基礎生物学研究所
ゲノムインフォマティクス・トレーニングコース2016年3月

RNA-seq 入門

～概論～

March 10-11, 2016 @ NIBB (Okazaki)

重信秀治 / Shuji Shigenobu

- サポートWiki
  https://github.com/nibb-gitc/gitc2016mar-rnaseq/wiki
- **shige@nibb.ac.jp**

RNA-seq

RNA-seq is a revolutionary tool for *transcriptomics* using deep-sequencing technologies.

(Wang 2009 with modifications)

RNA-seq is unraveling complexities of eukaryotic transcriptomes in model and non-model organisms

- Gene expression analysis
- Novel gene discovery (model org.)
  - Coding and non-coding genes
- Gene cataloguing (non-model org.)
- Anti-sense transcripts
- RNA editing
- Novel splicing variants & fusion genes
- Allele-specific expression

Beyond transcriptome

- DB for proteome analysis
- SNP finding
- *and more …*

Two major goals of RNA-seq

- Gene cataloguing
- Gene expression analysis

RNA-seq analysis pipeline for DE

Differential Expression analysis

Millions of short reads

↓ **pre-processing**
mapping

Reads aligned to reference

↓ **count by unit**
(gene, transcript, exon)

Table of counts

↓ **normalization**
DE testing

List of DE gene

↓ **systems biology**
- GO enrichment
- multivariate analysis
- network analysis

Biological insights

解析ツールの現状: RNA-seq

- 全てのプロセスをこなせる万能ツールはない。
- それぞれのステップに特化したツール群が次々に登場している。

基本戦略

- 各ステップに最適なツールをチョイス、組み合わせた、解析パイプラインの構築。

Pipeline

- 本コースで学ぶオススメの2つのパイプライン
  - Genome-based: TopHat/Cufflinks
  - Transcriptome-based: Trinity/Bowtie/eXpress/edgeR

基生研ゲノムインフォマティックス・トレーニングコース　2016春
RNA-Seq入門
2016.03.10-2016.03.11

NGS基本データフォーマット

基礎生物学研究所
生物機能解析センター
山口勝司

概要

序論

- データフォーマットとは？
- フォーマットを学ぶ理由
- 効率よい学習のポイント

NGS基本データフォーマット

- FASTA　FASTQ　SRA
- BED　GFF/GTF　WIG
- SAM/BAM

概要

序論

- データフォーマットとは？
- フォーマットを学ぶ理由
- 効率よい学習のポイント

NGS基本データフォーマット

- FASTA　FASTQ　SRA
- BED　GFF/GTF　WIG
- SAM/BAM

データフォーマットとは？

データを記録するルール
ルールがあれば情報を効率良く正確に共有できる

例：Webページ　→　HTMLフォーマット
を使用することで
ハード(PC/スマートフォン)
OS (Windows/Mac)
ソフト(IE/Chrome/Safari)
が違っても、どんな環境でも同じページを閲覧可能

次世代シーケンサー解析では
様々なフォーマットが使われる
これらの把握が解析に必須

フォーマットを学ぶ理由

NGS解析の基礎知識だから

研究者間のコミュニケーションや解析方法の理解に必須

例1）　同僚A　：A遺伝子の塩基配列データ見せて
　　あなた　：了解です。fastaで送りますね

← fasta形式が塩基配列情報を含むことを理解していれば、やりとりがスムーズ

例2）　マニュアル　：このソフトはfastaからtree/phylipファイルを生成します
　　あなた　：系統解析をするソフトなんだな

← 入力と出力の形式から行った解析がわかる

研究目的にあわせた解析に必要だから

フォーマットを知ると、そこから自力で必要な情報を獲得できる
これにより、独自性の高い研究が可能になります

例3）　1，巨大なfastaファイルから配列名だけ取り出したい
　　2，fasta形式では、配列名の頭に常に”>”がつく
　　3，”>”がある行だけ集めれば、配列名のリストができる！
　　（エクセルの”並べ変え”機能でできそうだ！）

← 専用のプログラムがなくても自分がほしい結果を得られる

効率良い学習のポイント

Wet 研究者がつまずく点

1：たくさん形式があって区別がつかない！

- 実態はなじみ深い生物学的情報です
- 各フォーマットが含む生物学的情報や解析で使われる場面に注目しましょう

2：意味不明な文字がでてくる！

- ＄や#など“意味不明文字”が頻出しますが、実は重要な情報が含まれています
- 「ヒトとコンピュータ、両方に扱いやすい表記」を考えた開発者の努力の結晶です
- 使い方を理解すれば強力な武器になります。がんばって理解しましょう

以上を踏まえて、各フォーマットを見ていきましょう

概要

序論

- データフォーマットとは？
- フォーマットを学ぶ理由
- 効率よい学習のポイント

NGS基本データフォーマット

- FASTA　FASTQ　SRA
- BED　GFF/GTF　WIG
- SAM/BAM

NGS基本データフォーマット

数十以上のフォーマットがあります
頻出フォーマットだけを紹介します

- 配列用

  FASTA, FASTQ, SRA

- アノテーション用

  BED, GFF/GTF, WIG

- マッピング(アライメント)用

  SAM/BAM

FASTA

概要	配列情報の標準フォーマット
内容	塩基配列　アミノ酸配列
例	公共DBからの配列情報ダウンロード

○規則

“>”で始まる行がタイトル行、改行後に配列
タイトル行は改行不可　配列中では改行可能

○ファイル例

```
>gi|31342400 Bos taurus crystallin, gamma S (CRYGS), mRNA  ←タイトル行
TGCACCAAACATGTCTAAAGCTGGAACCAAAATTACTTTCTTTGAAGACAAAA
ACTTTCAAGGCCGCCACTATGACAGCGATTGCGACTGTGCAGATTTCCACATG
TACCTGAGCCGCTGCAACTCCATCAGAGTGGAAGGAGGCACCTGGGCTGTGTA
TGAAAGGCCCA
>gi|31342400 Bos taurus crystallin, gamma S (CRYGS), mRNA
TGCACCAAACATGTCTAAAGCTGGAACCAAAATTACTTTCTTTGAAGACAAAA
ACTTTCAAGGCCGCCACTATGACAGCGATTGCGACTGTGCAGATTTCCACATG
TACCTGAGCCGCTGCAACTCCATCAGAGTGGAAGGAGGCACCTGGGCTGTGTA
TGAAAGGCCCA
```

FASTQ

概要	NGS結果データの実質的な標準形式
内容	塩基配列、一塩基ごとの品質情報 (Quality value)
例	マッピング、アセンブル での入力データ形式

○規則

1行目 ：“@” の後にタイトル（配列IDや説明）
2行目 ：塩基配列
3行目 ：“+” の後にタイトル（省略可）
4行目 ：配列のクオリティ
＊配列とクオリティには基本的に改行を入れない

○ファイル例

```
@SEQ_ID ←配列ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT  ←塩基配列
+       ←配列ID（省略）
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65  ←クオリティ
```

実習1−1　lessコマンドでEx1_1.fqの中身を見て、fastq形式を確認しよう

FASTQのポイント

塩基配列の信頼性も示せる

Quality value (Phred quality score)

```
+
!''*((((***+))%%%++)(%%%%).1***-
+*''))**55C
```

ABI キャピラリーシーケンサーで
この部分で表されていた値

QV= −10 $\log_{10}$ p （ p : 間違った 塩基決定である確率）
QV = 30 → p = 0.001
QV = 20 → p = 0.01

数値でなく謎の文字が書かれている！

実際のFASTQデータをみると、

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

謎の文字の正体 → “ASCIIコード”を使ってQVを1文字で表したもの

ASCII: American Standard Code for Information Interchange

コンピュータでは文字を数値で表す
通信のため文字と数値の対応関係を規定（1965年）
0～126の数値に文字を割り当て

A → 65　　Apple →65;112;112;108;101;

FASTQ → ASCIIコードを逆に使って、QV(数値)を文字で表す

65 → A

利点：10進数表記よりもファイルサイズを減らせる
（字数が半分、区切り文字も不要）

塩基：	G	A	T	T	G	G	T	G	A	A	T	T	文字が各塩基のQVを表現
文字：	!	?	@	A	>	=	;	9	7	4	0	,	

QVから文字への変換規則

問題点：ASCIIコードでは0-32はコンピューター用の特殊文字に割り当てられている

ASCIIコード表

数値	文字
0	null文字
1	SOH（ヘッダ開始）
2	STX（テキスト開始）
3	ETX（テキスト終了）
4	EOT（転送終了）
.....	.....
30	RS（レコード区切り）
31	US（ユニット区切り）
32	（スペース）
33	!
34	"

・NGSでは10−30を頻用

p = 0.001 → QV=30

・妥協案として特定値を加算してから文字に変換

Phred(QV)値＋X ＝ ASCII値とする

・X値は現在 X=33 でほぼ統一

例）QV 30を表す場合

30 + 33 = 63

→ ASCIIコードで63に該当する文字を当てる（“?”が該当）

・変換にはコード表と簡単な計算が必要

実習1−2 Ex1_2.fqのQV値を求め、すべての配列のp値(エラー確率)が 0.01以下となるように3’側をトリミングしよう

Ex1_2.fq

```
@SEQ_ID
GATTGGTGAATT
+
??@A>=;9740,
```

QV値＋33 ＝ ASCII値

ASCIIコード表

文字	10進	16進	文字	10進	16進	文字	10進	16進	文字	10進	16進	文字	10進	16進	文字	10進	16進	文字	10進	16進	文字	10進	16進
NUL	0	00	DLE	16	10	SP	32	20	0	48	30	@	64	40	P	80	50	`	96	60	p	112	70
SOH	1	01	DC1	17	11	!	33	21	1	49	31	A	65	41	Q	81	51	a	97	61	q	113	71
STX	2	02	DC2	18	12	"	34	22	2	50	32	B	66	42	R	82	52	b	98	62	r	114	72
ETX	3	03	DC3	19	13	#	35	23	3	51	33	C	67	43	S	83	53	c	99	63	s	115	73
EOT	4	04	DC4	20	14	$	36	24	4	52	34	D	68	44	T	84	54	d	100	64	t	116	74
ENQ	5	05	NAK	21	15	%	37	25	5	53	35	E	69	45	U	85	55	e	101	65	u	117	75
ACK	6	06	SYN	22	16	&	38	26	6	54	36	F	70	46	V	86	56	f	102	66	v	118	76
BEL	7	07	ETB	23	17	'	39	27	7	55	37	G	71	47	W	87	57	g	103	67	w	119	77
BS	8	08	CAN	24	18	(	40	28	8	56	38	H	72	48	X	88	58	h	104	68	x	120	78
HT	9	09	EM	25	19	)	41	29	9	57	39	I	73	49	Y	89	59	i	105	69	y	121	79
LF*	10	0a	SUB	26	1a	*	42	2a	:	58	3a	J	74	4a	Z	90	5a	j	106	6a	z	122	7a
VT	11	0b	ESC	27	1b	+	43	2b	;	59	3b	K	75	4b	[	91	5b	k	107	6b	{	123	7b
FF*	12	0c	FS	28	1c	,	44	2c	<	60	3c	L	76	4c	\¥	92	5c	l	108	6c	|	124	7c
CR	13	0d	GS	29	1d	-	45	2d	=	61	3d	M	77	4d	]	93	5d	m	109	6d	}	125	7d
SO	14	0e	RS	30	1e	.	46	2e	>	62	3e	N	78	4e	^	94	5e	n	110	6e	~	126	7e
SI	15	0f	US	31	1f	/	47	2f	?	63	3f	O	79	4f	_	95	5f	o	111	6f	DEL	127	7f

http://e-words.jp/p/r-ascii.html

* LFはNL、FFはNPと呼ばれることもある。
* 赤字は制御文字、SPは空白文字(スペース)、黒字と緑字は図形文字。
* 緑字はISO 646で割り当ての変更が認められており、例えば日本ではバックスラッシュが円記号になっている

解説

```
@SEQ_ID
GATTGGTGAATT
+
??@A>=;9740,
```

①p値が0.01の時のQV値を求める

$$QV = -10 \log_{10} p$$
$$= -10 \log_{10} 0.01$$
$$= -10 (-2)$$
$$= 20$$

QV < 20 部分をトリムすればよい

文字	10進	16進	文字	10進	16進	文字	10進	16進
SP	32	20	0	48	30	@	64	40
!	33	21	1	49	31	A	65	41
"	34	22	2	50	32	B	66	42
#	35	23	3	51	33	C	67	43
$	36	24	4	52	34	D	68	44
%	37	25	5	53	35	E	69	45
&	38	26	6	54	36	F	70	46
'	39	27	7	55	37	G	71	47
(	40	28	8	56	38	H	72	48
)	41	29	9	57	39	I	73	49
*	42	2a	:	58	3a	J	74	4a
+	43	2b	;	59	3b	K	75	4b
,	44	2c	<	60	3c	L	76	4c
-	45	2d	=	61	3d	M	77	4d
.	46	2e	>	62	3e	N	78	4e
/	47	2f	?	63	3f	O	79	4f

②各文字をコード表からASCII値になおし、33 を引いてQV値にする

塩基:	G	A	T	T	G	G	T	G	A	A	T	T
文字:	?	?	@	A	>	=	;	9	7	4	0	,
ASCII値:	63;	63;	64;	65;	62;	58;	59;	57;	55,	52;	48;	44;
QV値:	30;	30;	31;	32;	29;	25;	26;	24;	22,	19;	15;	11;

QV値＋33 ＝ ASCII値
ASCII値 ― 33 ＝QV値

fastqファイルを見る上での注意点

1，QV値はあくまでシーケンサーによる推定値　　目安として利用
2，古いSolexa/Illuminaデータでは規格が乱立！！ ←重要

解析ソフト ver. (CASAVA)	～1.3	1.3～1.5	1.5～1.8	1.8～
参考使用時期	～2009	2009～2010	2010～2012	2012～
QV値算出法	Solexa	Phred	Phred	Phred
X値	64	64	64	33
QV range	-5～40	0～40	3～40 (2=end of read)	0～40

Phred(QV)値＋X ＝ ASCII値

自分のデータがどのバージョン由来か確認し
解析ソフトの設定を補正する必要がある

FASTQのまとめ

概要：塩基配列情報と各塩基の信頼性を表現する

規則：

1行目 ： ”@” 配列名
2行目 ： 塩基配列
3行目 ： ”+”（配列名）
4行目：配列のクオリティ

ポイント：　クオリティは ASCII文字で表現されている

QV値＋33 ＝ ASCII値

fastqの仲間　SRA（Sequence Read Archve）

公共DBへの登録とダウンロードに使用。
バイナリ化（機械語化）された生シーケンスデータ
fastqに変換可能

後ほど詳しく説明

NGS基本データフォーマット

数十以上のフォーマットがあります
頻出フォーマットだけを紹介します

- 配列用

FASTA, FASTQ, SRA

- アノテーション用

BED, GFF/GTF, WIG

- マッピング（アライメント）用

SAM, BAM

BED, GFF/GTF

概要	**ゲノム上の特徴配列を表現する　（アノテーション情報）**
内容	遺伝子名　染色体上の位置　向き　エクソン構造
例	公共DBからアノテーション情報をダウンロード 解析したい領域の指定　アノテーション作業 遺伝子構造予測ソフトの結果出力

＜3形式の違い＞

BED	**ブラウザでの描画情報（色など）を記録可能**
GFF	**拡張性が高く様々な特徴情報を記録可能**
GTF	**GFFの厳格化版　一貫した規則で特徴情報を記録可能**

BED (Browser Extensible Data format)

ブラウザでの描画情報(色など)を記録可能

○規則

項目数 3-12　タブ区切り

省略する場合は何も書かない(タブを2個連続させる)

染色体/Scaffold名	指定領域		領域名	スコア/表記の濃淡	ストランド	太線表示		表示色 赤, 緑, 青 の強度 (0-255)	ブロック(exon等)の情報 コンマ区切りで表記		
	開始位置	終止位置				開始位置	終了位置		個数	サイズ	開始位置
chr22	1000	5000	cloneA	960	+	1000	5000	255,0,0	2	567,488,	0,3512
chr22	2000	6000	cloneB	900	-	2000	6000	0,0,255	2	433,399,	0,3601

1−3項目は必須

4−12項目は省略可

領域開始位置＝0とした位置

実習1−3 Ex1_3.bedはヒトゲノム(GRCh37)の一部をbed形式にしたものである
lessコマンドで開いてbed形式を確認しよう

GFF (General Feature Format / Gene Finding Format)

拡張性が高く様々な特徴情報を記録可能

ゲノムアノテーションの標準的形式

○規則

項目数 5-9 タブ区切り

省略する場合は “-” や“.” を入れる

セミコロンで区切られた タグ値の対

			指定領域					
染色体/Scaffold名	予測ソフト名等	領域の種類	開始位置	終止位置	スコア	ストランド	読枠	属性
chr22	Manual	exon	1001	5000	960	+	0	.
chr22	Manual	exon	2001	6000	900	-	0	NAME “pol1”;

必須　　省略可

属性カラムに様々な情報を追加できる → 拡張性高

GTF (General Transfer Format)

基本的にGFFと同じだが、仕様をより細かく規定

○規則

			指定領域					
染色体/Scaffold名	予測ソフト名等	領域の種類	開始位置	終止位置	スコア	ストランド	読枠	属性
chr22	Twinscan	CDS	380	401	.	+	0	gene_id “001”; transcript_id “001.1”;
chr22	Twinscan	CDS	501	650	.	+	2	gene_id “001”; transcript_id “001.1”;
chr22	Twinscan	CDS	700	707	.	+	2	gene_id “001”; transcript_id “001.1”;
chr22	Twinscan	start_codon	380	382	.	+	0	gene_id “001”; transcript_id “001.1”;
chr22	Twinscan	stop_codon	708	710	.	+	0	gene_id “001”; transcript_id “001.1”;

必須：CDS, start_codon, stop_codon

任意：5UTR, 3UTR, inter, inter CNS, intron_CNS, exon

それ以外は無効

遺伝子と転写産物のIDを表記する

実習1−4　　Ex1_4.gtfは 1_3と同じ領域をgtf形式にしたものである。
lessコマンドで開いてgtf形式を確認しよう

注意　GFF/GTFとBEDでは座標の表現が異なる

GFF/GTF：開始、終了ともに 1-based（1 から始まる）座標

BED ：開始はObased, 終了は 1-based 座標

具体例

GFF/GTF	1	2	3	4	5	6	7	8
	A	G	T	A	C	T	C	G
BED	0	1	2	3	4	5	6	7	8

黄色部分を示す時

GFF/GTF format：開始 3，終了 6 （長さは6-3+1=4）

BED format　　：開始 2，終了 6 （長さは6-2=4）

実習1−5　　Ex1_3.bedとEx1_4.gtfを開き、実際に座標がずれていることを確認しよう

WIG（Wiggle Format）

概要	ゲノム上の量的特徴を表現するための形式
内容	ゲノム上の座標に対する”数値”情報
例	GC含量、発現量などを表す

〇規則　　２形式から選べる

1）VariableStep　柔軟な指定が可能

```
variableStep chrom=chr2
300601        22.5
300701        30.5
300751        28.2
```

位置と値の組で領域を指定するため
間隔は位置ごとに変更可能

2）FixedStep　コンパクトな表現が可能

```
fixedStep  chrom=chr3 start=300601 step=100
22.5
30.5
25.8
```

定開始位置と間隔は先頭行で指定し、後は値のみを示していく

NGS基本データフォーマット

数十以上のフォーマットがあります
頻出フォーマットだけを紹介します

- 配列用

FASTA, FASTQ, SRA

- アノテーション用

BED, GFF/GTF, WIG

- マッピング(アライメント)用

SAM, BAM

SAM (Sequence Algnment/Map format)

概要	マッピング(アライメント)結果を表現
内容	マッピング情報 (位置, インデル, ミスマッチ) ペアフラグメントの状況, 塩基配列
例	SNP、発現量解析への入力データ形式

