Introduction to key concepts in Illumina sequencing data analysis

- イルミナシーケンスデータ解析入門その前に

癸生川絵里 (Eri Kibukawa)
Bioinformatics Support Scientist

illumina®

Agenda

- イルミナシーケンシング計画
  - 検討事項
- データ解析時の背景コンセプト
  - 主要な解析タイプ
  - 目的と前提

シーケンスデータ: リード

- FASTQ形式ファイルとしてまとめられる
- FASTQ: 配列とクオリティースコアが含まれるファイル
- 500万 ～ 30億リードの情報が1回のシーケンシングランで得られる
- FASTQファイルの例；

```
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13501:2240 1:N:0:CTTGTA
TGAAACCAGTGTTCTTAATTGGCATTTTACACACACACACACAGAATTTAAAAAAAAAAATCAAAGGAAATCATTCTAAATGTACTATGATAGCATGTTAAA
+
=55>7;?::BDADDD@EE88DCD?DFFEFFECBE6666BB=B;<;<-34:;<CB51>=BBEE>EE?3D@??CB->:=:AA8DDDDDDBBE9;,=?:/89<E
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13660:2247 1:N:0:CTTGTA
CCAAACATTAAGTAACTCTTAAAATGGCACACAGGTTTTAAAGCTATTGGTTTTTCCTTCCTAACTCTCTGAATTTTTCCCCTGGCCTTTGTAGATCAACT
+
FFEDFBGEGGGGDFGEFFFFGGDF=FBFFFGGGE7CEEDEFBFBFGEEGF@FCDDFDFFEGFEAGFGGGGD.;DDGG@FGE.EBEFGFGFCEFEBEFDEF8
@HWI-BRUNOP20X:994:B809UWABXX:1:1101:13966:2183 1:N:0:CTTGTA
TTGGGTAACTTGAATATAACATGGCTCCCTTGCTGTAAGCAAATGTTTTAGAGCTGAATTTTTCCTTTTTTTTTTTTTTTTTTTAAGCCAAGAAGTTCACC
+
HHHHHEHHHHHHFHHHHHHHHHHHHHHHHHHHHGGFHHHHHHHHHHFHHHFHEHHFHEHHHHFHHHFHHHEHEHHHHHHHHHH@?################
```

シーケンシング計画

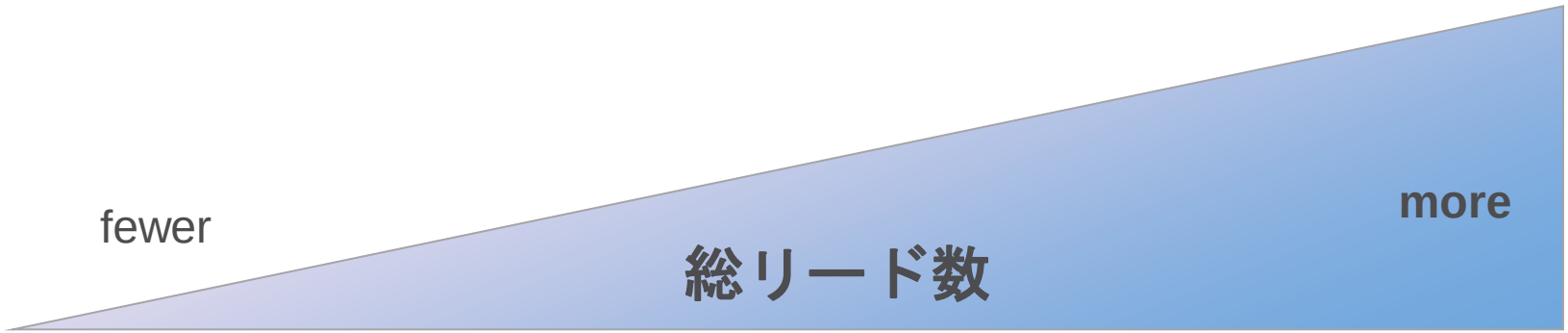
- リード数？
- ペアード (PE) かシングル (SR)か？
- リード長?

