ゲノムインフォマティクストレーニングコース

NGS解析入門
コース概要

基礎生物学研究所

情報管理解析室

内山　郁夫

ゲノムインフォマティクストレーニングコース
NGS解析入門　スケジュール

2月9日（水）

09:00-09:30	オンライン受付	
09:30-10:05	演習環境の構築	
10:05-10:45	コース概要	[内山]
10:45-12:00	UNIX基本コマンド（前編）	[西出]
12:00-13:00	（昼休憩／自己紹介タイム）	
13:00-14:30	UNIX基本コマンド（後編）	[西出]
14:30-14:45	（休憩）	
14:45-17:00	R入門	[内山]
17:15-19:00	統計学入門	[佐藤]
19:00-	オンライン懇親会（参加自由）	

2月10日（木）

09:00-10:00	NGS基本データフォーマット	[杉浦]
10:00-10:30	クオリティコントロールとNGS基本ツール	[山口]
10:30-10:40	（休憩）	
10:40-12:00	クオリティコントロールとNGS基本ツール（続き）	[山口]
12:00-13:00	（昼休憩）	
13:00-14:00	エディタとスクリプト	[杉浦]
14:00-15:00	UNIXによるテキストファイル処理	[中村]
15:00-17:00	演習	

講師

- **生物機能解析センター・情報管理解析室**
  - 内山郁夫　准教授（本コースオーガナイザー）
  - 西出浩世　技術職員
  - 中村貴宣　技術職員
  - 杉浦宏樹　技術職員
- **生物機能解析センター・生物機能情報分析室**
  - 重信秀治　教授（RNA-Seq入門オーガナイザー）
  - 山口勝司　技術職員
- **北海道大学大学院農学研究院**
  - 佐藤昌直　助教

次世代シーケンサ
Next Generation Sequencer (NGS)