○ファイル例

ヘッダー部

マッピング結果

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   163   ref  7   30   8M2I4M1D3M   =   37  39   TTAGATAAAGGATACTG     *
r002   0     ref  9   30   3S6M1P1i4M   *   0   0    AAAGATAAGGATAT       *
r003   0     ref  9   30   5S6M         *   0   0    GCCTAAGCTAA          *   SA:Z:ref,29,-,6H5M
r004   0     ref  16  30   6M14N5M      *   0   0    ATAGCTTCAGC          *
r003   2064  ref  29  17   6H5M         *   0   0    TAGGC                *   SA:Z:ref,9,+,5S6M
r001   83    ref  37  30   9M           =   7   -39  CAGCGGCAT            *   NM:i:1
```

実習1−6　　Ex1_7.samを開きsam形式を確認しよう

○規則

ヘッダー部

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

“@”で開始

@HD VN: (バージョン) SO: (ソート状況)

@SQ SN: (リファレンス名) LN: (リファレンスの長さ)

マッピング結果部分

項目間はタブで区切る

フラグメント名	FLAG	リファレンス配列名	アライメント開始位置	マッピングQV	CIGAR	ペアフラグメントの場所 Ref名	開始	長さ	配列	配列QV	オプション
r001	163	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1i4M	*	0	0	AAAGATAAGGATAT	*	
r003	0	ref	9	30	5S6M	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	*	SA:Z:ref,9,
r001	83	ref	37	30	9M	=	7	-39	CAGCGGCAT	*	NM:i:1

ポイント！ “CIGAR” “FLAG”

SAMのポイント1：CIGAR

数字と文字を組み合わせアライメント状況を示す

フラグメント名	FLAG	リファレンス配列名	アライメント開始位置	マッピングQV	CIGAR	ペアフラグメントの場所 Ref名	開始	長さ	配列	配列QV	オプション
r001	163	ref	5	30	3M2D2M	=	37	39	GCAAG	44>>>	

3M2D2M

塩基数

状況

3塩基一致、2個挿入、2塩基一致

```
ref   :ATGCGCATTAGCCTAA
read  :     GCA--AG
```

記号	状況
M	一致
I	挿入
D	欠失
N	イントロン(RNAvsDNAのみ)
S	クリップ(塩基情報残す)
H	クリップ(塩基情報削除)
P	他リードが 挿入を入れている

SAMのポイント2：FLAG　リードの状態を示す数値

理解すると「マップされなかったリードだけ選ぶ」などの操作が可能になる

数値（10進数）	意味
1	ペアリードがある
2	両方適切にマップされている
4	自分がマップされていない
8	ペア相手がマップされていない
16	逆鎖にマップされた（配列も逆鎖で表記）
32	ペア相手は逆鎖にマップされた
64	Read1の配列である
128	Read2の配列である
256	Multiple hitでトップヒットでないアライメント
512	マッピングQVが低い

複数の状況に合致する場合は数値を加算

ペアリード, 両方マップされた　→　1+2＝3
2進数の個々の有無で評価されている
加算した結果が、ほかの状況と一致しないようになっている

Paired end readでFLAG値の組み合わせを見てみる

	Read2の配列である 1	Read1の配列である 1	ペア相手は逆鎖にマップされた 1	逆鎖にマップされた 1	ペア相手がマップされていない 1	自分がマップされていない 1	両方適切にマップされている 1	ペアリードがある 1	2進数表記 11111111	samファイルの記載は10進数表記 255
通常のpaired end seqで consistentにアラインしていれば この4通りになる	0	1	0	1	0	0	1	1	01010011	83
	0	1	1	0	0	0	1	1	01100011	99
	1	0	0	1	0	0	1	1	10010011	147
	1	0	1	0	0	0	1	1	10100011	163
片方しかアラインしてない場合	0	1	0	0	1	0	0	1	01001001	73
	0	1	0	1	1	0	0	1	01011001	89
	0	1	0	0	0	1	0	1	01000101	69
	0	1	1	0	0	1	0	1	01100101	101
	1	0	0	0	1	0	0	1	10001001	137
	1	0	0	1	1	0	0	1	10011001	153
	1	0	0	0	0	1	0	1	10000101	133
	1	0	1	0	0	1	0	1	10100101	165
どっちもアラインしていない場合	0	1	0	0	1	1	0	1	01001101	77
	1	0	0	0	1	1	0	1	10001101	141

自動でflagを計算してくれるサイトがある

http://broadinstitute.github.io/picard/explain-flags.html

This utility explains SAM flags in plain English.
It also allows switching easily from a read to its mate.

Flag: [ ] Explain

Switch to mate

Explanation:
- ☐ read paired
- ☐ read mapped in proper pair
- ☐ read unmapped
- ☐ mate unmapped
- ☐ read reverse strand
- ☐ mate reverse strand
- ☐ first in pair
- ☐ second in pair
- ☐ not primary alignment
- ☐ read fails platform/vendor quality checks
- ☐ read is PCR or optical duplicate
- ☐ supplementary alignment

Summary:

SAMのまとめ

概要：各リードがマップされた場所と状態を表す

規則：ヘッダ部とアライメント部からなる　タブ区切り

ポイント：　FLAG値　→　リードのマップ状況
　　CIGAR値　→　リードのアライメント状況

触れなかった重要点

ペアフラグメント部分の“長さ”列　→　フラグメント間距離 + 両リード長

SAM formatの詳細な仕様書
http://samtools.github.io/hts-specs/SAMv1.pdf

BAM

- BAM

  SAMをバイナリ（機械語）化したもの

  容量が小さくなるが、人には理解できない

  SAMに戻すことも可能なので必要に応じて変換

- BAM indexing file

  BAMファイルに対して作られる検索用ファイル

  高速検索や可視化ソフトなどに必要

  後ほど詳しく説明

フォーマット各論まとめ

	FASTA	FASTQ	SAM
概要	配列情報の標準形式	NGS結果の標準形式	マッピング結果を示す
内容	塩基配列 アミノ酸配列	塩基配列と 一塩基to毎の品質情報	マッピング情報 ペアの状況, 塩基配列
例	公共DBからの配列情報ダウンロード	マッピング、アセンブル解析での入力データ形式	マップ結果の閲覧、集計 SNP、発現量解析への入力
特徴		QV値はASCII文字で表現 SRAから変換可能	CIGAR, FLAG値を利用 バイナリ化したのがBAM

	BED	GFF	GTF	WIG
概要	ゲノム上の特徴配列を表現する			ゲノム上の量的特徴を表現
内容	遺伝子名　染色体上の位置　向き　エクソン構造			ゲノム上の座標に対する ”数値”情報
例	公共DBからアノテーション情報をダウンロード 解析したい領域の指定　アノテーション作業 遺伝子構造予測ソフトの結果出力			GC含量、発現量などを表す
特徴	ブラウザでの描画情報を記録	拡張性高	GFFの厳格化版 一貫した規則	2つの形式 VariableStep/FixedStep

NGS基本ツール:
マッピングツール Bowtie2

基礎生物学研究所
ゲノムインフォマティクストレーニングコース
内山　郁夫 (uchiyama@nibb.ac.jp)

ショートリードのマッピング

ゲノム配列
（リファレンス reference 配列）

FASTAファイル（配列）

```
>chr
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCT
CTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAA
TTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAG
AGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGG
```

サンプル（ゲノムDNA／RNA）
（リード read 配列）

FASTQ ファイル
（配列＋クオリティ値）

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631
ATCCGGCTGGCGCACCGACCTATGTTCCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513
CCCFFDFDFHDFFHIIIEGIHJJJJGFHGGHGGHGGIIJDGIJHHGGGHIH
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530
CAGGACATCGCCTTTGATCGGTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530
CCCFFFDFAFHFHIJGHIJIJJIJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

リファレンス配列へのマッピング

SAM ファイル（マッピング結果）

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr  LN:4639675
@PG     ID:bowtie2      PN:bowtie2      VN:2.2.4        CL:"/bio/bin/bowtie2-alig
SRR1515276.40   0    chr 4423609 42    51M    *    0        0        GGAATTCCTCACTGCCA
SRR1515276.158  16   chr 501700  42    51M    *    0        0        ACGCACCGAGTGCAAAG
SRR1515276.212  4    *   0       0     *      *    0        0        GGCCGCTTTCAGCGTGT
SRR1515276.319  0    chr 2922768 42    51M    *    0        0        GCTTAAGTTGATTAAGG
SRR1515276.367  16   chr 2753873 42    51M    *    0        0        GCGTGTCCGTCCGCAGC
SRR1515276.411  0    chr 3440721 42    51M    *    0        0        ACGGCATAATTTCTTGA
```

ショートリードマッピングのための アライメントツール

- **ハッシュテーブルによるインデックス**
  - リード配列にインデックスづけ
    - MAQ, RMAP　など
  - リファレンス配列にインデックスづけ
    - SSAHA2, MOSAIK, NovoAlign　など
- **Burrows Wheeler 変換に基づくインデックス (FM index)**
  - リファレンス配列にインデックスづけ
    - Bowtie/Bowtie2, BWA, SOAP　など
- **splice-aware aligner**
  - ゲノム上離れた場所へのマッピングに対応
    - TopHat, HISAT, STAR, GSNAP など

（TopHatやHISATはBowtieを内部的に使っている）

インデックスを使った検索
ハッシュテーブル

ゲノム配列

ACACGTTACGGT.....

リード配列

CGTTGCA

①インデックス作成

ハッシュテーブル
各2-merの出現位置を記録

2-mer	positions
AC	1,3,8
CA	2
CG	4,9
GG	10
GT	5,11
TA	7
TT	6

② インデックスを使った
初期検索(seed検索)

CGTTGCA

③ 見つかったseedを
延長してアライメント

ACACGTTACGGT.....
CGTTGCA

k-mer ハッシュテーブル

ゲノム配列

ACACGTTACGGT

CGTTGCA
CGGTATG

リード配列

2-mer	positions
AC	1,3,8
CA	2
CG	4,9
GG	10
GT	5,11
TA	7
TT	6

3-mer	positions
ACA	1
ACG	3,8
CAC	2
CGG	9
CGT	4
GGT	10
GTT	5
TAC	7
TTA	6

kが大きいほどノイズが減る(効率↑)が、取りこぼしの可能性は増える(感度↓)

接尾辞配列(suffix array)

ACACGTTACGGT

接尾辞

1	ACACGTTACGGT
2	CACGTTACGGT
3	ACGTTACGGT
4	CGTTACGGT
5	GTTACGGT
6	TTACGGT
7	TACGGT
8	ACGGT
9	CGGT
10	GGT
11	GT
12	T

ソート

1	ACACGTTACGGT
8	ACGGT
3	ACGTTACGGT
2	CACGTTACGGT
9	CGGT
4	CGTTACGGT
10	GGT
11	GT
5	GTTACGGT
12	T
7	TACGGT
6	TTACGGT

ACG 3,8

CG 4,9

接尾辞
配列

Burrows-Wheeler 変換 (BWT)に基づくインデックス（FM-Index）

ACACGTTACGGT $\underset{\text{逆変換}}{\overset{\text{BWT}}{\rightleftarrows}}$ TTCAAACGCGTG

矢印（LF mapping）を辿って元の配列を再構築できる（逆変換）。

メモリ使用量、計算量とも効率のよい検索の実現

Bowtie コマンド

- Burrows-Wheeler 変換に基づくFM-indexを利用したショートリードのマッピングプログラム
- BowtieとBowtie2がある。後者はギャップを考慮した検索を行うので、感度がより高い。また、検索の方針が単純化されて分かりやすくなるなど、多くの点で改良されている。
- シーケンスのリード長が長い（50bp以上）時はBowtie2の方が一般に検索効率がよく、精度も高い。リード長が短い（50bp未満）時はBowtieの方が検索効率または精度がいい場合もある。

Bowtie2 アルゴリズムの詳細

1. Seed 配列の抽出
各リード配列およびその相補配列から *i* 塩基ごとに *L* 塩基の配列を抽出してseed配列とする（図では*i*=10, *L*=16)。

2. FM index を用いた検索
各seed配列がゲノム上に出現する位置がBW rangeとして得られる。最大1つのミスマッチを考慮した検索が可能。

3. ヒットの優先付け、位置の取得
BW rangeの幅が小さいヒットに高い優先度をつけて、ランダムに候補をピックアップし、ゲノム上の位置を取得。

4. アライメントの計算
得られた位置の周辺で、ギャップ入りのアライメントスコアを計算。これを各候補位置について繰り返して、最高スコアを与えるゲノム上の位置を出力。

インデックスの作成　bowtie2-build

- ゲノム配列に対してBWTに基づくインデックスを作成し、それを使って検索する（元の配列ファイルは使わない）。

- インデックスの作成

```
bowtie2-build 配列ファイル インデックス名
```

  - 配列ファイルはカンマで区切って複数を指定可能
  - 実行すると、インデックスとして、*インデックス名*.*n*.bt2 (*n*=1-4)および、*インデックス名*.rev.*m*.bt2 (*m*=1-2) の、計6つのファイルが作成される。

実習：bowtie2-build

- ゲノムデータ（FASTA形式）
  `eco_o139.fa`　腸管毒素原性大腸菌(ETEC) O139:H28のゲノム配列

- bowtie2用インデックスの作成（インデックス名は etec）

```
$ bowtie2-build eco_o139.fa etec
```

- インデックスから元の配列データを再構築

```
$ bowtie2-inspect etec |less
```

マッピングの実行　bowtie2

- マッピングの実行
  - single-end read の場合

```
bowtie2 -x インデックス名
        -U リードファイル -S 出力ファイル
```

  - paired-end read の場合

```
bowtie2 -x インデックス名
        -1 リードファイル1 -2 リードファイル2 -S 出力ファイル
```

（改行せずに1行で打つこと）

❖ リードファイルはカンマ区切りで複数を指定可能

❖ paired-endの場合、リード1とリード2は別々のファイルに格納されており、それらの中で、対応するリードは同じ順番で出現すること

実習：bowtie2

- リード配列（FASTQ 形式; paired-end）
  `etec_1.fq, etec_2.fq`
- リファレンス配列のインデックス名
  `etec`　（先ほど作ったもの）

- **bowtie2の実行**

```
$ bowtie2 -x etec -1 etec_1.fq -2 etec_2.fq
      -S etec_bowtie2.sam
```

マッピング結果ファイル(SAMファイル)

ヘッダ（@で始まる）

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:ETEC_chr     LN:4979619
@SQ     SN:pETEC_80     LN:79237
@SQ     SN:pETEC_35     LN:34367
@SQ     SN:pETEC_73     LN:70609
@SQ     SN:pETEC_6      LN:6199
@SQ     SN:pETEC_74     LN:74224
@SQ     SN:pETEC_5      LN:5033
@PG     ID:bowtie2      PN:bowtie2      VN:2.2.7        CL:"/bio/bin/bowtie2-align-s --wrapper basic-0 -x etec -1 etec_1.fastq -2 etec_2.fastq"
SRR345261.25    89      ETEC_chr        3758170 1       49M     =       3758170 0       ACACGCGGCATGGCTG...     ###?ED>EBDBDDE,E...     AS:i:-1 XS:i:-1 XN:i:0  XM
SRR345261.25    133     ETEC_chr        3758170 0       *       =       3758170 0       NNNNNNNNNNNNNNNN...     ################...     YT:Z:UP YF:Z:NS
SRR345261.50    73      ETEC_chr        4361458 1       49M     =       4361458 0       CAAGCGTTAATCGGAA...     :HEGDFHHHH@BGG=B...     AS:i:0  XS:i:0  XN:i:0  XM
SRR345261.50    133     ETEC_chr        4361458 0       *       =       4361458 0       NNNNNNNATNNNNNNN...     ################...     YT:Z:UP YF:Z:NS
SRR345261.75    73      ETEC_chr        4362922 1       49M     =       4362922 0       CGGTGGATGCCCTGGC...     DDDDBD6<B>DB>BB>...     AS:i:-2 XS:i:-2 XN:i:0  XM
SRR345261.75    133     ETEC_chr        4362922 0       *       =       4362922 0       NNNNNNTTTNNNTCGG...     ################...     YT:Z:UP YF:Z:NS
SRR345261.100   73      ETEC_chr        679991  42      49M     =       679991  0       GTGGTTTAATGAGTCC...     GGGGGGGGB=ED=EEG...     AS:i:0  XN:i:0  XM:i:0  XO
SRR345261.100   133     ETEC_chr        679991  0       *       =       679991  0       NNNNNNCACCGNTAGT...     ################...     YT:Z:UP YF:Z:NS
SRR345261.125   73      ETEC_chr        4376280 42      49M     =       4376280 0       CTCAGGATGAGGGTCA...     EEEE=B<<@BDEEDE:...     AS:i:0  XN:i:0  XM:i:0  XO
SRR345261.125   133     ETEC_chr        4376280 0       *       =       4376280 0       NNNNNTTTTCCNTTAG...     ################...     YT:Z:UP YF:Z:NS
SRR345261.150   89      ETEC_chr        779844  42      49M     =       779844  0       TTCAGGAAACCCTGAA...     B@8D>ECC?BG@ECC>...     AS:i:-5 XN:i:0  XM:i:1  XO
SRR345261.150   133     ETEC_chr        779844  0       *       =       779844  0       CNCNGGAGTACNTTGA...     ################...     YT:Z:UP YF:Z:NS
SRR345261.175   83      ETEC_chr        3605306 42      49M     =       3605113 -242    CCGCTTGCGCGGGCCA...     EDE<8??;?@DGGDDE...     AS:i:0  XN:i:0  XM:i:0  XO
SRR345261.175   163     ETEC_chr        3605113 42      49M     =       3605306 242     CCGGGTTCTGTCGTGG...     DGGDGFDGGGGGGEGD...     AS:i:-3 XN:i:0  XM:i:3  XO
SRR345261.200   77      *               0       0       *       *       0       0       AAAAAAAAAAAAAAAA...     ################...     YT:Z:UP
SRR345261.200   141     *               0       0       *       *       0       0       AAAAAAAAAAAAAAAA...     8@##############...     YT:Z:UP
SRR345261.225   83      ETEC_chr        2879707 1       49M     =       2879600 -156    CACAACACGAGCTGAC...     8?D8BEBGD@GG8GCE...     AS:i:0  XS:i:0  XN:i:0  XM
SRR345261.225   163     ETEC_chr        2879600 1       49M     =       2879707 156     CCCACCTTCCTCCAGT...     GGGBGDEGGG@GG<G8...     AS:i:-1 XS:i:-1 XN:i:0  XM
SRR345261.250   99      ETEC_chr        4361346 1       49M     =       4361525 228     GTACTTTCAGCGGGGA...     ECE=>EC?FDG<EGDA...     AS:i:0  XS:i:0  XN:i:0  XM
SRR345261.250   147     ETEC_chr        4361525 1       49M     =       4361346 -228    CCGGGCTCAACCTGGG...     #############BC...      AS:i:0  XS:i:0  XN:i:0  XM
```

リファレンス配列に関する情報

FLAG

マップされた染色体と位置（* はマップされなかった）

MAPQ

CIGAR

ペアの相手がマップされた染色体（同じなら=）と位置、フラグメントの長さ（右側のリードは負値）

リード配列

配列クオリティ値

オプション
- AS　アライメントスコア
- XS　他の位置でのベストスコア
- YF　リードがfiltering out された理由

同じ名前のリード＝ペアエンドのリード対

マッピングクオリティ(MAPQ)

- マッピングクオリティ(MAPQ)値は以下の式で計算される。

$$MAPQ = -10\log_{10}(P_e)$$

  ただし、P_eはリードが間違った位置にマップされている確率の推定値。
- MAPQは、リードがその位置にどの程度ユニークにマップされたかを示す指標であり、その位置でのアライメントスコアが、他のすべての位置におけるスコアよりずっと大きいときに大きくなる。
- Bowtie2では同じスコアのアライメントが複数の位置で得られた場合、ランダムに一つの位置を出力し、MAPQに低い値を設定する。
- MAPQが低いアライメントの位置は信用できないので、下流の解析の際には捨てた方が良い場合もある。
- P_eをどのように見積もるかについては決まりはなく、マッピングプログラムごとに異なっている。

ペアエンドリード対の検索オプション

- `-I` *int* フラグメント長の最小値(default: 0)
- `-X` *int* フラグメント長の最大値(default: 500)
- `--fr` / `--rf` / `--ff` リード1とリード2の相対的な向き(default:`fr`)

- 条件を満たさない(discordant)リード対もデフォルトでは出力される。その際、2カラム目(FLAG)の2ビット目(ペアが正しくアラインされたか？)に0がセットされる。

アライメントのモード

- `--end-to-end` リード配列全長に渡るアライメント(default)

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

- `--local` リード配列のうち、類似度の高い一部の領域のみを抜き出してアラインしたもの