- これらはあくまでガイドラインが存在するのみ

- その時々の科学コミュニティーにより基準、標準が提示されるといえる

どの程度のデータ量を設定するのか？

- 実施したいアプリケーション、必要な検出感度、ゲノムサイズ等により異なる



ゲノムサイズ	**小**		**大**
アプリケーション	**RNA発現解析**	Whole genome Resequencing	De Novo **アセンブリ**
検出感度	**低**		**高**

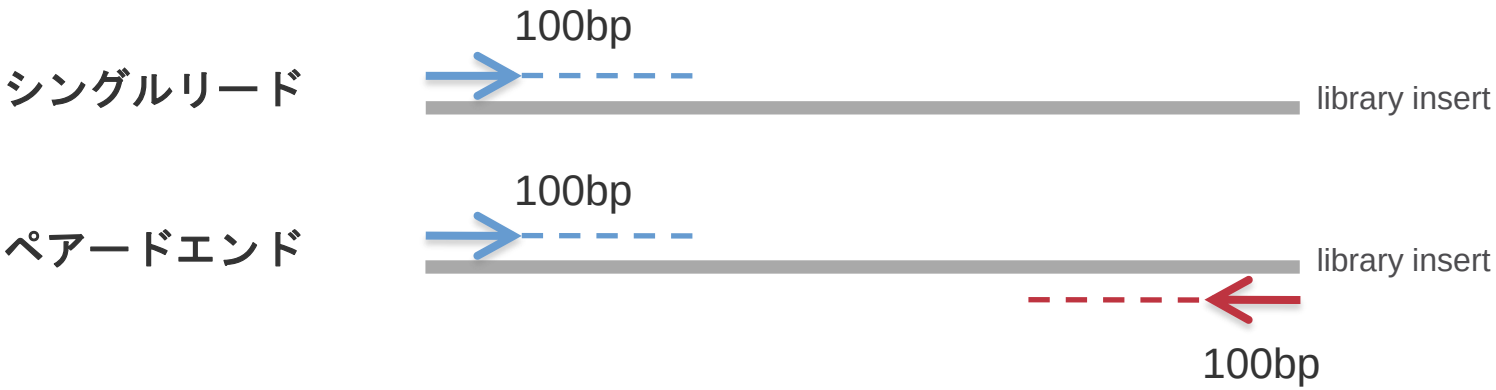
どの程度のデータ量を設定するのか？

(例)

アプリケーション	生物種	ゲノムサイズ	カバレッジ	必要データ量
リシーケンシング	Human	3.4 Gbp	20x	68 Gbp
De Novo アセンブル	Human	3.4 Gbp	70x	240 Gbp
リシーケンシング	Arabidopsis	0.125 Gbp	20x	2.5 Gbp
De Novo アセンブル	Arabidopsis	0.125 Gbp	60x	7.5 Gbp
RNA-Seq (遺伝子発現)	Human	N/A	必要検出感度による	5-100 M リード /サンプル