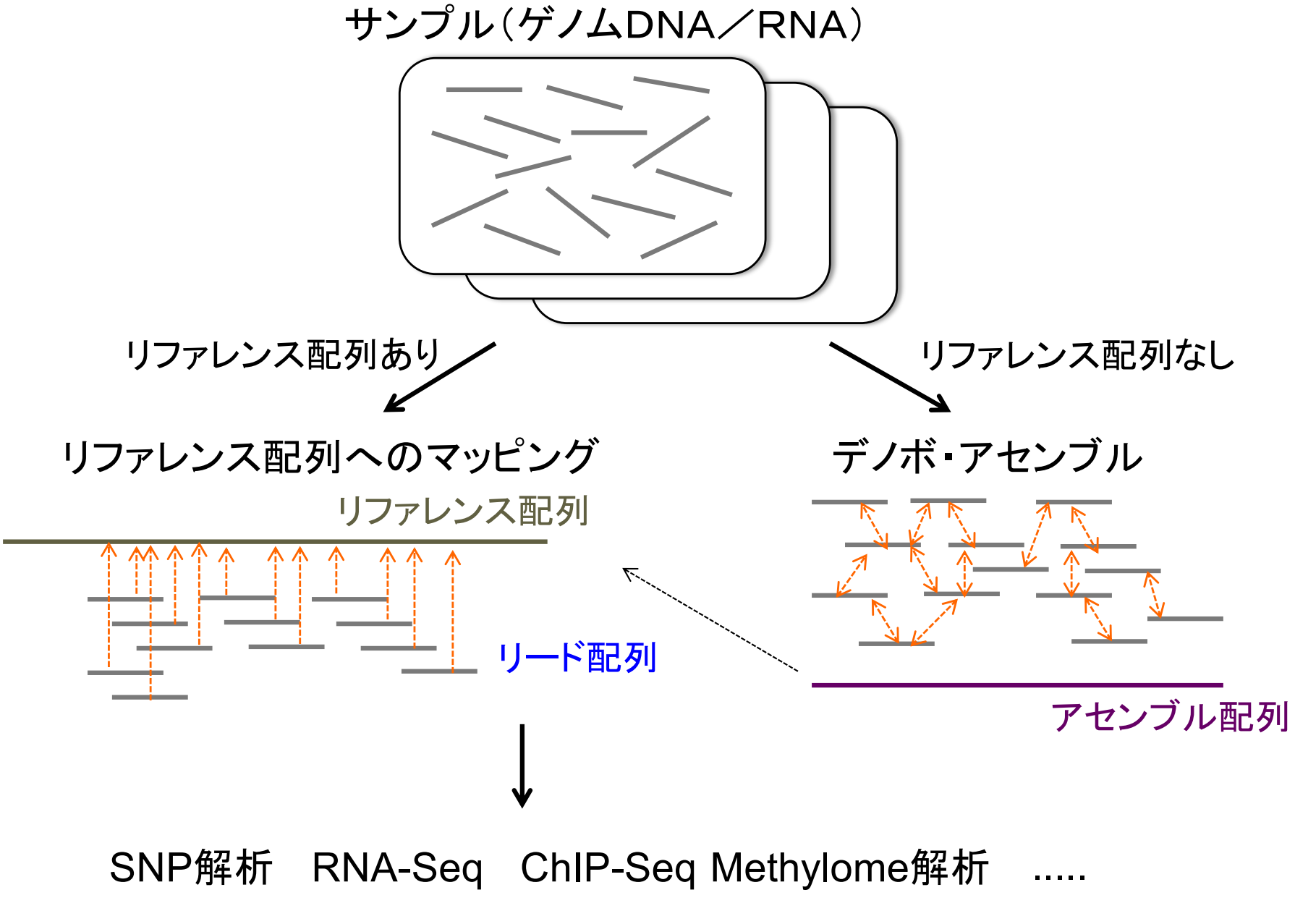
100 1,000 10,000 bp

（旧世代シーケンサ）
サンガーシーケンサ
(500~800 bp)

ショートリード シーケンサ
Roche 454
Illumina MiSeq/HiSeq
SOLiD
Ion Torrent

ロングリード シーケンサ
PacBio
Oxford Nanopore

次世代シーケンサデータ処理の概要



ちょっとやってみよう

「ターミナル」からサーバにログインした状態で、以下のコマンドを順にタイプしてみよう

```
$ cd data/0_intro
      (ディレクトリの移動)
$ ls
      (ファイルの表示)
$ bowtie2 -x ecoli_genome -U eco.fastq -S ecoli.sam
      (NGSリード配列(eco.fastq)をゲノム配列上にマッピング)
$ htseq-count ecoli.sam ecoli.gtf > ecoli.count
      (マッピングした結果を使って遺伝子ごとにリード数をカウント)
$ head ecoli.count
      (結果ファイル ecoli.count の先頭10行を表示)
```

データ処理の流れ

リファレンス配列
ecoli_genome.fasta

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

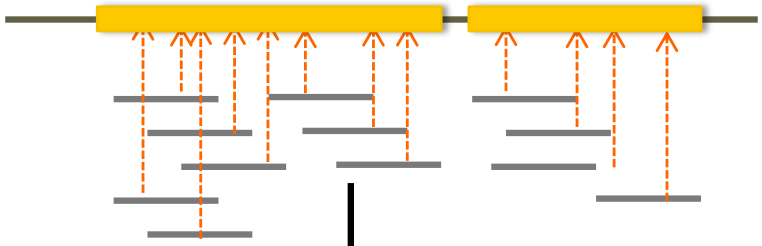
（インデックス：ecoli_genome）

リード配列 eco.fastq

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFFDFDFHDFFHIIIEGIHJJJJGFHGGHGGHGGIIJDGIJHHGGGHIH
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFFFDFAFHFHIJGHIJIJJIJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

① bowtie2

リファレンス配列へのマッピング



マッピング結果 ecoli.sam

```
@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
```

遺伝子アノテーション ecoli.gtf

```
chr   RefSeq   start_codon   190   192   1.000   +   .   gene_id "b0001"; transcript_
chr   RefSeq   CDS           190   252   1.000   +   0   gene_id "b0001"; transcript_
chr   RefSeq   stop_codon    253   255   1.000   +   .   gene_id "b0001"; transcript_
chr   RefSeq   exon          190   255   1.000   +   .   gene_id "b0001"; transcript_
```