```
Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
Reference: TAACTTGCGTTAAATCCGCCTGG
```

検索の精度と速度に関するオプション

- `-N` *int* seed 検索時にミスマッチを許す数(0 or 1)
- `-L` *int* seed の長さ
- `-i` *func* seed をとる間隔(リード長を基に決める式を指定)
- `-D` *int* 最高スコアが更新されないときアライメント計算を打ち切るまでの回数
- `-R` *int* seedが反復配列であるときにre-seedを行う最大回数

PresetOptions:
上記のオプションを同時に設定するpreset optionがある。高速(低感度)→高感度(低速)の順に4段階のオプションが用意されている。

- `end-to-end`モードの場合 (default: `sensitive`)
  `--very-fast` / `--fast` / `--sensitive` / `--very-sensitive`

- `local`モードの場合 (default: `sensitive-local`)
  `--very-fast-local` / `--fast-local` / `--sensitive-local` / `--very-sensitive-local`

その他のオプション

- `-p` *int* 指定した数のCPUコアを使って実行する
- `--phred64` Quality値がPhreadScore+64として定義されている
- `-f` リードがFASTA形式のファイルである
- `--no-unal` アラインできなかったリードは出力しない
- `--no-discordant` ペアの位置関係が正しくないリード対は出力しない

練習問題

1. 2つのリードファイル etec_1.fq, etec_2.fqを、それぞれ single end readのデータと見なして、bowtie2でetec をリファレンスとしてマッピングし、結果をファイルetec_bowtie2_single.sam に出力せよ。その際、リードファイルはカンマ区切りで複数指定できることを使え。出力ファイルの行数を、etec_bowtie2.samと比較せよ。また、それぞれのファイルの先頭20行をheadで出力し、比較せよ。
2. 再びetec_1.fq と etec_2.fqをpaired endとしてetecに対してマッピングするが、その際オプションとして `–I 100 –X 200` を指定しよう。これらのオプションはどういう意味を持っているか。このコマンドを出力ファイルを `etec_bowtie2_X200.sam` として実行せよ。出力ファイルの行数は、etec_bowtie2.samと比べて変化したか。ファイルの内容を以下のコマンドで比較し、どこが変わったか検討せよ。ただし、`diff`は2つのファイルを行ごとに比較して異なる行を出力するコマンドで、`'<'`で始まる行は最初のファイル、`'>'`で始まる行は2番目のファイルのみに出現する行を示す。また、`less` の`-S`オプションは、長い行を折り返さずに表示することを指示する。

```
$ diff etec_bowtie2.sam etec_bowtie2_X200.sam | less -S
```

samtools

SAM/BAMフォーマットの
マッピング結果を扱うコマンド群

Samtools

http://www.htslib.org

1. フォーマット変換
   SAM（テキスト） ⇔ BAM（バイナリ）の変換
2. データのソート, 索引付け
3. データ抽出
   特定のリードの選出
   統計情報収集（発現量解析）

Samtools の起動

```
$ samtools
```

```
Program: samtools (Tools for alignments in the SAM format)
Version: 1.3 (using htslib 1.3)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     index          index alignment

  -- Editing
     calmd          recalculate MD/NM tags and '=' bases
     fixmate        fix mate information
     reheader       replace BAM header
     rmdup          remove PCR duplicates
     targetcut      cut fosmid regions (for fosmid pool only)
     addreplacerg   adds or replaces RG tags

  -- File operations
     collate        shuffle and group alignments by name
     cat            concatenate BAMs
```

オプション/引数無しで起動するとSamtools の基本的な使い方が表示される

基本的な使い方： `$ samtools` *`command options`*

Samtools の起動: コマンド簡易マニュアル

```
$ samtools view
```

```
Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]

Options:
  -b       output BAM
  -C       output CRAM (requires -T)
  -1       use fast BAM compression (implies -b)
  -u       uncompressed BAM output (implies -b)
  -h       include header in SAM output
  -H       print SAM header only (no alignments)
  -c       print only the count of matching records
  -o FILE  output file name [stdout]
  -U FILE  output reads not selected by filters to FILE [null]
  -t FILE  FILE listing reference names and lengths (see long help) [null]
  -L FILE  only include reads overlapping this BED FILE [null]
  -r STR   only include reads in read group STR [null]
```

- コマンドを付けてオプション無しで実行するとそのコマンドのマニュアルが表示される
- 詳細は http://www.htslib.org/doc/samtools.html を参照のこと

SAM/BAM 変換

samtools view *options...*

- SAMファイルからBAMファイルの作成

```
$ samtools view -bS etec_bowtie2.sam > ectec_bowtie2.bam
```

- BAMをSAMに変換して less コマンドで表示

```
$ samtools view etec_bowtie2.bam | less
```

- BAMファイルを読もうとすると…？

```
$ less etec_bowtie2.bam
```

- SAMファイルに比べてBAMファイルのサイズは？

```
$ ls -l etec_bowtie2.*
```

BAM ファイルのソート

samtools sort *options...*

- ソート
  - マッピングデータをリファレンス配列上の<u>位置順</u>に並び替える
  - 位置順のキー：
    - 染色体
    - アライメントの先頭の塩基の位置

```
$ samtools sort etec_bowtie2.bam -o
  etec_bowtie2_sorted.bam
```

- ソートされたBAMファイルをSAMに変換してlessで表示

```
$ samtools view etec_bowtie2_sorted.bam | less
```

- 元のSAMファイルの表示と比較

```
$ less etec_bowtie2.sam
```

ソートされたBAM（SAM）ファイルに インデックス（索引）を付ける

samtools index *options...*

- インデックス
  - 索引、目次、見出し
  - ファイルのどの辺りに何が書いてあるかの指標
  - 分厚い本の「別冊目次」のイメージ
    - 欲しい情報を探すのにファイル（本）を先頭から総ナメして探さなくてもよい

```
$ samtools index etec_bowtie2_sorted.bam
```

- インデックスは .bai という拡張子付きの別ファイルで生成される。
- 「bamファイル名.bai」が作成されたのを ls コマンドで確認

```
$ ls etec_boetie2_sorted*
```

ここから先はソート & インデックス付与したbamファイルを使う

ソート & インデックス付与したbamファイルを使って

指定した領域内のマッピング結果を表示

```
$ samtools view ecoli_sorted.bam chr:200-500
```

染色体名：開始位置 - 終了位置

指定した flag値を含むマッピング結果を表示

- flag 値 83 の説明を表示　　samtools flags *options...*

```
$ samtools flags 83
```

- flag 値 83 を含むリードを抽出 (view -f)

```
$ samtools view -f 83 etec_bowtie2_sorted.bam
```

- flag 値 2（両側が適切にマップされた） を含まないリードを抽出 (view -F)

```
$ samtools view -F 2 etec_bowtie2_sorted.bam
```

マッピングしたアプリケーションや実行時のオプションによって
フラグ値が全て同じになったりする場合もあるので注意

マッピング統計情報収集 1

samtools flagstat *options...*

- n本マップされm本マップされなかった
- ペアのマップ状況

```
$ samtools flagstat etec_bowtie2_sorted.bam
```

全リード数 → 200000 + 0 in total (QC-passed reads+QC-failed reads)
0 + 0 duplicates
マップされた → 48481 + 0 mapped (24.24%:nan%)
ペア → 200000 + 0 paired in sequencing
100000 + 0 read1
100000 + 0 read2
適切にペアになった (向きなど) → 47050 + 0 properly paired (23.52%:nan%)
ペアがマップされた → 47846 + 0 with itself and mate mapped
片側のみマップされた → 635 + 0 singletons (0.32%:nan%)
ペアが異なる染色体にマップされた → 0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

マッピング統計情報収集 2

samtools idxstats *options...*

- 染色体毎にマップされたリード数を得る

```
$ samtools idxstats etec_bowtie2_sorted.bam
```

染色体名	染色体配列長	マップされたリード数	片側のみマップされたリード数
ETEC_chr	4979619	155633	876
pETEC_80	79237	1310	15
pETEC_35	34367	131	1
pETEC_73	70609	271	44
....			
*	0	0	600

マップされなかったリード数

染色体名が '*' として表示される

マッピング統計情報収集3

samtools depth *options...*

- 深度（マップされた回数）の統計情報を得る

```
$ samtools depth etec_bowtie2_sorted.bam
```

染色体名	位置	深度（マップされた回数）
ETEC_chr	4503791	93
ETEC_chr	4503792	93
ETEC_chr	4503793	298
ETEC_chr	4503794	297
ETEC_chr	4503795	298
ETEC_chr	4503796	301

マッピング統計情報収集4

samtools mpileup *options...*

- リファレンスの位置ごとにマッピングされた塩基を表示

```
$ samtools mpileup -f eco_o139.fa etec_bowtie2_sorted.bam
```

```
ETEC_chr    24242   G   14      ..,,,,,,,,,,,,        GFIDIIGIH@IEHG
ETEC_chr    24243   A   15      .$.$,,,,,,,,,,,,^K,   DEI=IIDID@DEIGC
ETEC_chr    24244   G   14      ,$,,,,,,,,,,,,^K.     EDIIIIH@GEFDAE
ETEC_chr    24245   C   14      ,,,,,,,,,,,,.^K.      DIIGIDDGBHDCHB
ETEC_chr    24246   C   15      ,,,,,,,,,,,,..^K,^K,  EIIIIEAE<GGH:BE
```

- bcftools と組み合わせてSNPコールに使える
- リファレンスを指定しないと表示は変わる

mpileup の読み方

記号	意味	
. (ピリオド)	純鎖 リファレンスと一致	
, (カンマ)	逆鎖 リファレンスと一致	
ACGTN	純鎖 リファレンスと不一致	
acgtn	逆鎖 リファレンスと不一致	
+数字 (数字と同数のACGTNまたはacgtn)	挿入部位	+1aは逆鎖に"a"が1塩基挿入"。 +2ATは順鎖に"AT"が2塩基挿入。 部位はこの塩基と、次の塩基の間。
-数字 (数字と同数のACGTNまたはacgtn)	欠損部位	+1Aは順鎖"a"が1塩基欠損。 -atは逆鎖に"at"が2塩基欠損。
^とその後の一文字	リードの開始位置と マップ品質値	^~は"~"文字がASCIIコードの126番目なので、 126-33=93がマップ品質値。
$	リードの終了	
*（アスタリスク）	短い欠損の最中	近傍に2以上の欠損があり、 近傍の欠損に含まれる場合。 例えば-2ATの次の位置には"*"が含まれる。
<または>	長い欠損の最中	"*"とほぼ同じ意味だが、長い欠損の場合"<"または">"となる。

https://cell-innovation.nig.ac.jp/wiki/tiki-index.php?page=SAMtools#　より引用

簡易ビューワ

samtools tview *options...*

```
$ samtools tview etec_bowtie2_sorted.bam eco_o139.fa
```

- 見方はmpileupと同じ
- ？：ヘルプ表示（何かキーを押すとヘルプは消える）
- g ：見たい場所を入力する窓を表示
  - ETEC_chr:10000
  - （染色体名：位置）等と入力してリターン
- q：終了

Samtools まとめ

samtools

view	リードを抽出, SAM/BAM変換
sort	ソート
index	インデックス作成
flagstat	マッピング全体の統計情報表示
idxstats	染色体毎のマッピング状況
depth	位置毎のマッピング深度
mpileup	位置毎にマッピングされた塩基を表示
tview	簡易ビューワ

実習：特定のリードの抽出

実習1

`etec_bowtie2_sorted.bam` から以下の遺伝子にマップされたリードを取り出し、数を数えよ

染色体名	開始位置-終了位置	遺伝子名
ETEC_chr	336 - 2798	thrA
ETEC_chr	55624 - 56613	pdxA
ETEC_chr	4518271 - 4522299	rpoB

*実習2（発展問題）

`samtools tview` で rpoB の開始位置にジャンプして付近を見てみよう

実習：特定のリードの抽出 2

実習3

`etec_bowtie2_sorted.bam` から、ペアが存在して両方ともマップされていないリードを抽出して数を数えよ

抽出された行を数えるには、パイプ "|" で `wc` コマンドに流し込むこと

*実習4（発展問題）

`etec_bowtie2_sorted.bam` から、ペアが存在して両方がマップされたが、両方が適切にはマップされていないリードを抽出して数を数えよ

数値（10進）	意味
1	ペアリードがある
2	両方適切にマップされた
4	自分がマップされていない
8	ペア相手がマップされていない
16	逆鎖にマップされた
32	ペア相手は逆鎖にマップされた
64	ペアリードの１番目である
128	ペアリードの２番目である
256	Multiple hitでトップヒットでない
512	マッピングQVが低い

実習解答

実習1

```
$ samtools view etec_bowtie2_sorted.bam
  ETEC_chr:336-2798 | wc
```

他の遺伝子についても適宜 位置を変更して実行

実習3

```
$ samtools view -f 13 etec_bowtie2_sorted.bam | wc
```

less などで、実際に選ばれるリードのflag値を確認すること。
13ではない値が選出されているのはなぜだろうか

*実習4

```
$ samtools view -F 14 etec_bowtie2_sorted.bam | wc
```

`$ samtools flags 14` でフラグの意味を確認のこと

基生研ゲノムインフォマティックス・トレーニングコース　2016春
RNA-Seq入門
2016.03.10-2016.03.11

NGS基本ツールIGV

基礎生物学研究所
生物機能解析センター
山口勝司

データ可視化ツール・IGVの紹介・実習

https://www.broadinstitute.org/igv/

なぜIGVを取り上げるか

データ可視化ツール
・自分のパソコン（ローカル環境）にインストールして使うタイプ
・サーバーに構築して、ネットワークで使うタイプ

The JBrowse Genome Browser

JBrowse is a fast, embeddable genome browser built completely with JavaScript and HTML5, with optional run-once data formatting tools written in Perl.

Featured Post

Exploring structural variation using JBrowse by Richard Finkers

Latest Release – JBrowse 1.11.6

コミュニティーに広く利用、あるいはウェブ公開を目的とするには良いが、ネットワーク・情報セキュリティの高度な知識も要求される。

より大容量なデータに対応できる。

管理者的な人がいて、その人がやってくれるなら、これも良いが。

もっとお手軽なものとしてIGVを紹介

可視化ツールに求められるものは何か

膨大なデータを如何に直感的に理解できるようにするか
sortや絞り込みができる表データと対比双璧

・配列、GC ratio、遺伝子情報
・遺伝子発現情報
・SNPの位置情報・頻度情報
・様々なデータの精度情報

レファレンス配列 / gene model / gene annotationとNGSデータを並べて比較
複数のデータセットを並べて比較

色々なデータ（variant, 発現, ChIP, BSseq等々）を、様々なスケールで
比較・統合的に解釈できるようにしたい

ゲノムviewerに自分のデータを乗せ、
統合的直感的に解釈できること

可視化ツールをどう選ぶか

選択の基準
genome data viewing に求められるもの
取捨選択の基準

1. 無料 / 有料 / 基本無料

2. 個人的レベルの使用 / コミュニティーレベルの使用

3. 見るだけ/自分から色々工夫

4.アクセスのしやすさ・使いやすさ
導入に必要なコンピュータスペック
マニュアルは分かりやすいか
情報の多さ
利用の簡便さ
使っている人が近くにいるか

Integrative Genomics Viewer(IGV)

お手軽ツール

・アカデミックウェアで無料
・コミュニティーでの利用者が多いから、情報も多い
・javaのプログラムなので、オールプラットフォーム対応
・マニュアルは親切、サンプルデータのある
・WEBサーバーではなく、PCレベルでできる
・データ閲覧環境の共有が可能

誰もが簡便に使えるものが良い。

Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - Hosted Genomes
  - FAQ
  - IGV User Guide
  - File Formats
  - Release Notes
  - IGV for iPad
  - Credits
- Contact

Search website

search

Broad Home
Cancer Program

BROAD INSTITUTE

© 2013 Broad Institute

Home

Integrative Genomics Viewer

Overview

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011)

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** Briefings in Bioinformatics 14, 178-192 (2013).

Downloads

Please register to download IGV. After registering, you can log in at any time using your email address.

Funding

Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.

IGV participates in the GenomeSpace initiative, which is funded by the National Human Genome Research Institute.

GENOMESPACE

---

nature biotechnology

Login

Search Go Advanced search

nature.com ▸ journal home ▸ archive ▸ issue ▸ opinion and comment ▸ correspondence ▸ abstract

◂ previous abstract next abstract ▸

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Integrative genomics viewer

James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz & Jill P Mesirov

Affiliations | Corresponding authors

Nature Biotechnology **29**, 24–26 (2011) | doi:10.1038/nbt.1754
Published online 10 January 2011

To the Editor:

Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Analysis of these large, diverse data sets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large data sets and to flexibly integrate multiple data types, including clinical data. However, the sheer volume and scope of data pose a significant challenge to the development of such tools.

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

Subscribe today, save **50%** and receive **51** weekly issues of ***Nature*** in **print, online** and **mobile**.

Citations to this article
Crossref (10) Scopus (12) Web of Science (0)

Science jobs from naturejobs

Faculty Position
Harvard Medical School

Ramalingaswami Re-Entry Fellowship
Ministry of Science & Technology, Government of India

Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - Hosted Genomes
  - FAQ
  - IGV User Guide
  - File Formats
  - Release Notes
  - IGV for iPad
  - Credits
- Contact

Search website

search

Broad Home
Cancer Program

BROAD INSTITUTE

© 2013 Broad Institute

Home

Overview

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. **Integrative Genomics Viewer**. Nature Biotechnology 29, 24–26 (2011)

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** Briefings in Bioinformatics 14, 178-192 (2013).

Downloads

Please register to download IGV. After registering, you can log in at any time using your email address.

Funding

Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.

IGV participates in the GenomeSpace initiative, which is funded by the National Human Genome Research Institute.

GENOMESPACE

Integrative Genomics Viewer

- Home
- Downloads
- Documents
  - Hosted Genomes
  - FAQ
  - IGV User Guide
    - User Interface
    - Starting IGV
    - Navigating
    - Loading a Genome
    - Viewing the Reference Genome
    - Loading Data and Attributes
    - Viewing Data
    - Viewing Alignments
    - Viewing Variants
    - Gene List View
    - Regions of Interest
    - Sample Attributes
    - Sorting, Grouping, and Filtering
    - Saving and Restoring Sessions
    - Server Configuration
    - External Control of IGV
    - igvtools
    - Motif Finder
  - File Formats
  - Release Notes
  - IGV for iPad
  - Credits
- Contact

Home › IGV User Guide

IGV User Guide

This guide describes the Integrative Genomics Viewer (IGV).

- To start IGV, go to the IGV downloads page: http://www.broadinstitute.org/igv/download.

Look at a printer-friendly HTML version of the whole User Guide.

- User Interface
  - Starting IGV
  - Navigating
  - Loading a Genome
  - Viewing the Reference Genome
  - Loading Data and Attributes
  - Viewing Data
  - Viewing Alignments
  - Viewing Variants
  - Gene List View
  - Regions of Interest
  - Sample Attributes
  - Sorting, Grouping, and Filtering
  - Saving and Restoring Sessions
  - Server Configuration
  - External Control of IGV
  - Motif Finder
  - igvtools

User Interface ›

Nature Biotech. 29:24–26 (2011) Supplement figureからの抜粋

Downloads

Integrative Genomics Viewer (IGV) (Version 2.3)

Install IGV

Options for installing and running IGV:

1. (Mac only) Download and run the Mac application; or
2. (Windows) Download and run the self-extracting archive; or
3. (All systems) Use the Java Web Start buttons (Mac users: see below for limitations); or
4. (All systems) Download the binary distribution and run IGV from the command line.

Note: IGV 2.3.x requires Java 7. Users with Java 6 (JRE 1.6) should first try to upgrade Java to the latest version. If this is not possible you will need to run a 2.2.x version available in the archive.

Mac

Download and unzip the Mac App archive, then double-click the IGV application to run it. The application can be moved to the "Applications" folder, or anywhere else.

Download Mac App

Windows

1. Download the Windows package and execute the self-extracting archive.
2. It will prompt you for a location to extract the folder, choose anywhere you like (e.g. your home folder).
3. On completion, open the new folder.
4. Double-click the file "igv.bat", it might appear as just "igv" depending on your settings.

Download Windows Package

Java Web Start (All Platforms)

The buttons below use Java Web Start (JWS) to install and launch IGV directly from our web site.

***Mac Users**: The Java Web Start option does not work for some users due to security settings. The recommended solution is to use the bundled Mac App from the link above. Alternatively you can try to work around this by right-clicking on the buttons and saving the "jnlp" file, then right-clicking on the saved "jnlp" file and select "Open With > Java Web Start".

Chrome: Chrome does not automatically launch the Java Web Start files by default. Instead, the launch buttons below will download a "jnlp" file. This should appear in the lower left corner of the browser. Double-click the downloaded file to run, or if on a Mac right-click and select "Open With > Java Web Start"

Windows users: To run with more than 1.2 GB of memory you must install 64-bit Java. ***Most Windows installs do not include 64-bit Java by default, even if the operating system is 64-bit***. Attempting to use the 2GB or greater launch options with 32-bit Java will result in the error *"could not create virtual machine"*.

Launch	Launch	Launch	Launch
Launch with 750 MB	Launch with 1.2 GB Maximum usable memory for Windows OS with 32-bit Java.	Launch with 2 GB Maximum usable memory for 32-bit MacOS.	Launch with 10 GB For large memory machines with 64-bit Java.

Binary Distribution

Download and unzip the binary distribution archive in a folder of your choosing. IGV is launched from a command prompt -- follow instructions in the "readme" file. To launch igv on Mac or Linux platforms use the shell script "igv.sh". On Windows use "igv.bat".

Download Binary Distribution

いくつかの生物種は登録されている

新たな生物種をロードすることも可能

ゲノムViewerなので次世代DNAシーケンサーのデータに限定されない。
マイクロアレイの結果や、ゲノムアノテーションの情報も随時表示できる。

対応するファイル形式に応じて、表示方法が決まる。

File Formats

- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG
- chrom.sizes

File Formats

IGV supports a number of different file formats for experimental data and genome annotations. For a complete list of supported formats see http://www.broadinstitute.org/igv/FileFormats. The following table shows the recommended file formats for a number of common data types.

Source Data	Recommended File Formats
ChIP-Seq, RNA-Seq	WIG, TDF
Copy number	CN, SNP, TDF, canary_calls (Birdsuite)
Gene expression data	GCT, RES, TDF
Genome annotations	GFF, BED, GTF, PSL, UCSC table format
GISTIC data	GISTIC
LOH data	LOH, TDF
Mutation data	MUT, MAF
Variant calls	VCF
RNAi data	GCT
Segmented data	SEG, CBS
Sequence alignment data	BAM, SAM, PSL
Any numeric data	IGV, WIG, TDF
Sample metatadata	Tab-delimited sample info file

公開情報のviewerとして

その他の便利機能

セッションの保存

表示しているデータの読み込み状況を、それごと保存。

セッションをロードすることで、意図した画面を表示できる。

データセットが揃っていること、フォルダー構造が同一である必要がある。

バッチ処理

重要領域の画面スナップショットを自動で取ったりできる。