ペアードで実施するかシングルで実施するか？

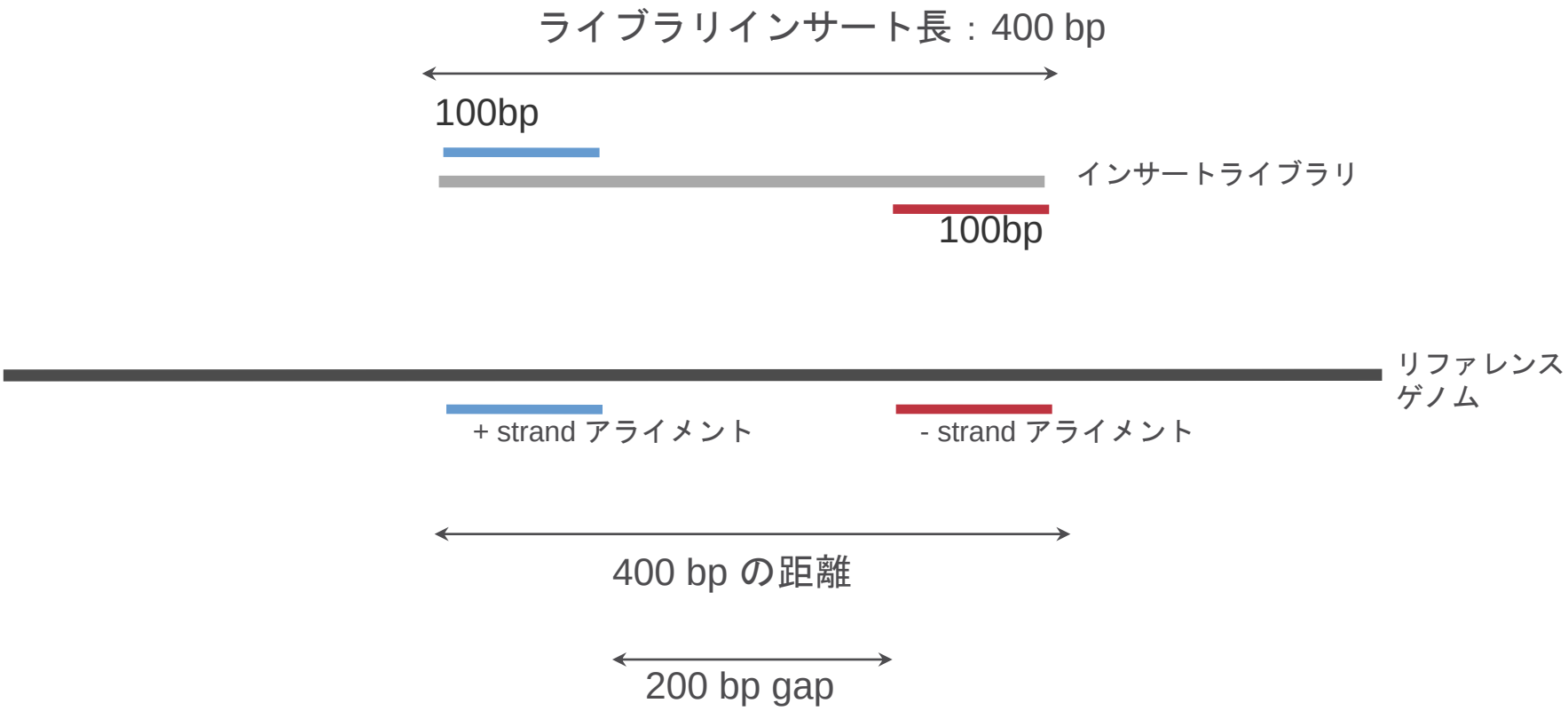
- ペアエンド（PE）やシングルリード（SR）



- ライブラリのインサートサイズから、アライメント後の PE のリード間距離がどの程度であるか予め分かり、マップ結果と比較し変異検出等に利用することができる。

ペアードで実施するかシングルで実施するか？

- ペアエンドはアライメント結果にさらなる情報を加えることができる



ペアードで実施するかシングルで実施するか？

Application	PE or SR?	Note
SNP 検出 (リシーケンシング)	SR または PE	coverage depthがキー
Indel, 構造変異検出 (リシーケンシング)	PE	PE を前提とした検出解析方法のため
De Novo ゲノムor トランスクリプトーム アセンブル	PE	アセンブルの際にPE情報が利用される
RNA-Seq (発現)	PE (あるいはSR)	新規転写産物、遺伝子構造を決めるためにPE情報が必要

リード長はどのように考えたらいいか？

small RNA

RNA

RNA

Indel と
構造変異検出

リシーケンシング

De Novo アセンブル
(ゲノムあるいはトランスクリプトーム）

遺伝子 構造決定
(ゲノム アノテーション)

Coverage depth

- ゲノム上の位置あたりにマップされたリード数の平均
- この図のカバレッジは？

Coverage depth 計算

- ゲノム上の位置あたりにマップされたリード数の平均
- この図のカバレッジは？　- 4.5x

(例) サイズ 0.1 Gbp のゲノムで考えたとき、カバレッジ 30x を得るには
リード長100 bp で何リード必要と試算できるか？

Coverage depth 計算

(問) サイズ 0.1 Gbp のゲノムで考えたとき、カバレッジ 30x を得るには
リード長100 bp で何リード必要と試算できるか？

(解)　30 x 0.1Gbp = 3Gbpのデータ量（塩基数）が必要

3G bp / 100 bp reads = 30Kリード　=　3万リード

＊ 弊社サイトMyIlluminaにて “**Coverage Calculation Tech Note**” で検索していただきますと、
より詳細な説明が記載された、テクニカルノートをダウンロードいただけます。