② htseq-count

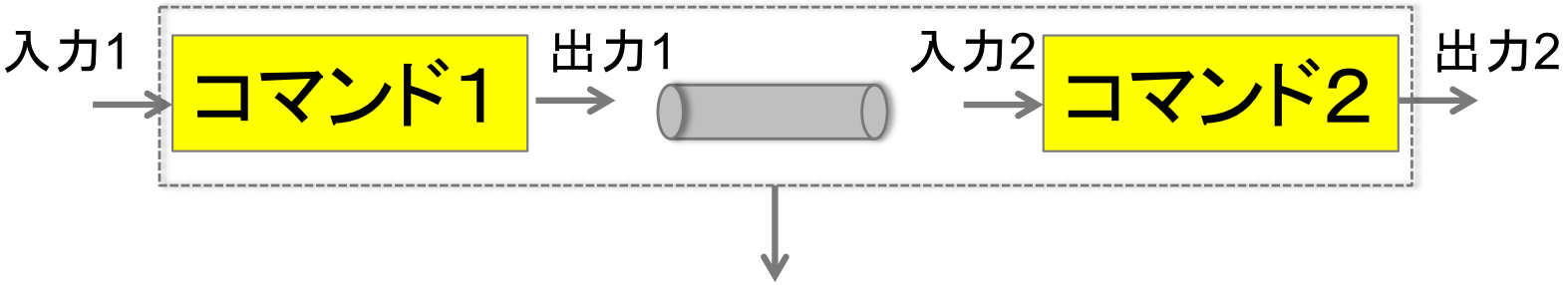
遺伝子ごとの集計

集計結果 ecoli.count

```
b0001   11
b0002   117
b0003   33
b0004   44
```

複数のコマンド(プログラム)を組み合わせた複雑な処理の実行

コマンドのパイプライン



スクリプト:コマンドS

```
コマンド1
コマンド2
```

スクリプトによる実行

テキストデータ

リファレンス配列 ecoli_genome.fasta

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

リード配列 eco.fastq

```
@SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
ATCCGGCTGGCGCACCGACCTATGTTCCGGGCGAATACAAGCTGGGTGAAG
+SRR1515276.1 HWI-ST808:151:D2D13ACXX:2:1207:3625:88631 length=51
@@@AD>DDFF7DC?FFEBF@DFII<DF@AAA6AEFBDBDCA?>A?B=>B::
@SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CACCGTGTAGTACCAGCATCCTGCGTACAATCAGCAATCCCAGTCCTCCCC
+SRR1515276.2 HWI-ST808:151:D2D13ACXX:2:1207:3871:88513 length=51
CCCFFDFDFHDFFHIIIEGIHJJJJGFHGGHGGHGGIIJDGIJHHGGGHIH
@SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CAGGACATCGCCTTTGATCGGTTCAGACTTCGGACCAACCTGCATTTTCAG
+SRR1515276.3 HWI-ST808:151:D2D13ACXX:2:1207:3950:88530 length=51
CCCFFFDFAFHFHIJGHIJIJJJIJJHEHIIJGHIFEHIIA@FIFHGGIIGI
```

遺伝子アノテーション ecoli.gtf

```
chr  RefSeq  start_codon  190  192  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  CDS          190  252  1.000  +  0  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  stop_codon   253  255  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
chr  RefSeq  exon         190  255  1.000  +  .  gene_id "b0001"; transcript_id "b0001";
```

マッピング結果 ecoli.sam

```
@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
SRR1515276.434  0    chr 4198737 42    51M    *    0       0       GCGCGGTACGCATCTGG
```

集計結果 ecoli.count

```
b0001   11
b0002   117
b0003   33¥
b0004   44
```

発現量データ(表形式のデータ)の解析

表データ

	条件1	条件2	条件3	条件4
遺伝子1	58.3	161.9	24.3	46.3
遺伝子2	1061.9	1073.9	106.9	222.9
遺伝子3	236.0	207.9	153.4	116.1
遺伝子4	16.2	38.3	0.0	0.0

条件1　(58.3, 1061.9, 236.0, 16.2, ...)

条件2　(161.9, 1073.9, 207.9, 38.3, ...)