```
new
load myfile.bam
snapshotDirectory mySnapshotDirectory
genome hg18
goto chr1:65,289,335-65,309,335
sort position
collapse
snapshot
goto chr1:113,144,120-113,164,120
sort base
collapse
snapshot
```

Inversions

An inversion is a large section of DNA that is reversed in the subject genome compared to the reference genome.

When an inversion shows up in paired-end reads, the reads are distinctively variant from the reference genome.

This appears in IGV as shown below.

Interpreting Color by Insert Size

The inferred insert size can be used to detect structural variants, such as:

- deletions
- insertions
- inter-chromosomal rearrangements

IGV uses color coding to flag anomalous insert sizes. When you select Color alignments>by insert size in the popup menu, the default coloring scheme is:

- for an insert that is larger than expected
- for an insert that is smaller than expected
- 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y for paired end reads that are coded by the chromosome on which their mates can be found

Deletions

A deletion is a large section of DNA that is absent in the subject genome compared to the reference genome.

The "expected" insert size is the insert size obtained in sequencing the subject genome. The "inferred" insert size is the insert size that would result in the reference genome, assuming the same pair of reads.

RNA-Seqのデータ表示させる

GCTファイルでgene ローカスの発現情報を図示

```
#
#
Name        Description              2D_2D2L     4D_2D2L
ANAC001     |@Chr1:3630-5899         -2.60184    -2.60956
DCL1        |@Chr1:23145-33153       -0.742675   -1.5642
MIR838A     |@Chr1:23145-33153       0           0
AT1G01073   |@Chr1:44676-44787       0           0
IQD18       |@Chr1:52238-54692       -1.93871    -1.13128
AT1G01115   |@Chr1:56623-56740       0           0
GIF2        |@Chr1:72338-74737       -0.251287   -0.616679
AT1G01180   |@Chr1:75582-76758       0.45929     -0.809567
AT1G01210   |@Chr1:88897-89745       1.6964      0.857196
FKGP        |@Chr1:91375-95651       -0.174589   0.725947
AT1G01240   |@Chr1:99893-101834      -0.226384   -0.936641
AT1G01260   |@Chr1:108945-111609     -0.161848   0.315699
CYP703A2    |@Chr1:112262-113947     0           0
CNX3        |@Chr1:114285-116108     0.111249    -0.551359
AT1G01300   |@Chr1:116942-118764     -0.68348    0.108578
```

Gene listを定義して
サンプルごと
条件ごと
の発現・発現変動を
カラーマップできる

IGV実習

IGVの使用法を学ぶと共に
先のファイルフォーマットも
確認しよう

以下のファイルを確認

buc.genome.fasta
buc.gtf
buc_cg.wig
illumina_ex_B2_Read_bowtie2.mate.sort.bam
illumina_ex_B2_Read_bowtie2.mate.sort.bam.bai
illumina_ex_B4_Read_bowtie2.mate.sort.bam
illumina_ex_B4_Read_bowtie2.mate.sort.bam.bai

登録されていない生物種・配列でも、自分でimportすればOK

Buchneraゲノムを使う

Genome -> Load Genome from File
buc.genome.fasta
を読み込む

ゲノム構造を記述したgtfファイルを読み込む

File-> Load from File
buc.gtf ファイルを読み込む

buc.gtf

gn:buc	KEGG:GENES	CDS	197	2083	.+	geneid "BU001"; transcript_id "BU001-RA"; gene_name "gidA";
gn:buc	KEGG:GENES	CDS	2278	3102	.+	geneid "BU002"; transcript_id "BU002-RA"; gene_name "atpB";
gn:buc	KEGG:GENES	CDS	3139	3378	.+	geneid "BU003"; transcript_id "BU003-RA"; gene_name "atpE";
gn:buc	KEGG:GENES	CDS	3497	3982	.+	geneid "BU004"; transcript_id "BU004-RA"; gene_name "atpF";
gn:buc	KEGG:GENES	CDS	3982	4515	.+	geneid "BU005"; transcript_id "BU005-RA"; gene_name "atpH";
gn:buc	KEGG:GENES	CDS	4530	6068	.+	geneid "BU006"; transcript_id "BU006-RA"; gene_name "atpA";
gn:buc	KEGG:GENES	CDS	6101	6973	.+	geneid "BU007"; transcript_id "BU007-RA"; gene_name "atpG";
gn:buc	KEGG:GENES	CDS	6997	8394	.+	geneid "BU008"; transcript_id "BU008-RA"; gene_name "atpD";
gn:buc	KEGG:GENES	CDS	8421	8837	.+	geneid "BU009"; transcript_id "BU009-RA"; gene_name "atpC";
gn:buc	KEGG:GENES	CDS	8911	11322	.-	geneid "BU010"; transcript_id "BU010-RA"; gene_name "gyrB";
gn:buc	KEGG:GENES	CDS	11449	12549	.-	geneid "BU011"; transcript_id "BU011-RA"; gene_name "dnaN";
gn:buc	KEGG:GENES	CDS	12554	13918	.-	geneid "BU012"; transcript_id "BU012-RA"; gene_name "dnaA";
gn:buc	KEGG:GENES	CDS	14369	14512	.+	geneid "BU013"; transcript_id "BU013-RA"; gene_name "rpmH";
gn:buc	KEGG:GENES	CDS	14525	14872	.+	geneid "BU014"; transcript_id "BU014-RA"; gene_name "rnpA";
gn:buc	KEGG:GENES	CDS	15011	16609	.+	geneid "BU015"; transcript_id "BU015-RA"; gene_name "yidC";

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
All
Go
どれか1つの
染色体
scafoldを選ぶ
gn:buc
buc.gtf
2 tracks
537M of 635M

ビューイングの拡大縮小
IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
13 kb
180 kb
182 kb
184 kb
186 kb
188 kb
190 kb
buc.gtf
BU167-RA
gn:buc:180565-181800
gn:buc:180565-181800
geneid=BU167
transcript_id=BU167-RA
gene_name=folC
カーソルのgene modelに
持っていくことで、gtfファイル
の記述がpop upされる。
トラックの移動
マウスのドラッグ移動が可能
← →による移動も可能
Feature trackを指定して、
Ctr+F gene model単位で右に移動
Ctr+B gene model単位で左に移動
Shift+Ctr+F exon単位で右に移動
2 tracks
gn:buc:181,552
294M of 646M

シーケンスデータを読み込ませる
File -> Loead from File

illumina_ex_B2_Read_bowtie2.mate.sort.bam

同ファイル名のindexファイル
illumina_ex_B2_Read_bowtie2.mate.sort.bam.bai
が同じフォルダーにないといけない。

もう一つ別の系統のシーケンスデータも読み込ませる
illumina_ex_B4_Read_bowtie2.mate.sort.bam

各Trackはドラッグアンドドロップで移動できる

WIGファイルを読み込む
ウィンドウサイズ50baseでCG ratioを計算したWIGファイルを作成してある
buc_cg.wig

