGG	AB7A63@9;?=?@>!<B>3<=A=BCB@66<?=@A	XA:i:1	MD:Z:15T20	NM:i:1
DRR000349.28587 HWI-EAS421_90325:1:1:340:332 length=36	4	*	0	0	*	*	0	0	ATTTTTGGGACCTAATTTGANGGTAATTTATTCTTT	BB=CCBBB=ABABA???@A?!4??>8?BB@BABABB	XM:i:0		
DRR000349.28588 HWI-EAS421_90325:1:1:340:837 length=36	4	*	0	0	*	*	0	0	GAAAAAGGCTGCTTATATGCNTGCCTGTTTACACTG	BAA@?BA>BA>@BBBAB??@A?!4??>8?BB=@8A@9AB@	XM:i:0		
DRR000349.28589 HWI-EAS421_90325:1:1:340:1106 length=36	4	*	0	0	*	*	0	0	ATGGAGGGCGGTAAGAAGCANGCTGGTCTCCGAGGG	##########	XM:i:0		
DRR000349.28590	16	Os12t0291733-00	211	255	36M	*	0	0	CGCCTGCCTGACGTTNTCGAAGCCGATGATGCGAAC	##7B@=69?A6BA@7!;7@AAB;@BAB3A=@@B@@	XA:i:1	MD:Z:1A13G20	NM:i:2

課題 6-1　例データから得られた SAM ファイル(RiceEST.sam)から、各参照配列にマップされたリード数を表にまとめなさい。マッチング状況が 36M になったリードをマップされたリードとみなします。

マップされたリード数から、参照配列の発現量を見積もる場合には、参照配列の長さ当たりに換算する方がよいはずです（長い遺伝子の方が、発現している量が低くても断片配列が読まれ易くなるため)。例データのような、シングル・リードをマッピングして遺伝子発現量を見積もる指標としては RPKM(read per kilobase / million)があります。これはマップされた参照配列の長さ当たりのリード数を、100 万リード当たりに換算したものです。

課題 6-2　課題 6-1 から得られた情報より、参照配列 1bp 当たりのリード数が最も高い、つまり最も発現している遺伝子を同定しなさい。