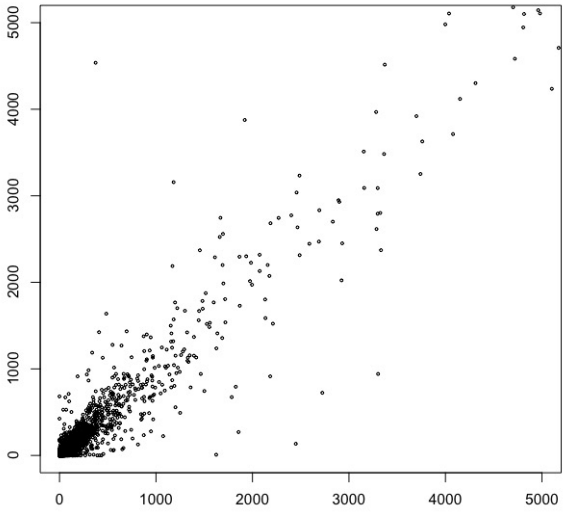
データ解析、統計解析

条件1と条件2の発現量比

$$\left(\frac{58.3}{161.9} \quad \frac{1061.9}{1073.9} \quad \frac{236.0}{207.9} \quad \frac{16.2}{38.3} \right)$$

データ可視化

散布図
(scatter plot)



ゲノムインフォマティクストレーニングコース
NGS解析入門　スケジュール

2月9日（水）

09:00-09:30	オンライン受付	
09:30-10:05	演習環境の構築	
10:05-10:45	コース概要	[内山]
10:45-12:00	UNIX基本コマンド（前編）	[西出]
12:00-13:00	（昼休憩／自己紹介タイム）	
13:00-14:30	UNIX基本コマンド（後編）	[西出]
14:30-14:45	（休憩）	
14:45-17:00	R入門	[内山]
17:15-19:00	統計学入門	[佐藤]
19:00-	オンライン懇親会（参加自由）	

2月10日（木）

09:00-10:00	NGS基本データフォーマット	[杉浦]
10:00-10:30	クオリティコントロールとNGS基本ツール	[山口]
10:30-10:40	（休憩）	
10:40-12:00	クオリティコントロールとNGS基本ツール（続き）	[山口]
12:00-13:00	（昼休憩）	
13:00-14:00	エディタとスクリプト	[杉浦]
14:00-15:00	UNIXによるテキストファイル処理	[中村]
15:00-17:00	演習	

準備編を通しての目標

- **インフォマティクスに対する心的障壁を取り除く**
- **ゲノムインフォマティクスの基礎的技術と考え方を身に付ける**
  - UNIXコマンドラインの操作や環境に慣れる
  - 統計的な考え方やデータ処理の流れを理解する
  - NGSデータの基本的な見方、扱い方に習熟する
  - タブ区切りテキストを処理する程度の簡単なプログラミングを学ぶきっかけをつかむ
- **独習するための基盤を身に付ける**
  - 今後独習する為に必要な基礎的なスキル
  - 今後何を学べば良いかの指針を得る
- **インフォマティクス専門家と対話できる程度の基礎知識を身に付ける**

オススメ勉強法

- コマンドやプログラムは自分で試してみる。copy & pasteでなくタイピングすること。（熊楠メソッド）
- 気軽に質問する。講師はもちろん、隣や前後の受講生にも。その一方で、ヘルプやマニュアルドキュメントをうまく活用する。
- 自分の研究との接点を常に意識する。自分の研究に応用する。

コースページ

https://github.com/nibb-unix/gitc202202-unix/wiki

質問について

- **講義内容に関する講師への質問**
  - Zoomの「手を挙げる」機能を使うか、Zoomのチャットに書き込んでください。
  - 受講生、聴講生のいずれも質問していただけます。
- **実習中のトラブルなど、個別対応が必要な質問**
  - Slack の講義質問チャンネルに書き込んでください。フロアサポートスタッフが対応します。
  - Slackで質問できるのは受講生のみですが、聴講生も閲覧は可能です。
  - よくある質問とその回答は、コースページのFAQの項目にも記載しますので、適宜参照してください。

それでは始めましょう