```
variableStep chrom=gn:buc
25      16
26      16
27      16
28      16
29      16
30      14
31      12
32      12
33      10
34      10
35      8
36      8
37      8
38      8
```

File -> Load from File でそのまま読み込ませることも可能だが、
テキストファイルで膨大なサイズだと非常に時間がかかる。
Tools -> Run igvtools

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
IGVtoolsが起動
13 kb
Command Tile
Input File G:¥training_course¥2013春¥配布_Data¥05_NGS_IGV¥buc_cg.wig
Output File G:¥training_course¥2013春¥配布_Data¥05_NGS_IGV¥buc_cg.wig.tdf
Genome G:¥training_course¥2013春¥配布_Data¥05_NGS_IGV¥buc.genome.fasta
Browse
Tile and Count Options
Zoom Levels 7
Window Functions
Min
Max
Mean
Median
2%
10%
90%
98%
Probe to Loci Mapping
Window Size 25
Sort Options
Temp Directory
Max Records 500000
Close
Run
Messages
buc.gtf
6 tracks loaded
gn:buc:181,667
197M of 736M

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
WIGファイルがバイナリー化されたtdfファイルができるので、
それをFile -> Load from file
188 kb
190 kb
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
buc.gtf
7 tracks
gn:buc:184,355
452M of 636M

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:177,949-191,848
Go
trackを右クリックでポップアップが出る
buc_cg.wig.tdf
Rename Track...
Change Track Color (Positive Values)...
Change Track Color (Negative Values)...
Change Track Height...
Change Font Size...
Type of Graph
Heatmap
Bar Chart
Points
Line Plot
Windowing Function
Mean
None
Set Data Range...
Set Heatmap Scale...
Log scale
Autoscale
Show Data Range
Kaplan-Meier Plot...
Remove Track
Save image...
適切な表示形式に変更する
例）
Type of Graph -> Line Plot
Windowing Function -> Set Data Range
-> min 0
-> mid 50
-> max 100
buc.gtf
7 tracks
gn:buc:178,077
506M of 636M

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:153,880-167,779
Go
13 kb
buc_cg.wig.tdf
トラックの移動
マウスのドラッグ移動が可能
← →による移動も可能
Feature trackを指定して、
Ctr+F gene model単位で右に移動
Ctr+B gene model単位で左に移動
Shift+Ctr+F exon単位で右に移動
buc.gtf
7 tracks loaded
gn:buc:153,951
418M of 630M

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
buc.genome.fasta
gn:buc
gn:buc:153,880-167,779
Go
13 kb
154 kb
156 kb
158 kb
160 kb
162 kb
164 kb
166 kb
NAME
DATA TYPE
DATA FILE
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.sort.bam
illumina_ex_B4_Read_bowtie2.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.sort.bam
buc.gtf
7 tracks loaded
484M of 630M
illumina_ex_B2_Read_bowtie2.mate.sort.bam
Rename Track...
Copy read details to clipboard
Group alignments by
Sort alignments by
Color alignments by
Shade base by quality
Show mismatched bases
Show all bases
View as pairs
Go to mate
View mate region in split screen
Set insert size options ...
Re-pack alignments
Show coverage track
Load coverage data...
Collapsed
Expanded
Squished
data trackを右クリックすると、
track内の表示方式変更可能
View as pairs
を選んで見る。

IGV
File View Tracks Regions Help
buc.genome.fasta
gn:buc
gn:buc:316,735-323,248
Go
317,000 bp
323,000 bp
NAME
DATA TYPE
DATA FILE
pair endを維持した表示になる
insertion, deletionを見たい場合はこちらが良い。
buc_cg.wig.tdf
illumina_ex_B2_Read_bowtie2.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.sort.bam
illumina_ex_B4_Read_bowtie2.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.sort.bam
buc.gtf
BU289-RA
BU290-RA
BU291-RA
BU292-RA
BU293-RA
BU294-RA
7 tracks
gn:buc:318,057
254M of 493M

IGV
File View Tracks Regions Help
buc.genome.fasta
gn:buc
gn:buc:86,265-90,071
VCF ファイルを読み込ませてみる
IGVtoolsでindexを作成
File -> Load from File
87,000 bp
88,000 bp
89,000 bp
90,000 bp
buc_cg.wig.tdf
bowtie2_B2_var.flt.vcf
/home/kyamag...ate.sort.bam
bowtie2_B4_var.flt.vcf
/home/kyamag...ate.sort.bam
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
BU084-RA
BU085-RA
BU087-RA
9 tracks loaded
gn:buc:86,508
394M of 701M

IGV
File View Tracks Regions Help
buc.genome.fasta
gn:buc
gn:buc:86,265-90,071
Go
3,801 bp
87,000 bp
88,000 bp
popupが邪魔な時はここをクリック、
あるいはCtrlキーを押していれば
表示されない。
read dataにカーソルを持っていくと、その情報が表示される
buc_cg.wig.tdf
bowtie2_B2_var.flt.vcf
/home/kyamag...ate.sort.bam
bowtie2_B4_var.flt.vcf
/home/kyamag...ate.sort.bam
illumina_ex_B2_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B2_Read_bowtie2.
.sort.bam
illumina_ex_B4_Read_bowtie2.
.sort.bam Coverage
illumina_ex_B4_Read_bowtie2.
.sort.bam
buc.gtf
BU084-RA
Left alignment
Read name = HWI-ST676_0085:5:1:14919:65430#CGATGT
Alignment start = 86599 (+)
Cigar = 101M
Mapped = yes
Mapping quality = 42
Pair start = gn:buc:86763 (-)
Pair is mapped = yes
Insert size = 265
Pair orientation = F2R1
Second in pair
MD = 101
XG = 0
NM = 0
XM = 0
XN = 0
XO = 0
AS = 0
YS = 0
YT = CP
Right alignment
Read name = HWI-ST676_0085:5:1:14919:65430#CGATGT
Alignment start = 86763 (-)
Cigar = 101M
Mapped = yes
Mapping quality = 42
Base = T
Base phred quality = 33
Pair start = gn:buc:86599 (+)
Pair is mapped = yes
Insert size = -265
Pair orientation = F2R1
First in pair
MD = 101
XG = 0
NM = 0
XM = 0
XN = 0
XO = 0
AS = 0
YS = 0
YT = CP
9 tracks loaded
gn:buc:86,815
563M of 701M

IGV紹介のまとめ

可視化ツールとして十分な機能を持つ

・無料

・比較的簡単・お手軽

・自分で見るためにも良し、人に見せるためにも良し

・利用範囲は次世代DNAシーケンサーに限定しない
広くゲノミクスの解析に有用

ごく一部のみの機能を紹介しました。
ウェブサイトを見ながら復習をお勧めします。

統計学入門

慶應義塾大学　先端生命科学研究所
佐藤昌直

これらをこれから学習していくためには

- 汎用される統計の仕組みを知る
- 測定、実験計画を見直す
- 教科書を読めるように統計用語・表記に慣れる
- 道具を準備する - R

私が重視しているポイント

- 研究全体における統計の役割、**実験と統計との連携**を意識する
- 遺伝子発現解析に必要な**統計の基礎概念**を解説する
- ***"statistical mind"***を養う

基本的な統計の用途

- 仮説検定
- 予測 (モデル構築)

遺伝子発現解析における統計の役割

遺伝子発現解析における統計の役割

仮説検定 - *t* 検定を例に

ねらい

*t*検定から検定の背景知識を得る:

- 検定の流れを知る
- 勉強のとっかかりを作る

用語の意味の整理

- 統計量、確率分布、自由度、*p*値

統計における検定の手続き

1. 仮説を立てる
2. 統計量を求める
3. 求めた統計量を確率分布に照らし合わせる
4. 判定: 求めた確率と棄却限界値との比較

ポイント

2. 統計量を求める:

統計量: データから導いた
<u>具体的</u>な数値

↔ **母数**: 未知の数値

我々ができること: 少数の測定値（**標本**）から
「**母集団**」を推定すること

ポイント

1. 仮説を立てる:

帰無仮説

最終的に棄却される仮定:

「AとBに差がある」かを検定する場合は
「AとBには差がない」と仮定する

例1. 野生型と変異体Aの遺伝子xの発現量に違いがあるか？

例2. 野生型と変異体Aの遺伝子発現プロファイル間の相関係数は0.35だった。これらは有意に相関していると考えられるか？

3. 確率分布と照らし合わせる

棄却限界値によって
規定される面積
(通例: 全体の5%)

4. 判定: 帰無仮説が棄却されるか？

最終的に棄却される仮定:

「AとBに差がある」かを検定する場合は「AとBには差がない」という仮定

代表値

平均値: 相加平均。すべてのデータを足して、データ数で割って得られる値

$$\bar{x} = \hat{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

中央値: データを小さいものから順に並べたときに中央にくる値

ポイント

2. 統計量を求める:

統計量: データから導いた <u>具体的</u>な数値

↔ **母数**: 未知の数値

我々ができること: 少数の測定値（**標本**）から「**母集団**」を推定すること

ばらつき: **分散／偏差**

分散:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

標準偏差:

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

***n* -1?**

なぜ、平均を求める時と分散を求める時では分母が変わるのか？

自由度: 統計量を求めるのに使うことができる**「独立」**な標本数

統計的検定の手続き　　t検定

1. **仮説**を立てる	2つのサンプル間で遺伝子発現量（平均値）の違いがある？
2. **統計量**を求める	平均、標準誤差、自由度からt統計量を求める
3. 求めた統計量を**確率分布**に照らし合わせる	t分布からp値を求める
4. 判定: 求めた確率と**棄却限界値**との比較	有意差の判定

母集団を推定する統計量

(標本)平均:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

-（バー）は平均を表す

標準偏差: $\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

^（ハット）は推定を表す

ポイント

t検定:

2サンプルの平均の検定

- 平均値 = μ_1, μ_2
- データは正規分布

ほぼ全ての検定方法に前提がある

母集団を推定する統計量

1. (真の値に近い)代表値

2. ばらつきの範囲

統計量その2：
平均値も**あくまで推定値**

(平均) 標準誤差:
「統計量」の偏差

$$SEM = \frac{s}{\sqrt{n}}$$

統計量その1

平均値: 相加平均。すべてのデータを足して、データ数で割って得られる値

$$\overline{x} = \hat{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

統計量その3:
平均の差とその誤差

<u>*t*統計量</u>

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{SEM_{|\hat{\mu}_1 - \hat{\mu}_2|}}$$

確率分布-*t* 分布

- 得られたt統計量がどのくらいの確率で起きうるか
- t分布（確率分布）を標本のt統計量と自由度を使って参照

【おさらい】自由度: 統計量を求めるのに使うことができる独立な標本数

ポイント

データの分布、仮説検定に即した確率分布を使う

我々の測定では

母分散が**未知**
したがって確率密度は**自由度**によって変化

→正規分布ではなく、t分布

例) 3つの観察で得られた平均値と100観察から得られた平均値はどちらが確からしいか

*p*値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 多くの場合、0.05が危険率

統計における検定の手続き

1. 仮説を立てる
2. 統計量を求める
3. 求めた統計量を確率分布に照らし合わせる
4. 判定: 求めた確率が棄却限界値より大きいか、小さいか

研究の手順（危険な例）

実験を計画する

↓

実験を行う

↓

結果の分析

統計解析の結果によって
結論を導く

多重検定の補正

+ 統計検定における重要な思考

ポイント

現実には: 実験デザインはデータを取得する「前」に練ってある必要がある

実験を計画する

↓

実験を行う

↓

結果の分析

統計解析の結果によって
結論を導く

ほぼ全ての検定方法に
前提がある

p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 多くの場合、**0.05**が危険率

p値とは：

- 標本に基づいた統計量が帰無仮説の下、起きうる確率
- 多くの場合、**0.05**が危険率
  = 100回に5回起きる

多重検定の補正

1. **Bonferroniタイプ**
2. **False discovery rate (FDR):**
   - Benjamini-Hochberg
   - Storey

多重検定の補正

- p = 0.05の検定を100回*繰り返すと、
  5回はランダムに間違い

*NGS解析では数万回以上繰り返すことになります

Bonferroniタイプの多重検定の補正

危険率を検定数で調整

危険率 = α/k

α: 元の危険率、
k: 検定数

False Discovery Rate (FDR)

帰無仮説

観察

ポイント

*p*値、*q*値の違い

*p*値の視点: **FP/(TN+FP)**

*q*値の視点: **TP/(TP+FP)**

		Statistical test	
		positive	**negative**
Real	**+**	**True positive**	**False negative**
	-	**False positive**	**True negative**

False Discovery Rate (FDR)

q値:

補正されたp値。そのq値以下の検定のうち、どのくらいの割合でfalse positiveが含まれているか。

復習／発展学習

- 検定の手順
  - 統計量
  - 自由度
  - p値
- 統計解析の結果は確率に判断して得られたもの、トランスクリプトーム解析ではそれを多数行う
  → 多重検定の補正
- 検定方法、多重検定の補正における仮定
  例) 時系列データの比較にFDRは使えない

データのばらつきと
実験デザイン・統計学的観点

我々の実験対象の例

- ある遺伝子型の生物の
- ある環境での　　**＋ 制御不能な実験要因**
- ある遺伝子の発現量 **＋ 生化学反応のノイズ**

我々に実行できる事

少数の測定値（**標本**）から
「**母集団**」を推定すること

母集団を推定する統計量

1. (真の値に近い)代表値

2. ばらつきの範囲

測定データはバラつく

- **実験（測定）を反復する**
- 何を**「真」**と考えるか
- 論文として発表できるデータには**再現性**が必要

反復？

例: バクテリア増殖定量

Katagiri, Thilmony R, and He S (2002) The Arabidopsis Thaliana-Pseudomonas Syringae Interaction. The Arabidopsis Book.

反復？

反復？

我々が1データポイントの測定で得ているもの

- 生物学的にばらつきの中のある1点
- 測定技術のばらつきの中のある1点

我々が1データポイントから得ているもの

- 生物学的にばらつきの中のある1点
- 測定技術のばらつきの中のある1点

測定における2要因

- Precision - 精度
- Accuracy - 正確度

測定における2要因

- **Precision - 精度**

  ある1測定を繰り返した際のばらつきの尺度

ポイント

測定における2要因

Real value: 真の値

Measured mean: 測定値から得られた平均

Harm van Bakel & Frank C.P. Holstege (2004) *EMBO reports* (2004) 5, 964 - 969

測定における2要因

- **Accuracy - 正確度**

  ある測定値が「真の値」にどれだけ近いかの尺度

我々にできる事

少数の測定値（**標本**）から「**母集団**」を推定すること

生体サンプルを繰り返し取る: **biological replicates**	同一サンプルを繰り返し測る: **technical replicates**

我々が1データポイントから得ているもの

- **生物学的にばらつきの中のある1点**
- **測定技術のばらつきの中のある1点**

“マーカー遺伝子”測定

- 何が再現されうるか？再現されたとするか？

Walley *et al.* (2008) *PLOS Pathog.*

明瞭な違いを示す遺伝子: **明瞭な再現性**

定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？
  - いつ行っても再現できる?
  - どこで行っても再現できる?
  - 誰が行っても再現できる?

定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？
  - ~~いつ行っても再現できる?~~
  - ~~どこで行っても再現できる?~~
  - ~~誰が行っても再現できる?~~

バラつきの定量と**割当て**

何を知るための実験か？
再現性のあるデータとは何か？
どのように反復を行うのが適切か？

解析の流れ

分散分析・線形モデル:

多変数データを系統立てて解析する
- 実験デザインと統計の連携

目標

- 線形モデルの**概念**を掴む
- **実験デザインがどう統計に影響するか**を考えるきっかけとする

リマインド:
母集団を推定する統計量

あるRT-qPCR実験

- genotype A, Bについて
- 6検体ずつ3回反復して計測

- genotype: A, B
- replicate: 1, 2, 3
- value:

  計18個/ genotype

t検定: 平均値の検定

$$x_i = \overline{x} + \underbrace{(x_i - \overline{x})}_{\text{偏差: 平均値からのばらつき}}$$

$$x_i = \overline{x} + (x_i - \overline{x})$$

線形モデルの枠組みで考えてみる

$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \underline{\varepsilon_i}$$

残差 (観察値-推定値):
想定要因では説明できない
データの変動

ポイント

“トランスクリプトーム”測定

- 何が再現されうるか？再現されたとするか？

網羅的測定:
再現性の
再定義

Chen *et al.* (2015) edgeR User’s Guide page 63

ポイント

考慮するのは1要因で良いか？

観察値を複数要因の影響に起因するものとして分解

$$x_i = \bar{x} + (x_i - \bar{x})$$

$$x_i = \bar{x} + \varepsilon_i$$

$$\downarrow$$

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

***genotype*と*replicate*の**
影響を同時に
考えられないか？

例: 2遺伝子型の測定を3反復したデータ

$(\bar{x}_{i\bullet} - \bar{O})$ 遺伝子型による変動

線形モデルの仕組み

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

$$O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}$$

$(\bar{y}_{\bullet j} - \bar{O})$ 反復ごとの変動

各計測値は $\boxed{O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}}$ と表せる

線形モデルとは

応答変数　〜　説明変数1 + 説明変数2 + ... + 誤差

と観察値を説明する（かもしれない）
変数でそれらの関係性を書き下すこと

- 実際には: Rでlmなどの関数を使う

分散分析・線形モデルの枠組み

$$O_{ij} = x_i + y_j + \varepsilon_{ij}$$

$$O_{ij} = \bar{O} + (\bar{x}_{i\bullet} - \bar{O}) + (\bar{y}_{\bullet j} - \bar{O}) + \varepsilon_{ij}$$

教科書・論文風に書くと

$$O_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

応答変数　　説明変数

実験デザインの重要性

- -omicsデータは“batch effect”という体系的なバイアスが混入する。
  例: 実験時期、餌

OPINION
Tackling the widespread and critical impact of batch effects in high-throughput data
Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly and Rafael A. Irizarry

Nature Reviews Genetics (2010) 11, 733-

- 線形モデルで推定・除去

ポイント

実験デザインの重要性

- 線形モデルで推定・除去

$$O_{ij} = \alpha_i + \beta_j + \varepsilon_{ij}$$

α_i : 遺伝子型／処理など注目している効果の要因

β_j : 反復（実験日時）／実験者などバイアス要因

- α_i の推定値、標準誤差のみを使う

定量的測定が可能且つ要求される時代の再現性のあるデータとは何か？

- 何が再現されうるか？再現されたとするか？

- ~~いつ行っても再現できる?~~
- ~~どこで行っても再現できる?~~
- ~~誰が行っても再現できる?~~

バラつきの定量と**割当て**

実験デザインの重要性:
genotype+replicate+potモデルを当てはめるには？

実験デザインの重要性:
genotype+replicate+potモデルを当てはめるには？

実験デザインの重要性:
genotype+replicate+potモデルを当てはめるには？

↑

genotypeとpotが独立ではない
（切り分けられない）

まとめ

- 計測データセットに影響を与える要因が一つではない場合、分散分析・線形モデルの枠組みが有効
- 理屈は難しいかもしれないが、Rで簡単に実行できるので実験デザインと連動したモデルを立てることが重要

ポイント

実験デザインの重要性

- 要因効果を推定するための実験デザイン
  - 各実験要因を適切に反復させた実験デザイン
- 実験デザインとモデル
  - 要因: データ取得**「前」**に想定しておくもの
  - データの変動を説明しない要因を解析時に減らすことは可能。実験デザイン時に計画しなかった要因を増せない。

復習／発展学習

- 回帰（最小二乗法）
- 実験計画法
- 交互作用
- Bioconductor: limma、edgeRパッケージ

RNA-seqの解析パイプライン

RNA-seq Analysis Pipeline

Shuji Shigenobu
NIBB, Japan
<shige@nibb.ac.jp>

RNA-seq analysis pipeline for DE

Differential Expression analysis

Millions of short reads

↓ **pre-processing**
mapping

Reads aligned to reference

↓ **count by unit**
(gene, transcript, exon)

Table of counts

↓ **normalization**
DE testing

List of DE gene

↓ **systems biology**
- GO enrichment
- multivariate analysis
- network analysis

Biological insights

NGS reads

	#gene	m1	m2	m3	h1	h2	h3
2	AT1G01010	35	77	40	46	64	60
3	AT1G01020	43	45	32	43	39	49
4	AT1G01030	16	24	26	27	35	20
5	AT1G01040	72	43	64	66	25	90
6	AT1G01050	49	78	90	67	45	60
7	AT1G01060	0	15	2	0	21	8
8	AT1G01070	16	34	6	9	20	1
9	AT1G01080	170	191	382	127	98	184
10	AT1G01090	291	346	563	171	116	453
11	AT1G01100	113	125	246	78	27	361
12	AT1G01110	0	1	1	0	0	0
13	AT1G01120	226	189	270	147	83	174
14	AT1G01130	9	11	1	0	2	
15	AT1G01140	181	120	142	161	73	
16	AT1G01150	0	2	0	0	0	
17	AT1G01160	117	125	215	86	46	
18	AT1G01170	74	57	82	36	22	
19	AT1G01180	46	7	26	24	18	
20	AT1G01190	0	3	2	1	2	
21	AT1G01200	5	0	2	0	0	
22	AT1G01210	178	203	98	205	83	
23	AT1G01220	26	49	40	21	15	
24	AT1G01225	4	10	6	6	0	
25	AT1G01230	72	51	58	70	18	
26	AT1G01240	81	89	45	62	24	
27	AT1G01250	1	1	5	1	2	
28	AT1G01260	15	52	37	33	27	
29	AT1G01290	7	16	23	30	5	
30	AT1G01300	75	115	232	89	109	

Table of counts

	gene	signal (A)	logFC (M)	p-value	FDR
2	AT5G48430	-9.17	6.30	3.7E-27	9.8E-23
3	AT2G08986	-9.15	5.04	2.8E-21	3.6E-17
4	AT5G31702	-9.74	5.68	5.0E-21	4.3E-17
5	AT1G51850	-10.07	5.32	2.2E-19	1.4E-15
6	AT2G39380	-9.82	4.97	2.9E-19	1.5E-15
7	AT2G44370	-10.23	5.48	1.1E-18	4.7E-15
8	AT3G46280	-8.19	4.93	1.3E-18	5.0E-15
9	AT3G55150	-10.44	5.86	1.8E-18	5.8E-15
10	AT2G19190	-8.70	4.55	8.5E-18	2.5E-14
11	AT1G51820	-9.40	4.40	1.3E-17	3.3E-14
12	AT2G39530	-9.16	4.36	1.7E-17	4.1E-14
13	AT1G40104	-9.58	4.61	2.2E-17	4.9E-14
14	AT2G30750	-9.97	4.64	1.5E-16	3.0E-13
15	AT2G17740	-9.12	4.18	1.8E-16	3.3E-13
16	AT2G07981	-10.07	4.75	1.0E-15	1.8E-12
17	AT1G30720	-9.21	3.91	3.5E-15	5.8E-12
18	AT5G31720	-9.12	3.76	1.9E-14	2.7E-11
19	AT1G51800	-8.47	3.99	1.9E-14	2.7E-11
20	AT1G30720	-9.06	3.75	4.7E-14	6.5E-11
21	AT4G12520	-6.58	4.39	1.5E-13	1.9E-10
22	AT1G07160	-10.76	5.02	2.0E-13	2.5E-10
23	AT2G39200	-7.57	3.98	6.0E-13	7.2E-10
24	AT2G10630	-9.30	3.50	8.8E-13	1.0E-09
25	AT1G65500	-9.60	3.50	9.6E-13	1.0E-09
26	AT5G39580	-9.24	3.45	1.3E-12	1.3E-09
27	AT5G44380	-9.51	3.45	1.7E-12	1.7E-09
28	AT5G13220	-7.20	3.88	2.8E-12	2.7E-09
29	AT2G45220	-8.45	3.54	3.9E-12	3.6E-09
30	AT1G30730	-9.80	3.45	6.0E-12	5.4E-09
31	AT1G64170	-9.91	3.39	1.2E-11	1.1E-08
32	AT4G20860	-8.10	3.49	1.5E-11	1.2E-08
33	AT4G23550	-10.08	3.49	1.5E-11	1.2E-08
34	AT4G12490	-6.75	3.88	1.6E-11	1.2E-08
35	AT5G64120	-7.09	3.76	2.1E-11	1.6E-08

List of DE genes
Table of significance score

Biological insights

Systems biology:
clustering and
network analysis

RNA-seq analysis pipeline for DE

	data type	format
Millions of short reads	sequences	fastq
↓ **pre-processing** **mapping**		
Reads aligned to reference	alignment	SAM/BAM
↓ **count by unit** (gene, transcript, exon)		
Table of counts	table	text (tab delimited)
↓ **normalization** **DE testing**		
List of DE gene	various	various
↓ **systems biology** - GO enrichment - multivariate analysis - network analysis		
Biological insights		

Two Basic Pipelines

- Choice of reference
  - **Genome** – standard for genome-known species
  - **Transcript** – the only way for genome-unknown species
    -- can be used for genome-known species

Common workflow

A Pipeline: Genome-based (1)

"Tuxedo pipeline"

Novel gene modeling by Cufflinks is not included in this pipeline.

A Pipeline: Genome-based (2)

Anders S, et al. (2013) *Nat Protoc*

A Pipeline: Transcript-based

Read QC and Pre-processing

- Read QC
  - Tools: FastQC etc.
- Pre-processing
  - Filter or trim by base quality
  - Remove artifacts
    - adaptors
    - low complexity reads
    - PCR duplications (optional)
  - Remove rRNA and other contaminations (optional)
  - Sequence error correction (optional)
  - Tools: `cutadapt`, `trimmomatic`

Martin et al (2011) *Nat Rev Genet*

RNA-seq analysis pipeline for DE

Input

- Typical primary data = matrix of #genes x #samples

column x number of samples (libraries)

row x number of genes (probes)

	A	B	C	D	E	F	G
1	#gene	m1	m2	m3	h1	h2	h3
2	AT1G01010	35	77	40	46	64	60
3	AT1G01020	43	45	32	43	39	49
4	AT1G01030	16	24	26	27	35	20
5	AT1G01040	72	43	64	66	25	90
6	AT1G01050	49	78	90	67	45	60
7	AT1G01060	0	15	2	0	21	8
8	AT1G01070	16	34	6	9	20	1
9	AT1G01080	170	191	382	127	98	184
10	AT1G01090	291	346	563	171	116	453
11	AT1G01100	113	125	246	78	27	361
12	AT1G01110	0	1	1	0	0	0
13	AT1G01120	228	189	270	147	83	174
14	AT1G01130	9	11	1	0	2	9
15	AT1G01140	181	120	142	161	73	134
16	AT1G01150	0	2	0	0	0	0
17	AT1G01160	117	125	215	86	46	212
18	AT1G01170	74	57	82	36	22	29
19	AT1G01180	46	7	26	24	18	58
20	AT1G01190	0	3	2	1	2	2
21	AT1G01200	5	0	2	0	0	0
22	AT1G01210	178	203	98	205	83	143
23	AT1G01220	26	49	40	21	15	34
24	AT1G01225	4	10	6	6	0	3
25	AT1G01230	72	51	58	70	18	77
26	AT1G01240	81	89	45	62	24	33
27	AT1G01250	1	1	5	1	2	2
28	AT1G01260	15	52	37	33	27	54
29	AT1G01290	7	16	23	30	5	19
30	AT1G01300	75	115	232	89	109	224

Import count table / diagnostics

<u>Look into the input data first.</u>

- Quick view of the table (tools: R, MS Excel etc.)
  - Check: Format, data structure, data size etc.
- Scatter plot, MA plot (tools: R, MS Excel etc.)

Let's try: data import and quick check

```
> dat <- read.delim("~/data/SS/arab2.txt", row.names=1)
> head(arab2)                                  # look at the first several lines
          m1 m2 m3 h1 h2 h3                    # for checking
AT1G01010 35 77 40 46 64 60
AT1G01020 43 45 32 43 39 49
AT1G01030 16 24 26 27 35 20
AT1G01040 72 43 64 66 25 90
AT1G01050 49 78 90 67 45 60
AT1G01060  0 15  2  0 21  8

> dim(dat)                                     # get numbers of rows and columns
[1] 26221     6

> colSums(dat)                                 # get column sums
     m1      m2      m3      h1      h2      h3
1902032 1934029 3259705 2129854 1295304 3526579
```

演習問題 ex3

Let's try: Scatter plot

```
> plot(dat$m1 + 1, dat$m2 + 1, log="xy")
```

MA plot

M: log fold-change
A: log intensity average

$$M = log_2(R/G) = log_2(R) - log_2(G)$$
$$A = \frac{1}{2}log_2(RG) = \frac{1}{2}(log_2(R) + log_2(G))$$

R: expression level of sample 1
G: expression level of sample 2

演習問題 ex4

Table of counts

data import

diagnostics

normalization

DE testing

evaluation

List of DE gene

Normalization

What is normalization? Why it is required?

- Normalization means to adjust transcriptome data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between genes.
- Normalization is an essential step in the analysis of DE from RNA-seq data to make them really comparable.

Normalization: two types

- Between-libraries
  - Comparing expression (counts) of genes between libraries
  - Adjust by the total number of reads
    - RPM (Reads Per Million mapped reads)
- Within-library
  - Comparing expression (counts) of genes within a library (should be possible with NGS – in contrast to microarray)
  - longer transcripts have higher counts => RPM + adjust by length
    - RPKM (Reads Per Kb per Million mapped reads)
    - FPKM (Fragments Per Kb of exons per Million fragments Mapped)

DEG: RNA-seq specific issues

- RNA-seq count data is Non-Gaussian
- Normalization: composition effects
- *N* (biological replicates) is so small
- Multiple comparisons (多重検定の問題)

RNA-seq data is Non-Gaussian

- RNA-seq data
  - Discrete-valued data（離散値）
  - Not normally distributed random variables
  - **Poisson distribution** for technical replicates
  - **Negative binomial distribution** for biological replicates.
    （負の二項分布）

RNA-seq issue: Normalization

- Simple normalization
  - RPM or RPKM works well, but not best
- Composition effects
  - A small number of highly expressed genes can consume a significant amount of the total sequence.
  - Strategies
    - estimate scaling factors from data and statistical models
    - quantile normalization
    - …

Implenation examples: edgeR and Cuffdiff

edgeR

- **Model**: An over dispersed Poisson model, **negative binomial (NB) model** is used
- **Normalization: TMM method** (trimmed mean of M values; Robinson et al., 2010), RLE (Anders et al., 2010) and upperquantile (Bullard et al., 2010)

Cuffdiff

- **Model**: FPKM, Geometric, quartile
- **Normalization:** Pooled (default), per-condition, blind, Poisson

RNA-seq analysis pipeline for DE

基生研ゲノムインフォマティックス・トレーニングコース　2016春
RNA-Seq入門
2016.03.10-2016.03.11

RNA-Seqパイプライン
ゲノムベースの解析法

基礎生物学研究所
生物機能解析センター
山口勝司

genomeをレファレンスとする場合

レファレンスがゲノム配列の場合
イントロン配列のスプライシングを考慮した
アライメントを行う必要がある。
TopHatを用いる
他　Blat, SpliceMap, MapSplice, GSMAP, QPALMA

実際こんな感じにアラインされる

本トレーニングコースでの流れ

RNA-SeqにおけるPreprocessingの必要性

RNA-Seq解析において通常mappingはglobal matchが用いられる。
・部分的な配列でのmapを許容するとfalse positive mapが多くなる
・Global matchにおいて末端に余計な配列があるとmapしない

通常イルミナRNA-Seqライブラリーは200baseくらいの長さから存在するうち両端にアダプター63baseずつ

Preprocessing tools

現行では以下の2ツールが有名
・Cutadapt
・Trimmamotic

Paired end readに対応(ver. 1.8以降)
片方のreadが非常に短くしか残らない場合、そのpair read自体をcutする。

http://cutadapt.readthedocs.org/en/stable/guide.html

MacOSXでのcutadaptのインストール

Cutadapt install手順

Cython をダウンロード
http://cython.org/#download

cd Cython-0.23.4
sudo python setup.py install

cd ..

git clone
https://github.com/marcelm/cutadapt
cd cutadapt
sudo python setup.py install

現状最新はver. 1.9.2

Cutadapt

Removing adapters

Cutadapt supports trimming of multiple types of adapters:

Adapter type	Command-line option
3' adapter	-a ADAPTER
5' adapter	-g ADAPTER
Anchored 3' adapter	-a ADAPTER$
Anchored 5' adapter	-g ^ADAPTER
5' or 3' (both possible)	-b ADAPTER

Here is an illustration of the allowed adapter locations relative to the read and depending on the adapter type:

Cutしたいアダプター配列の
位置関係など詳細に指定可能

fastqファイルはgz圧縮してあってもよい
fastaファイルも可

$ cutadapt

cutadapt version 1.9.1
Copyright (C) 2010-2015 Marcel Martin <marcel.martin@scilifelab.se>

cutadapt removes adapter sequences from high-throughput sequencing reads.

Usage:
  cutadapt -a ADAPTER [options] [-o output.fastq] input.fastq

For paired-end reads:
  cutadapt -a ADAPT1 -A ADAPT2 [options] -o out1.fastq -p out2.fastq in1.fastq in2.fastq

最適な
QV値
minimum-length値
O値
を設定して行う。

crude_fastqフォルダーに生シーケンス配列
trim_fastqフォルダーにcutadaptにかけた配列を用意してあります

Single readの場合

```
$ cutadapt \
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC \
-o hoge_read1.cut.fastq \
hoge_read1.fasta
```

Paired end readの場合

```
$ cutadapt \
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC  \
-A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATC
-o hoge_read1.cut.fastq \
-p hoge_read2.cut.fastq \
hoge_read1.fasta \
hoge_read2.fasta
```

実習1
wcコマンドでread数を確認してみよう

実習2
crude_fastqのどれか1つのデータでcutadaptを試して見よう

```
例）
$ cutadapt \
-q 20 \
-O 5 \
--minimum-length 50 \
-a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC  \
-A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATC \
-o trim_4D_rep1_R1.fastq \
-p Trim_4D_rep1_R2.fastq \
4D_rep1_R1.fastq \
4D_rep1_R1.fastq
```

Tophat

TopHat
A spliced read mapper for RNA-Seq

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

» **TopHat 2.1.1 release 2/23/2016**
Please note that TopHat has entered a low maintenance, low support stage as it is now largely superseded by **HISAT2** which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and **much more efficient** way.
Version 2.1.1 is a maintenance release which includes the following changes, some of them thanks to GitHub contributors:
- TopHat can be now built on more recent Linux distributions with newer GNU C++ (5.x), as the included SeqAn library was finally upgraded to a newer version.
- improved the detection of linker options for the Boost::Thread library which prevented the TopHat build from source on some systems.
- incorporated Luca Venturini's code to support large bowtie2 indexes (.bt2l).
- `bam2fastx` usage message (-h/--help) was updated in order to better document the functions of this program which can be used as a standalone utility for converting reads from BAM/SAM to FASTQ/FASTA; the -v/--version option was also added to this utility for easier integration in other pipelines.

» **TopHat 2.1.0 release 6/29/2015**
- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:
- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

» **TopHat source code moved to GitHub 3/31/2015**
TopHat is now available as a public GitHub repository where users are welcome to submit bug reports (issues) and developers are encouraged to submit patches (pull requests).

» **TopHat 2.0.14 release 3/24/2015**
Version 2.0.14 is a maintenance release with the following changes:
- pipeline speed improvements thanks to contributions from Vérorique Legrand and Michaël Pressigout of Institut Pasteur
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)

Site Map
Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol

News and updates
New releases and related tools will be announced through the Bowtie **mailing list.**

Getting Help
Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases
version 2.1.1 2/23/2016
Source code
Linux x86_64 binary
Mac OS X x86_64 binary

Related Tools
Cufflinks: Isoform assembly and quantitation for RNA-Seq
Bowtie: Ultrafast short read alignment
TopHat-Fusion: An algorithm for Discovery of Novel Fusion Transcripts
CummeRbund: Visualization of RNA-

TopHat2になりalignerとして
Bowtie2に対応
indelを考慮したアライメント
が可能になった　2012.4

Kim *et al. Genome Biology* 2013, **14**:R36
http://genomebiology.com/2013/14/4/R36

Genome Biology

METHOD Open Access

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

Daehwan Kim[1,2,3*], Geo Pertea[3], Cole Trapnell[5,6], Harold Pimentel[7], Ryan Kelley[8] and Steven L Salzberg[3,4]

Abstract

TopHat is a popular spliced aligner for RNA-sequence (RNA-seq) experiments. In this paper, we describe TopHat2, which incorporates many significant enhancements to TopHat. TopHat2 can align reads of various lengths produced by the latest sequencing technologies, while allowing for variable-length indels with respect to the reference genome. In addition to *de novo* spliced alignment, TopHat2 can align reads across fusion breaks, which can occur after genomic translocations. TopHat2 combines the ability to identify novel splice sites with direct mapping to known transcripts, producing sensitive and accurate alignments, even for highly repetitive genomes or in the presence of pseudogenes. TopHat2 is available at http://ccb.jhu.edu/software/tophat.

TopHat
A spliced read mapper for RNA-Seq

Johns Hopkins University Center for Computational Biology CCB

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

OSI certified

» **TopHat 2.1.1 release 2/23/2016**

Please note that TopHat has entered a low maintenance, low support stage as it is now largely superseded by **HISAT2** which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and **much more efficient** way.

Version 2.1.1 is a maintenance release which includes the following changes, some of them thanks to GitHub contributors:

- TopHat can be now built on more recent Linux distributions with newer GNU C++ (5.x), as the included SeqAn library was finally upgraded to a newer version.
- improved the detection of linker options for the Boost::Thread library which prevented the TopHat build from source on some systems.
- incorporated Luca Venturini's code to support large bowtie2 indexes (.bt2l).
- `bam2fastx` usage message (-h/--help) was updated in order to better document the functions of this program which can be used as a standalone utility for converting reads from BAM/SAM to FASTQ/FASTA; the -v/--version option was also added to this utility for easier integration in other pipelines.

» **TopHat 2.1.0 release 6/29/2015**

- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:

- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

» **TopHat source code moved to GitHub 3/31/2015**

TopHat is now available as a public GitHub repository where users are welcome to submit bug reports (issues) and developers are encouraged to submit patches (pull requests).

» **TopHat 2.0.14 release 3/24/2015**

Version 2.0.14 is a maintenance release with the following changes:

- pipeline speed improvements thanks to contributions from Véronique Legrand and Michaël Pressigout of Institut Pasteur
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)

Site Map

- Home
- Getting started
- Manual
- Index and annotation downloads
- FAQ
- Protocol

News and updates

Getting startedで、とりあえず使ってみる

Tools Users Google Group. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases

version 2.1.1 2/23/2016
- Source code
- Linux x86_64 binary
- Mac OS X x86_64 binary

Related Tools

Cufflinks: Isoform assembly and quantitation for RNA-Seq
Bowtie: Ultrafast short read alignment
TopHat-Fusion: An algorithm for Discovery of Novel Fusion Transcripts
CummeRbund: Visualization of RNA-

Getting started

- Install quick-start
- Test the installation
- Preparing your reference
- Preparing your reads
- Running TopHat
- Examining your results

Install quick-start

Download and extract the latest Bowtie 2 (or Bowtie) releases.

Note that you can use either Bowtie 2 (the default) or Bowtie (--bowtie1) and you will need the following Bowtie 2 (or Bowtie) programs in your PATH:

- `bowtie2 (or bowtie)`
- `bowtie2-build (or bowtie-build)`
- `bowtie2-inspect (or bowtie-inspect)`

Installing a pre-compiled binary release

In order to make it easy to install TopHat we provide a few binary packages to save users from the occasionally frustrating process of building TopHat themselves, which requires a certain development environment and the Boost libraries installed. To use the binary packages, simply download the appropriate one for your platform, unpack it, and make sure the `TopHat` binaries are in a directory in your PATH environment variable (or create a symbolic link to the included tophat2 script somewhere in your PATH, see below)

Note: if you want to be able to install and run this new version without overwriting a previous TopHat version already installed on your system, make sure you unpack the new version into a different directory from the old version, then instead of copying the new programs in a directory in your PATH just create a symbolic link from the `tophat2` wrapper script in this new directory to a directory in your shell's PATH. For example, assuming the ~/bin directory is in your PATH and you unpack tophat-2.0.0.Linux_x86_64.tar.gz under your home directory:

```
cd
tar xvfz tophat-2.0.0.Linux_x86_64.tar.gz
cd ~/bin
ln -s ~/tophat-2.0.0.Linux_x86_64/tophat2 .
```

Now you can start the new version of TopHat with the `tophat2` command, while the previous version, if present, can still be launched with the regular "tophat" command (assuming this is how you used it before).

Building TopHat from source

In order to build TopHat2 you must have the following installed on your system:

- the Boost C++ libraries (we recommend version 1.47 or higher so you can use it for building Cufflinks as well)

インストールの方法・
必要ツールなどの記載・
テストデータ等での極く簡単な
解析手順に関する記載がある

必要ツール
・bowtie2
・samtools

TopHat2はあらかじめコンパイルした
バイナリーファイルが配布されている
ので、自分でmakeする必要はない。
自分でソースからmakeする場合は
・SAMtools lib
・Boost C++ library
が必要

testデータが用意されている

```
tar zxvf test_data.tar.gz
cd test_data
tophat -r 20 test_ref reads_1.fq reads_2.fq
```

TopHat

A spliced read mapper for RNA-Seq

Johns Hopkins University Center for Computational Biology CCB

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort among Daehwan Kim and Steven Salzberg in the Center for Computational Biology at Johns Hopkins University, and Cole Trapnell in the Genome Sciences Department at the University of Washington. TopHat was originally developed by Cole Trapnell at the Center for Bioinformatics and Computational Biology at the University of Maryland, College Park.

OSI certified

TopHat 2.1.1 release 2/23/2016

Please note that TopHat has entered a low maintenance, low support stage as it is now largely superseded by **HISAT2** which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and **much more efficient** way.

Version 2.1.1 is a maintenance release which includes the following changes, some of them thanks to GitHub contributors:

- TopHat can be now built on more recent Linux distributions with newer GNU C++ (5.x), as the included SeqAn library was finally upgraded to a newer version.
- improved the detection of linker options for the Boost::Thread library which prevented the TopHat build from source on some systems.
- incorporated Luca Venturini's code to support large bowtie2 indexes (.bt2l).
- `bam2fastx` usage message (-h/--help) was updated in order to better document the functions of this program which can be used as a standalone utility for converting reads from BAM/SAM to FASTQ/FASTA; the -v/--version option was also added to this utility for easier integration in other pipelines.

TopHat 2.1.0 release 6/29/2015

- TopHat-Fusion algorithm improvements for more sensitive and accurate discovery of fusions, thanks to contributions from Gordon Bean and Ryan Kelley at Illumina.
  - This release implements a new algorithm for counting fusion-supporting read pairs that reduces the number of false-positive potential fusions. This algorithm computes the inner distance between read pairs by first converting the pair positions to transcript coordinates using the transcript information in refGene.txt and ensGene.txt. Pairs with small inner distance (suggesting the pair could come from a plausible pair-end insert) are counted as supporting evidence for the fusion. The default threshold for the inner distance is 250 base pairs; this parameter can be set using the `--fusion-pair-dist <int>` flag.
- fixed a few issues with GFF parsing of some annotation files
- fixed a runtime-error when using --no-discordant option.

Several fixes/improvements thanks to contributors on GitHub:

- new `--max-num-fusions` option allowing the user to specify the maximum number of reported fusions in tophat-fusion-post
- adjusting lower limit for `--fusion-multipairs`
- fixed a few typos, cleaning up python code etc.

TopHat source code moved to GitHub 3/31/2015

TopHat is now available as a public GitHub repository where users are welcome to subm ed to submit patches (pull requests).

TopHat 2.0.14 release 3/24/2015

Version 2.0.14 is a maintenance release with the following changes:

- pipeline speed improvements thanks to contributions from Véronique Legrand and M
- added support for xz compressed read files (thanks to a patch submitted by Ashton Trey Belew)

パラメータの意味など
詳しく知るためには、
必ずManualを見る

Site Map

- Home
- Getting started
- Manual
- Index and annotation downloads
- FAQ
- Protocol

News and updates

New releases and related tools will be announced through the Bowtie **mailing list**.

Getting Help

Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases

version 2.1.1 2/23/2016

- Source code
- Linux x86_64 binary
- Mac OS X x86_64 binary

Related Tools

Cufflinks: Isoform assembly and quantitation for RNA-Seq
Bowtie: Ultrafast short read alignment
TopHat-Fusion: An algorithm for Discovery of Novel Fusion Transcripts
CummeRbund: Visualization of RNA-

Manual

- What is TopHat?
- Prerequisites
- Using TopHat

・75base以上のreadに最適化
・リファレンス annotationなしでも
スプライスジャンクションを見つける

What is TopHat?

TopHat is a program that aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions. It is built on the ultrafast short read mapping program Bowtie. TopHat runs on **Linux** and **OS X**.

What types of reads can I use TopHat with?

TopHat was designed to work with reads produced by the Illumina Genome Analyzer, although users have been successful in using TopHat with reads from other technologies. In TopHat 1.1.0, we began supporting Applied Biosystems' Colorspace format. The software is optimized for reads 75bp or longer.

How does TopHat find junctions?

TopHat can find splice junctions without a reference annotation. By first mapping RNA-Seq reads to the genome, TopHat identifies potential exons, since many RNA-Seq reads will contiguously align to the genome. Using this initial mapping information, TopHat builds a database of possible splice junctions and then maps the reads against these junctions to confirm them.

Short read sequencing machines can currently produce reads 100bp or longer but many exons are shorter than this so they would be missed in the initial mapping. TopHat solves this problem mainly by splitting all input reads into smaller segments which are then mapped independently. The segment alignments are put back together in a final step of the program to produce the end-to-end read alignments.

TopHat generates its database of possible splice junctions from two sources of evidence. The first and strongest source of evidence for a splice junction is when two segments from the same read (for reads of at least 45bp) are mapped at a certain distance on the same genomic sequence or when an internal segment fails to map - again suggesting that such reads are spanning multiple exons. With this approach, "GT-AG", "GC-AG" and "AT-AC" introns will be found ab initio. The second source is pairings of "coverage islands", which are distinct regions of piled up reads in the initial mapping. Neighboring islands are often spliced together in the transcriptome, so TopHat looks for ways to join these with an intron. We only suggest users use this second option (--coverage-search) for short reads (< 45bp) and with a small number of reads (<= 10 million). This latter option will only report alignments across "GT-AG" introns

Illumina has provided the RNA-Seq user community with a set of genome sequence indexes (including Bowtie indexes) as well as GTF transcript annotation files. These files can be used with TopHat and Cufflinks to quickly perform expression analysis and gene discovery. The annotation files are augmented with the `tss_id` and `p_id` GTF attributes that Cufflinks needs to perform differential splicing, CDS output, and promoter user analysis. We recommend that you download your Bowtie indexes and annotation files from this page. More information about Illumina's iGenomes project can be found here.

Organism	Data source	Version	Size	Last Modified
Homo sapiens	Ensembl	GRCh37	17297 MB	May 14 17:23
	NCBI	build36.3	15814 MB	May 14 19:36
		build37.1	15850 MB	May 14 19:04
		build37.2	21450 MB	May 14 17:54
	UCSC	hg18	17349 MB	May 14 15:31
		hg19	21058 MB	May 14 15:36
Mus musculus	Ensembl	NCBIM37	14428 MB	May 14 22:13
	NCBI	build37.1	15260 MB	May 15 17:53
		build37.2	15725 MB	May 14 22:52
	UCSC	mm9	14537 MB	May 14 21:12
		mm10	14193 MB	Jun 14 11:29
Rattus norvegicus	Ensembl	RGSC3.4	13725 MB	May 15 22:33
	NCBI	RGSC_v3.4	14234 MB	May 15 23:58
	UCSC	rn4	13710 MB	May 15 22:32
Bos taurus	Ensembl	Btau_4.0	13315 MB	May 11 14:18
		UMD3.1	14042 MB	May 11 12:41
	NCBI	Btau_4.2	13357 MB	May 11 14:11
		Btau_4.6.1	13448 MB	May 11 16:09
		UMD_3.1	13990 MB	May 11 16:08

Site Map

- Home
- Getting started
- Manual
- Index and annotation downloads
- FAQ
- Protocol

News and updates

New releases and related tools will be announced through the Bowtie **mailing list**.

Getting Help

Questions and comments about TopHat can be posted on the **Tuxedo Tools Users Google Group**. Please use tophat.cufflinks@gmail.com for private communications only. Please do not email technical questions to TopHat contributors directly.

Releases

version 2.0.12 6/24/2014
- Source code
- Linux x86_64 binary
- Mac OS X x86_64 binary

Related Tools

Cufflinks: Isoform assembly and quantitation for RNA-Seq
Bowtie: Ultrafast short read alignment
TopHat-Fusion: An algorithm for

メジャーな生物種では
indexファイルやannotation
ファイル等が配布されて
いるので有効活用できる

NATURE PROTOCOLS | PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

Nature Protocols **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract • Accession codes • References • Author information

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

protocol論文も出ている

ただし今となっては少し古い

Freeではない

tophat基本コマンド

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

```
> tophat -G gene.gtf -o out_dir genome read_1.fastq read_2.fastq
```

-G/--GTF <GTF/GFF3 file>
まずgtfに基づき、トランスクリプトにmapさせ、ゲノム位置として戻す。
mapしないリードはゲノムから探す

tophatの出力

prep_reads.info
align_summary.txt
deletions.bed
insertions.bed
junctions.bed
accepted_hits.bam
unmapped.bam

sam/bam フォーマットのファイル
accepted_hits.bamファイルがこの後必要

実習3
tophatを用いて2D_1のfastqファイルをgenome_chr4にmapさせよ、
GTPファイルとしてgenes_chr4.gtfを用いる

例）
```
$ tophat -p 4 -G genes_chr4.gtf -o 2D_1 genome_chr4 2D_1_R1.fastq 2D_1_R2.fastq
```

出力を確認しよう。
例えば、align_summary.txtを見ればどの程度mapしたか分かる。
これでRNA-Seqのリード配列がゲノム配列にアラインできた。

cufflinksを用いてアラインされたreadを数える

定義した方法でのカウントが可能
gene単位
トランスクリプト単位
エキソン単位

- cufflinks
-BEDTools
-HTseq
が利用できる

今回はCufflinksを利用
そもそもTopHat → Cufflinksの解析系は同じ開発元、非常に良く使われている。

ローカスアノテーション情報を記載したgtfファイルを用意しておけば、
それに基づいて、genes単位、isoforms単位での解析を進めてくれる。

簡易的に、特定ローカスの解析などを進めたい場合や、
gtfファイルがない場合などは、BEDToolsも有用
gtfファイルを自分で作製するのは結構大変だが、bedファイルは比較的容易

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

http://cole-trapnell-lab.github.io/cufflinks/

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

Cufflinks was originally developed as part of a collaborative effort between the Laboratory for Mathematical and Computational Biology, led by Lior Pachter at UC Berkeley, Steven Salzberg's computational genomics group at the Institute of Genetic Medicine at Johns Hopkins University, and Barbara Wold's lab at Caltech. The project is now maintained by Cole Trapnell's lab at the University of Washington.

Cufflinks is provided under the OSI-approved Boost License

News

*To get the latest updates on the Cufflinks project and the rest of the "Tuxedo tools", please subscribe to our **mailing list***

Cufflinks has moved to GitHub	DECEMBER 10, 2014
Cufflinks 2.2.1 released	MAY 05, 2014
Cufflinks 2.2.0 released	MARCH 25, 2014
Cufflinks 2.1.1 released	APRIL 11, 2013

Protocol

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | **Contributions** | **Corresponding author**

Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621

Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

PDF | Citation | Reprints | Rights & permissions | Article metrics

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation[1, 2, 3]. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Cufflinks is available for Linux and Mac OS X. You can find the full list of releases below.

The Cufflinks source code for each point release is available below as well. If you want to grab the current code, check out the Cufflinks GitHub repository.

Star 68 Fork 48

Cufflinks Releases

Version	Date			
2.2.1	May 05, 2014	Linux	Mac OS X	Source
2.2.0	March 25, 2014	Linux	Mac OS X	Source
2.1.1	April 11, 2013	Linux	Mac OS X	Source
2.1.0	April 10, 2013	Linux	Mac OS X	Source

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

- Install quick-start
  - Installing a pre-compiled binary release
- Building Cufflinks from source
  - Installing Boost
  - Installing the SAM tools
  - Installing the Eigen libraries
  - Building Cufflinks
  - Testing the installation
- Common uses of the Cufflinks package
- Using pre-built annotation packages

自分でソースからmakeする場合は
・Samtools
・Boost C++ library
が必要

```
cufflinks ./test_data.sam
```

これでツールが動くことを確認

Install quick-start

Installing a pre-compiled binary release

In order to make it easy to install Cufflinks, we provide a few binary packages to save users from occasionally frustrating process of building Cufflinks, which requires that you install the Boost libraries. To use the binary packages, simply download the appropriate one for your machine, untar it, and make sure the cufflinks,cuffdiff and cuffcompare binaries are in a directory in your PATH environment variable.

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

Bowtie is provided under the OSI-approved Artistic License 2.0.

TopHat: alignment of short RNA-Seq reads

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is provided under the OSI-approved Artistic License 2.0.

CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.

CummeRbund is provided under the OSI-approved Artistic License 2.0.

Monocle: Differential expression for single-cell RNA-Seq and qPCR.

Monocle is a toolkit for analyzing single-cell gene expression experiments. Monocle was designed for RNA-Seq, but can also work with single cell qPCR. It performs differential expression analysis, and can find genes that differ between cell types or between cell states. When used to study an ongoing biological process such as cell differentiation, Monocle learns that process and places cells in order according to their progress through it. Monocle finds genes that are dynamically regulated during that process.

Monocle is provided under the OSI-approved Artistic License (version 2.0)

Cufflinksの関連ツール
Bowtie, TopHatは説明済み

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

The Cufflinks RNA-Seq workflow

The Cufflinks suite of tools can be used to perform a number of different types of analyses for RNA-Seq experiments. The Cufflinks suite includes a number of different programs that work together to perform these analyses. The complete workflow, performing all the types of analyses Cufflinks can execute, is summarized in the graph below. The left side illustrates the "classic" RNA-Seq workflow, which includes read mapping with TopHat, assembly with Cufflinks, and visualization and exploration of results with CummeRbund. A newer, more advanced worfklow was introduce with Cufflinks version 2.2.0, and is shown on the right. Both are still supported. You can read about the classic workflow in detail in our protocol paper.

Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression.

Cuffcompare

After assembling a transcriptome from one or more samples, you'll probably want to compare your assembly to known transcripts. Even if there is no "reference" transcriptome for the organism you're studying, you may want to compare the transcriptomes assembled from different RNA-Seq libraries. Cuffcompare helps you perform these comparisons and assess the quality of your assembly.

Cuffmerge

When you have multiple RNA-Seq libraries and you've assembled transcriptomes from each of them, we recommend that you merge these assemblies into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you've assembled. Cuffmerge performs this merge step.

Cuffquant

Quantifying gene and transcript expression in RNA-Seq samples can be computationally expensive. Cuffquant allows you to compute the gene and transcript expression profiles and save these profiles to files that you can analyze later with Cuffdiff or Cuffnorm. This can help you distribute your computational load over a cluster and is recommended for analyses involving more than a handful of libraries.

Cuffdiff

Comparing expression levels of genes and transcripts in RNA-Seq experiments is a hard problem. Cuffdiff is a highly accurate tool for performing these comparisons, and can tell you not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.

Cuffnorm

Sometimes, all you want to do is normalize the expression levels from a set of RNA-Seq libraries so that they're all on the same scale, facilitating downstream analyses such as clustering. Expression levels reported by Cufflinks in FPKM units are usually comparable between samples, but in certain situations, applying an extra level of normalization can remove sources of bias in the data. Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results you obtain with other downstream tools.

cufflink

cufflinks
cuffmerge
cuffcompare
cuffquant
cuffnorm
cuffdiff
の6つのプログラムから構成

cuffquant, cuffnormは
ver2.2.0(20140325)
から実装

MacOSX版のバイナリーはver2.2.0以降は
バグがありsegmentation errorでまともに
動かないようです。

今回の実習ではver2.1.1を使用し、
cuffquant, cuffnormは簡単な説明のみ
に留めます。

INSTALL MANUAL GETTING STARTED TOOLS HELP HOW IT WORKS PROTOCOL BENCHMARKS CODE FEED

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Cufflinks is an ongoing research project as well as a suite of tools. Here are the papers that describe the science behind the programs. If you use Cufflinks, **please cite these papers** in your work!

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Jeltje van Baren, Steven Salzberg, Barbara Wold, Lior Pachter.

Nature Biotechnology, 2010

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

doi:10.1038/nbt.1621

Note: *This is the original Cufflinks paper. Please cite this paper if you use Cufflinks in your work.*

Improving RNA-Seq expression estimates by correcting for fragment bias

Adam Roberts, Cole Trapnell, Julie Donaghey, John L. Rinn, Lior Pachter.

Genome Biology, 2011

どうやって動いているか

まず動いて使えそうな感じに
なったら詳細を把握していく

cufflinks基本コマンド

Cufflinksコマンド

cufflinks -o out_directory -G hoge.gtf tophat_directory/accepted_hits.bam

cufflinksを実行してパラメータを確認しよう。

考慮すべきパラメーター例

-o	出力の指定、TopHatの出力と同じ場所にしておくのが分かりやすいだろう
-p	CPUスレッド数の指定(デフォルトは1)、結構時間がかかるので使える数を指定すると良いだろう
-G	GTFファイルに記載されたアノテーションのみについて計算
-g	GTFファイルに記載されたアノテーションをガイドにしてアセンブルする
-M	無視したいトランスクリプト（rRNAなど）を指定

cufflinks出力

出力

skipped.gtf
transcripts.gtf
genes.fpkm_tracking
isoforms.fpkm_trancking

実習4
先のtophatの結果を用いてcufflinksにかけてみよう

例）
```
> cufflinks -p 4 -o 2D_1 -G genes_chr4.gtf accepted_hits.bam
```

出力を確認しよう。
geneごと、isoformごとにFPKM値が計算されているのが分かる。
-gを用いてcufflinksにかけると新規の発現領域が存在するのが分かる

cuffcompareコマンド

Cufflinks includes a program that you can use to help analyze the transfrags you assemble. The program cuffcompare helps you:

Compare your assembled transcripts to a reference annotation

Track Cufflinks transcripts across multiple experiments (e.g. across a time course)

From the command line, run cuffcompare as follows:

cuffcompare [options]* <cuff1.gtf> [cuff2.gtf] ... [cuffN.gtf]

今回はすでにあるgtfファイルの情報を用いるので、意識的に使う必要はない。

cuffmergeコマンドと出力

個々のサンプルのアセンブルモデルを統合する。

```
Usage:
    cuffmerge [Options] <assembly_GTF_list.txt>

Options:
    -h/--help                                Prints the help message and exits
    -o                     <output_dir>      Directory where merged assembly will be written  [ default: ./merged_asm  ]
    -g/--ref-gtf                             An optional "reference" annotation GTF.
    -s/--ref-sequence      <seq_dir>/<seq_fasta> Genomic DNA sequences for the reference.
    --min-isoform-fraction <0-1.0>           Discard isoforms with abundance below this       [ default:           0.05 ]
    -p/--num-threads       <int>             Use this many threads to merge assemblies.       [ default:              1 ]
    --keep-tmp
```

統合ファイルリストを事前に作製する必要がある(例 assemblies,txt)

cuffmerge -s $REFSEQ -g $GTF assemblies.txt

例 assemblies,txt

```
~/arabi_2D_2/transcripts.gtf
~/arabi_2D_3/transcripts.gtf
~/arabi_2D2L_2/transcripts.gtf
~/arabi_2D2L_3/transcripts.gtf
```

出力

merged.gtf

Cufflinks includes a script called cuffmerge that you can use to merge together several Cufflinks assemblies. It handles also handles running Cuffcompare for you, and automatically filters a number of transfrags that are probably artfifacts. If you have a reference GTF file available, you can provide it to the script in order to gracefully merge novel isoforms and known isoforms and maximize overall assembly quality. The main purpose of this script is to make it easier to make an assembly GTF file suitable for use with Cuffdiff.

cuffdiffコマンド

DE gene等を統計計算で取り出す
コマンド入力して使用法を確認してみよう

```
Usage:   cuffdiff [options] <transcripts.gtf> <sample1_hits.sam> <sample2_hits.sam> [... sampleN_hits.sam]
Supply replicate SAMs as comma separated lists for each condition:
sample1_rep1.sam,sample1_rep2.sam,...sample1_repM.sam
General Options:
  -o/--output-dir              write all output files to this directory          [ default:     ./ ]
  -L/--labels                  comma-separated list of condition labels
  --FDR                        False discovery rate used in testing             [ default:   0.05 ]
```

```
cuffdiff -o out_file merged.gtf bam1,bam2,bam3 bam4,bam5,bam6
```

Version 2.2.0以降は先のcuffquantで得られたcxbファイルをbamファイルの代わりに用いる。
cuffdiffにかかる時間やメモリー使用量が軽減される。

cuffdiffの出力

bias_params.info
run.info
read_groups.info
var_model.info
cds.read_group_tracking
cds.fpkm_tracking
cds.count_tracking
genes.read_group_tracking
genes.fpkm_tracking
genes.count_tracking
isoforms.read_group_tracking
isoforms.count_tracking
isoforms.fpkm_tracking
tss_groups.read_group_tracking
tss_groups.fpkm_tracking
tss_groups.count_tracking

gene_exp.diff
cds_exp.diff
cds.diff
isoform_exp.diff
promoters.diff
splicing.diff
tss_group_exp.diff

diffの付いたファイルがそれぞれの
違いの情報を記載したファイル

.diffファイルの内容

Column number	Column name	Example	Description
1	Tested id	XLOC_000001	A unique identifier describing the transcript, gene, primary transcript, or CDS being tested
2	gene	Lypla1	The gene_name(s) or gene_id(s) being tested
3	locus	chr1:4797771-4835363	Genomic coordinates for easy browsing to the genes or transcripts being tested.
4	sample 1	Liver	Label (or number if no labels provided) of the first sample being tested
5	sample 2	Brain	Label (or number if no labels provided) of the second sample being tested
6	Test status	NOTEST	Can be one of OK (test successful), NOTEST (not enough alignments for testing), LOWDATA (too complex or shallowly sequenced), HIDATA (too many fragments in locus), or FAIL, when an ill-conditioned covariance matrix or other numerical exception prevents testing.
7	$FPKM_x$	8.01089	FPKM of the gene in sample *x*
8	$FPKM_y$	8.551545	FPKM of the gene in sample *y*
9	log2 ($FPKM_y/FPKM_x$)	0.06531	The (base 2) log of the fold change *y*/*x*
10	test stat	0.860902	The value of the test statistic used to compute significance of the observed change in FPKM
11	p value	0.389292	The **uncorrected** *p*-value of the test statistic
12	q value	0.985216	The **FDR-adjusted** *p*-value of the test statistic
13	significant	no	Can be either "yes" or "no", depending on whether *p* is greater then the FDR **after** Benjamini-Hochberg correction for multiple-testing

cuffquantコマンドと出力(ver2.2.0以降)

bamの内容からgene/transcriptレベルで定量化し、バイナリー出力する

cuffquant -o out_directory hoge.gtf accepted_hits.bam

cuffquantを実行してパラメータを確認しよう。

考慮すべきパラメーター例
-o　　出力ディレクトリーの指定
-p　　CPUスレッド数の指定(デフォルトは1)、結構時間がかかるので使える数を指定
-M　　無視したいトランスクリプト（rRNAなど）を指定
他にもestimationに関わる　-b　-u パラメータがある。

出力
abundances.cxb

```
> cuffquant -p 4 -o 2D_1 genes_chr4.gtf accepted_hits.bam
```

新たにcxbファイルが作製されていることが分かる。
出力ファイルはこの１つだけ

cuffdiffの前にcuffquantを行い、cxbファイルを作製することで
cuffdiffを速くできる。

cuffnormコマンドと出力(ver2.2.0以降)

Cuffnormコマンド

Cuffnorm, which simply computes
a normalized table of expression values for genes and transcripts.

```
> cuffnorm -o out_file genes_chr4.gtf bam1,bam2,bam3 bam4,bam5,bam6
```

cuffnorm [options]* <transcripts.gtf>
<sample1_replicate1.sam[,...,sample1_replicateM.sam]>
<sample2_replicate1.sam[,...,sample2_replicateM.sam]>...
[sampleN.sam_replicate1.sam[,...,sample2_replicateM.sam]]

sam/bamかcxbファイルどちらも入力可能。ただし混在は不可

cuffnormの出力(ver2.2.0以降)

cds.attr_table
cds.count_table
cds.fpkm_table
cuffnorm.tree
genes.attr_table
genes.count_table
genes.fpkm_table
isoforms.attr_table
isoforms.count_table
isoforms.fpkm_table
run.info
samples.table
tss_groups.attr_table
tss_groups.count_table
tss_groups.fpkm_table

たくさんのサンプルで発現プロットやクラスター図を書きたい場合便利。

tophat -> cufflinksの解析系を使用する際の注意

It does not perform differential expression analysis. To assess the significance of changes in expression for genes and transcripts between conditions, use Cuffdiff. Cuffnorm's output files are useful when you have many samples and you simply want to cluster them or plot expression levels of genes important in your study.

Cuffnorm will report both FPKM values and **normalized**, estimates for the number of fragments that originate from each gene, transcript, TSS group, and CDS group. Note that because these counts are already normalized to account for differences in library size, they should not be used with downstream differential expression tools that require **raw** counts as input.

tophat -> cufflinksは一連の解析系

cufflinksの出力はすでにノーマライズされたものなので、rawデータを要求するedgeRなどの別のツールのinputには利用できない。

nature protocols

Home | Current issue | Archive | Authors & referees | About the journal

NATURE PROTOCOLS | PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber & Mark D Robinson

Affiliations | Contributions | Corresponding authors

Nature Protocols 8, 1765–1786 (2013) | doi:10.1038/nprot.2013.099
Published online 22 August 2013

Full text | PDF | Citation | Reprints | Rights & permissions | Article metrics

Abstract

Abstract • Accession codes • References • Author information • Supplementary information

RNA sequencing (RNA-seq) has been rapidly adopted for the profiling of transcriptomes in many areas of biology, including studies into gene regulation, development and disease. Of particular interest is the discovery of differentially expressed genes across different conditions (e.g., tissues, perturbations) while optionally adjusting for other systematic factors that affect the data-collection process. There are a number of subtle yet crucial aspects of these analyses, such as read counting, appropriate treatment of biological variability, quality control checks and appropriate setup of statistical modeling. Several variations have been presented in the literature, and there is a need for guidance on current best practices. This protocol presents a state-of-the-art computational and statistical RNA-seq differential expression analysis workflow largely based on the free open-source R language and Bioconductor software and, in particular, on two widely used tools, DESeq and edgeR. Hands-on time for typical small experiments (e.g., 4–10 samples) can be <1 h, with computation time <1 d using a standard desktop PC.

versionによる違いまとめ

Cutadaptの実習

1. Cutadaptを用いてpaired-endのtest data のアダプタートリミングをせよ。

paired endとしてのトリミングと、single readとしてのトリミングを試し、
両者のread数を比較してみよう。

tophat, cufflinksの実習

1. TopHatを用いて、paired-endのtest data（トリム済フォルダー下）
2D_rep1_R1.fastq,2D_rep1_R2.fastq
をリファレンスgenome_chr4にマップさせよ
オプション -Gの有無に
よる違いを確認しよう。

2.Cufflinksを用いて、
2D_rep1のカウントをしよう。
-Gと-gの違いを確認しよう。

結果をIGVで可視化してみよう

TAIR10の配列を呼び出し、TopHatで得られたBAMファイルを読み込む

Excelを使って結果を確認してみよう

gene_exp.diffファイルを読み込んでみる
tab区切りテキストファイルなのでそのまま読み込める
Excelのsort機能を使ってq値でsortしてみる

q値でsort

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_chan	test_stat	p_value	q_value	significant
XLOC_000047	XLOC_000047	KEA1	1:284609-291094	q1	q2	OK	12.8356	47.6879	1.89347	4.44122	5.00E-05	0.000325	yes
XLOC_000091	XLOC_000091	BXL2	1:564204-567769	q1	q2	OK	112.839	21.5634	-2.38762	-6.02938	5.00E-05	0.000325	yes
XLOC_000148	XLOC_000148	PSB27	1:898875-899655	q1	q2	OK	194.744	691.64	1.82844	7.10401	5.00E-05	0.000325	yes
XLOC_000310	XLOC_000310	PSBP-1	1:2047824-2049418	q1	q2	OK	588.195	3147.84	2.42	7.92975	5.00E-05	0.000325	yes
XLOC_000404	XLOC_000404	NPQ1	1:2706923-2709531	q1	q2	OK	21.2494	78.5734	1.88662	3.26377	5.00E-05	0.000325	yes
XLOC_000419	XLOC_000419	CSD1	1:2827060-2838469	q1	q2	OK	503.523	181.545	-1.47173	-5.38312	5.00E-05	0.000325	yes
XLOC_000450	XLOC_000450	CSP41B	1:3015327-3018234	q1	q2	OK	113.687	650.406	2.51627	8.83387	5.00E-05	0.000325	yes
XLOC_000487	XLOC_000487	LRR XI-23	1:3252239-3255693	q1	q2	OK	26.4081	49.6396	0.910512	2.30664	5.00E-05	0.000325	yes
XLOC_000598	XLOC_000598	ATGLX1	1:3995168-3997907	q1	q2	OK	60.1583	162.387	1.4326	3.26419	5.00E-05	0.000325	yes
XLOC_000600	XLOC_000600	AT1G11860	1:4001112-4003442	q1	q2	OK	319.6	756.582	1.24323	4.18318	5.00E-05	0.000325	yes
XLOC_000614	XLOC_000614	AT1G12080	1:4084161-4085045	q1	q2	OK	1884.29	67.9613	-4.79316	-9.20293	5.00E-05	0.000325	yes
XLOC_000616	XLOC_000616	CHL1-1	1:4105232-4109545	q1	q2	OK	107.267	57.7917	-0.892267	-2.70294	5.00E-05	0.000325	yes
XLOC_000624	XLOC_000624	AT1G12230	1:4147961-4151056	q1	q2	OK	102.049	50.9296	-1.00268	-2.40566	5.00E-05	0.000325	yes
XLOC_000680	XLOC_000680	CYP71B7	1:4467219-4469033	q1	q2	OK	17.1443	84.588	2.30272	4.53043	5.00E-05	0.000325	yes
XLOC_000724	XLOC_000724	AT1G13930	1:4761011-4762666	q1	q2	OK	94.6747	2483.48	4.71324	10.4968	5.00E-05	0.000325	yes
XLOC_000749	XLOC_000749	AT1G14345	1:4899144-4899979	q1	q2	OK	38.3992	157.145	2.03295	4.49341	5.00E-05	0.000325	yes
XLOC_000765	XLOC_000765	AT1G14670	1:5037611-5040528	q1	q2	OK	84.8105	44.439	-0.932415	-2.66978	5.00E-05	0.000325	yes
XLOC_000835	XLOC_000835	NDF1	1:5489297-5493772	q1	q2	OK	20.0548	104.567	2.3824	4.27443	5.00E-05	0.000325	yes
XLOC_000884	XLOC_000884	HCF173	1:5723087-5727312	q1	q2	OK	7.34039	112.227	3.93442	5.2414	5.00E-05	0.000325	yes
XLOC_000916	XLOC_000916	FUG1	1:5885082-5890470	q1	q2	OK	48.9638	105.457	1.10687	3.5512	5.00E-05	0.000325	yes
XLOC_001003	XLOC_001003	NDF6	1:6460597-6462224	q1	q2	OK	45.3045	185.555	2.03412	2.97075	5.00E-05	0.000325	yes
XLOC_001030	XLOC_001030	LHCA6	1:6612748-6613972	q1	q2	OK	52.6816	153.395	1.54188	4.09397	5.00E-05	0.000325	yes
XLOC_001063	XLOC_001063	PUP14	1:6832346-6833837	q1	q2	OK	37.731	91.5568	1.27892	3.13218	5.00E-05	0.000325	yes
XLOC_001076	XLOC_001076	ATLFNR2	1:6942716-6945018	q1	q2	OK	87.7487	1025.37	3.54662	10.0816	5.00E-05	0.000325	yes
XLOC_001099	XLOC_001099	AT1G20390	1:7065493-7071561	q1	q2	OK	45.6232	15.9769	-1.51378	-4.22277	5.00E-05	0.000325	yes
XLOC_001170	XLOC_001170	AT1G21680	1:7613004-7615339	q1	q2	OK	27.146	80.96	1.57647	3.93831	5.00E-05	0.000325	yes

GTFファイルに記載された遺伝子ごとの発現カウントに対して倍率、p値、q値が計算される。

RをつかってMA plotを書いて見よう

gene_exp.diffファイルを読み込んでみる
tab区切りテキストファイルなのでread.delim関数で読み込む
M, Aをそれぞれ計算する
plot関数を使って描画
colorのパラメータをsignifitureの値で色分けさせてみる。

例)

```
dat <- read.delim("gene_exp.diff")
A<-1/2*(log2(dat$value_1+1)+log2(dat$value_2+1))
M<-log2(dat$value_1+1)-log2(dat$value_2+1)
plot(A,M,col=dat$significant, pch=16, cex=0.4, ylim=c(-8,8))
```

簡易スクリプトを使って、結果を成形してみよう

Awkは便利な簡易スクリプト
　1行記述でもできる

例）
　q_valueが0.05以下のもののみリストアップするには？
　　q_valueの記載は13列目だから・・・

```
awk '$13<=0.05 {print $0}' gene_exp.diff
```

　と記述すればOK
　$で列番号を指定できる
　$0は行全体を意味する

その他
　grep, head, sort, cut, uniq等のUnixコマンドも活用しよう

スクリプトで連続的にcutadaptにかけてみよう

```
#!/bin/sh

UNIV_ADAPTER_COMP=AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT
IDX_CONS=AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
QV=30
O=7
MINCUT=50

for k in {2D_rep1,2D_rep2,2D_rep3,2D2L_rep1,2D2L_rep2,2D2L_rep3}

do

INPUT1=$k\_R1.fastq.gz
INPUT2=$k\_R2.fastq.gz

OUTPUT1=$INPUT1\.clnq_$QV\_$O\_$MINCUT.fastq.gz
OUTPUT2=$INPUT2\.clnq_$QV\_$O\_$MINCUT.fastq.gz

cutadapt \
-q $QV \
-O $O \
-a $IDX_CONS \
-A $UNIV_ADAPTER_COMP \
--minimum-length $MINCUT \
-o $OUTPUT1 \
-p $OUTPUT2 \
$INPUT1 \
$INPUT2

done
```

適切なパラメータ
例）
QV値
mincut値
O値
を設定追記して実行してみよう

実践演習課題

データセット
2D_rep1, 2D_rep2, 2D_rep3と2D2L_rep1, 2D2L_rep2, 2D2L_rep3で、
TopHat→Cufflinksの系を用いて、
2D（2days dark条件で育てた芽生え)
2D2L（その後2days light条件で育てた芽生え）
でのDE gene等を確認せよ。

GTFファイルとしてgenes_chr4.gtf
fastaファイルとしてgenome_chr4.fa
を利用する。
（アラビドプシスTAIR10の配列だが計算時間を考慮して、
それぞれChr4のみになっている）

マッピングデータのIGVでの可視化、
エクセルでの確認、
Rを用いたM-A plotの描画、
簡易スクリプトを用いたデータ抽出をせよ。
また解析をできる限り自動化できるよう、スクリプトを考えよ。

RNA-seq解析パイプライン：Transcript-based pipeline

Shuji Shigenobu
重信　秀治

基礎生物学研究所
生物機能解析センター

Two Basic Pipelines

- Choice of reference
  - **Genome** – standard for genome-known species
  - **Transcript** – the only way for genome-unknown species
    -- can be used for genome-known species

A Pipeline: Transcript-based

Mapping – alignment software

- For mapping reads onto transcript reference

  short read mapper (unspliced read aligner) is used

  - **Bowtie2**

    http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

bowtie2

Bowtie is an ultrafast, memory-efficient short read aligner.

http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

```
(example)

$ bowtie2 –x transcript.fa –U reads.fq –a –S out.sam
```

- Input
  - Reads (fastq) and reference (bowtie2-db)
- Output
  - Alignment in SAM format : **`out.sam`**

今回マッピングのセクションがこの前にあるので軽くで良い

Let's Try Bowtie2

Align 75-bp Illumina reads with a transcript reference using Bowtie2.

Prepare reads and reference genome

Sequences for this exercise are stored in **~/data/SS/**.

```
IlluminaReads1.fq – Illumina reads in fastq format
minimouse_mRNA.fa – a set of transcript sequnences
```

Build index of reference sequence

```
$bowtie2-build minimouse_mRNA.fa myref
```

Align reads with reference

```
$bowtie2 –x myref –U IlluminaReads1.fq –a –S out.sam
```

Count Reads by Transcript/gene

- The simplest way: just count reads by contig.

But…

- <u>Mapping ambiguity</u> should be taken into consideration.

Estimate Abundance

- **Multimapping issues**
  - Isoforms
  - Very similar paralogs
  - Repetitive sequences
  - => cannot align reads uniquely
- Mapping ambiguity should be taken into consideration.

- Critical for RNA-seq de novo analysis
- Software: RSEM and eXpress (EM algorithm)

eXpress

eXpress is a streaming tool for quantifying the abundances of a set of target sequences from sampled subsequences.

http://bio.math.berkeley.edu/eXpress/

```
(example)

$ express transcripts.fasta hits.bam
```

- Input
  - alignment (bam|sam) and reference (fasta)
- Output
  - Count estimation table: **`results.xprs`**

Let's Try eXpress

Prepare alignments and reference genome

Sequences for this exercise are stored in **~/data/SS/**.

```
IlluminaReads1.fq – Illumina reads in fastq format
out.sam – this file should be generated in the previous bowtie practice
```

Run eXpress

```
$ express minimouse_mRNA.fa out.sam
```

```
Output : results.xprs, params.xprs
```

Table of counts

data import

diagnostics

normalization

DE testing

evaluation

List of DE gene

edgeR

- A Bioconductor package for differential expression analysis of digital gene expression data
- **Model**: An over dispersed Poisson model, negative binomial (NB) model, is used
- **Normalization**: TMM method (trimmed mean of M values) to deal with composition effects
- **DE test**: exact test and generalized linear models (GLM)

edgeR

演習問題 ex5

- input: count data (not RPKM)
- output: gene table with DE significance statistics (FDR)

```
(example)

$ R
> library(edgeR)                      #load edgeR library
> dat <- read.delim("count_data.txt", …) #import count table to R
> group <- c(rep("M", 3), rep("H", 3))   #assign groups
> D <- DGEList(dat, group=group)      #import data to edgeR
> D <- calcNormFactors(D)             #normalization (TMM)
> D <- estimateCommonDisp(D)          #estimate common dispersion
> D <- estimateTagwiseDisp(D)         #estimate tagwise dispersion
> de <- exactTest(D, pair=c("M", "H")) #DE test
> topTags(de)
Comparison of groups: H-M
            logConc    logFC      P.Value          FDR
AT5G48430 -15.36821 6.255498 9.919041e-12 2.600872e-07
AT5G31702 -15.88641 5.662522 3.637593e-10 4.083773e-06
AT3G55150 -17.01537 5.870635 4.672331e-10 4.083773e-06
…
```

de novo RNA-seq

1. **Build** reference
2. **Characterize** reference

RNA-seq analysis pipeline (***de novo*** strategy)

de novo assemblers of RNA-seq

De novo assemblers use reads to assemble transcripts directly, which does not depend on a reference gnome.

- Trinity
- Oases
- TransAbyss
- EBARDenovo
- …

http://trinityrnaseq.sourceforge.net/

Trinity example

- Input: Illumina short reads in FASTQ | FASTA format
- Output: assembled contigs in FASTA format

```
# Run Trinity
$ Trinity --seqType fq --left left.fq --right right.fq \
          --CPU 8 --max_memory 20G
```

(Trinity is supported on only Linux)

RNA-seq analysis pipeline (*de novo* strategy)

optional

ORF prediction

- Special consideration in ORF prediction after de novo RNA-seq assembly
  - Sometimes partial: Start Met or terminal codon may be missing.
  - Ideally one ORF is present per contig, but erroneously joined contigs may include multiple ORFs.
  - Possible frame shifts.
    - Frame shifts do not occur so often in Illumina, while it happens very frequently in 454 and IonProton.
- Recommended software: TransDecoder

optional

Functional Annotation of Predicted ORFs

- BLAST
  - NCBI NR (or UniProt)
  - species of interest (model organisms, close relatives etc)
  - specific DB (SwissProt, rRNA DB, CEGMA etc)
  - self (assembly v.s. assembly)
- Motif search
  - Pfam, SignalP etc.
- Ortholog analysis
  - vs model organism
  - ortholog database (OrthoDB, eggNOG, OrthoMCL etc)
  - close relatives
- Gene Ontology term assignment

optional

Quick annotation by BLASTX

- Query: assembled contigs

  (nucleotide sequences in multi-fasta format)
- DB: Protein sequences of a model organism

Format DB

```
$ makeblastdb –in protein.fa -dbtype prot
```

Search

```
$ blastx -query trinity_contigs –db protein.fa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

optional

Let's try BLASTX

- Query: minimouse_mRNA.fa
- DB: human.protein.faa (human RefSeq protein)

1. Format DB

```
$ makeblastdb –in human.protein.faa -dbtype prot –parse_seqids
```

2. Search

```
$ blastx -query minimouse_mRNA.fa –db human.protein.faa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 0 > blastxout.txt
```

```
$ blastx -query minimouse_mRNA.fa –db human.protein.faa \
  -num_threads 8 -evalue 1.0e-8 –outfmt 7 > blastxout.tab
```

多変量解析

（ 特徴空間分割・次元圧縮）

慶應義塾大学　先端生命科学研究所
佐藤昌直

モチベーション:

多次元（例: 多パラメーター）をより少ない指標を使って理解する

↓

N個のサンプルをM個（M < N)のグループに分類する

→ 人間が新たな解釈を与える

解析の流れ

発現データ（生データ）

↓ ノーマライゼーション
前処理: 線形モデル

発現データ（バイアス除去）

↓

個々の遺伝子の解析 ⟶ 全体としての解析

有意差検定　　次元圧縮

下記のデータセットに含まれる数値を俯瞰してみましょう。 データの特徴を読み取れるでしょうか？

```
inputMatrix<- read.delim("~/data/MS/Sato_A_thaliana-P_syrin
gae_arvRpt2_6h_expRatio_small.txt", header=TRUE, row.name=1
)
head(inputMatrix)  #読み込みデータの一部を表示
image(t(inputMatrix))            #カラーコードによって可視化
heatmap(as.matrix(inputMatrix)) #階層クラスタリングで解析し、簡易
表示
```

この高次元（多パラメーター）の問題をどう扱うか？

1. クラスタリングによる分類

14 プロファイル
→ 3 クラスター

多変量解析のポイント

ポイント

教師有りか無しか
(supervised or unsupervised)？

どのような距離行列を使うか？

この高次元（多パラメーター）の問題をどう扱うか？

2. パラメーター数を減らす

多パラメーター → 3パラメーター（次元圧縮）

トランスクリプトームデータのある一部について可視化してみる

7次元の遺伝子発現データセット

7遺伝子の発現プロファイル間の類似性は7次元空間での距離によって決まる

コンピューターにどうデータを渡せばこの問題をどう扱えるか？

人間　**遺伝子発現プロファイル間のパターンの比較**

↓

問題の定義
(生物学の問題を数学の問題に置き換える)

コンピューター　**ある次元の空間におけるデータポイントの分布の比較**

距離の基準にするか？

距離尺度

ユークリッド距離

$$\left|\vec{A}-\vec{B}\right| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2}$$

相関係数
Pearson correlation coefficient

$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

Uncentered Pearson correlation coefficient = cosθ

$$\frac{\vec{A}\times\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|} = \frac{a_1b_1+a_2b_2+a_3b_3}{\sqrt{a_1^2+a_2^2+a_3^2}\sqrt{b_1^2+b_2^2+b_3^2}} = \cos\theta$$

ポイント

遺伝子発現プロファイルの**形**と**大きさ**

- **形**: ベクトルの方向
- **大きさ**: ベクトルのサイズ

どの距離係数を使うか？

- どんなプロファイルを同じプロファイルと定義するか？
- 距離係数計算の背後にあるものを意識して選択する。

Centering

距離係数計算の過程には

- **Centering**: 平均値をゼロにする
- **Normalization***: ベクトルの大きさを1にする

＊トランスクリプトームデータのnormalizationとは異なることに注意

これらはcentering後は
全く同じプロファイルになる

Normalization

これらはcentering, normalization後は
全く同じプロファイルになる

これらはnormalization後は
全く同じプロファイルになる

相関係数
Pearson correlation coefficient

$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

ポイント

多変量解析における注意点

方法依存的に抽出される特徴なので、
どんな特徴を認識したいのか注意が必要

Agglomerative hierarchical clustering

多変量解析の実際

階層クラスタリング

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering

Agglomerative hierarchical clustering

クラスター定義手法

Agglomerative hierarchical clustering

Average linkage

Complete linkage

階層クラスタリングの利点

- クラスター化してより少数のカテゴリーを示す
- 人間が認識可能なパターンを示す

Single linkage

階層クラスタリングの欠点

- Bottom-up: 非常に「手順」依存性
- 一つの距離のみを指標とした
  クラスタリング

「手順依存的」な方法の欠点を補うには？

- 偶然、観察されているクラスターを推定する
  → **同じ手順を繰り返す**
  → **クロスバリデーション**

多変量解析(I)のまとめ

教師有りか無しか

(supervised or unsupervised)？

- 事前情報、前提はあるか？
- ある場合はk-means法などの利用を検討

どのような距離行列を使うか？

- プロファイルの大きさ
- プロファイルの角度　など

クロスバリデーション

- あるクラスターは必然か偶然か？
- leave-one out validation:サンプルを一つ抜いてクラスタリングしてみる
- 少数の特定遺伝子がクラスタリングに影響していないか？
- Bootstrap: 遺伝子サブセットでクラスリングを繰り返してみる

主成分分析

主成分分析とは？

モチベーション:

多数の遺伝子で構成される多次元データ（サンプル）の中で相関のある遺伝子群を使って**新たな軸**を作り、データを見直す

→ **人間が新たな解釈を与える**

PCAは何をするのか？

階層クラスタリング、k-means法:
プロファイル間の類似性は
空間での**1つの距離**によって決まる

PCAは何をするのか？

PCAは何をするのか？

PCAで得られる重要な統計量

- 寄与率
- 因子負荷量
- 主成分得点

PCAの概略(2次元)

1. 各サンプル（1..n）の観察値(x_n, y_n)を

$$\boxed{\begin{aligned} u_n &= a_1 x_n + b_1 y_n \\ v_n &= a_2 x_n + b_2 y_n \end{aligned}}$$

とおく

2. $a^2 + b^2 = 1$, uとvの相関係数0という制約の下でこれを解いて a_n, b_n を求める。

寄与率

- 各主成分が説明する分散の割合

負荷量 loadings

- 得られた主成分と元データのパラメーターの相関
- 各パラメーターがもとのデータの情報をどれだけ有するか

主成分分析(まとめ)

- 主成分分析は**データの分散を説明する新たな軸を計算する**方法
  - **寄与率**
  - **因子負荷量**
  - **主成分得点**

主成分得点 scores

- 各パラメーターの値を各主成分について標準化したもの

標準化: 平均0, SD=1

注意点

1.デフォルトのprincompでは返り値loadingsは因子負荷量ではない。

2. 相関を使うか、分散共分散行列を使うか

多次元尺度構成法

Multi-dimensional scaling(MDS),
Principle coordinate analysis

MDSは何をするのか？

多次元尺度構成法とは？

モチベーション(PCAと同様)：

多数の遺伝子で構成される多次元データ（サンプル）の中で各サンプル間の違いを表現する座標を作る

→ 人間が新たな解釈を与える

サンプル間の距離をまず計算する

この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ

この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ

点kの位置が意味を持つことはないのか？

この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ

$$d_{ij}^2 = d_{ik}^2 + d_{jk}^2 - 2d_{ik}d_{jk}\cos\theta$$

例:入力データが野生型・変異体プロファイルの比であったら？

多変量解析(2)のまとめ

PCA/MDS

- データをそれがもつ次元に分解して評価・可視化する
- **重心の置き方に違い**: 入力データをどのように前処理するか

多変量解析をもう一歩進めて:

人間の解釈をアシストするデータ取得を心がける

多変量解析の枠組み

モチベーション:

多次元(例: 多パラメーター)をより少ない指標を使って理解する

↓

N個のサンプルをM個(M < N)のグループに分類する

→ **人間が新たな解釈を与える**

コントロール、指標サンプルは含められるか？

研究の目的、実験デザイン、多変量解析

目的	実験デザイン	多変量解析
•何を知りたいか明確に	•**線形モデル**	•入力データ前処理
•実施の制約	•比較、因子	•距離尺度
•予算	•検出力	•アルゴリズム
•時間、労力		

今回のトレーニングコースで扱わなかった重要項目

- 確率分布
- 回帰、相関
- 線形モデルにおける交互作用
- 非線形クラスタリング・次元圧縮
  - self-organization mapなど