Coverage depth について考える

- カバレッジが大きければシーケンスされたサンプル配列の信頼性はあがる
- 例えば カバレッジ= １ のとき；

  これはSNPなのか？ シーケンシングエラーなのか？

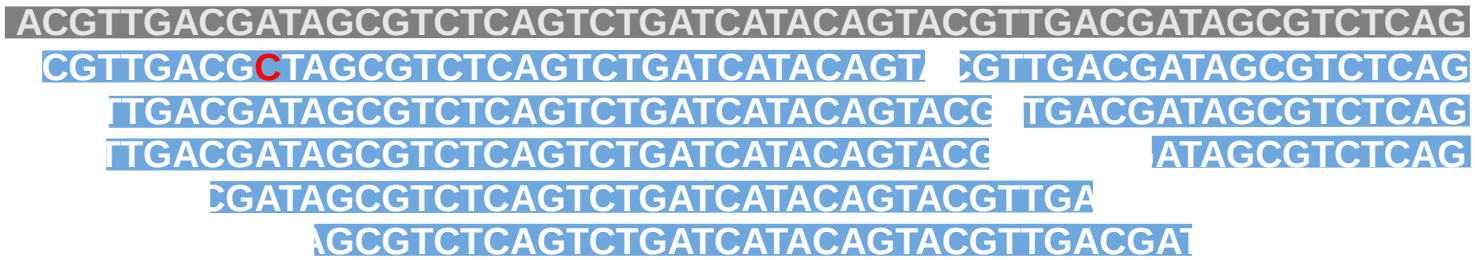
  リードが本来の位置でないところにマップされてしまっているのか？

Coverage depth について考える

- カバレッジが大きければシーケンスされたサンプル配列の信頼性はあがる

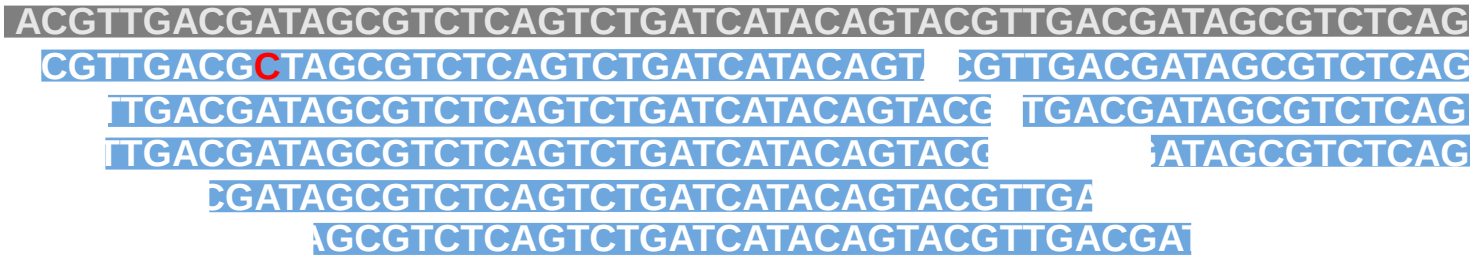
例えばSNP検出

- ほとんどの SNP検出器はこのような違いをSNPとしては検出しない
- しかしながらもしこれらのリードサンプルが癌組織由来である場合は、
  SNPである可能性は増加



例えばSNP検出

- 一般的なSNP検出器の前提;
  - 全てのサンプルはディプロイド（二倍体）であると仮定
  - サンプルは最高でも 2 アリルまでであると仮定
  - allelic ratio は 50-50 程度であると仮定
    - データ解析に使用するソフトが当該のサンプルタイプと実験タイプに適したものを選ぶ必要がある。



Key concepts in bioinformatic analysis

イルミナデータ解析の主要２タイプ

リファレンス シーケンス を
使用

↓

アライメント(マッピング)、
カウンティング

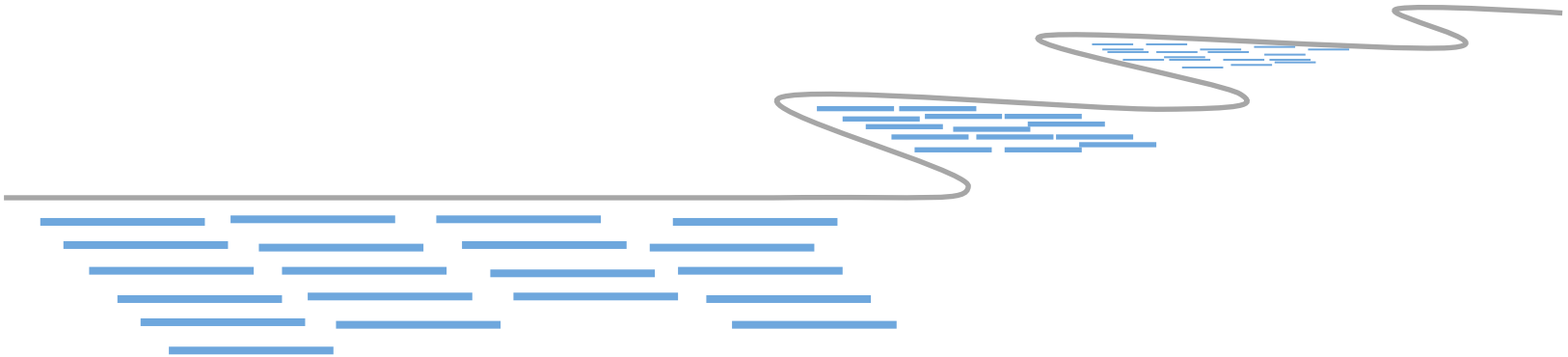
リードをつなげてできるだけ正確に
長くする
(ゲノム or トランスクリプトーム)

↓

De Novo アセンブリ

アライメントやリシーケンシング アプリケーション

- 通常 genomic DNA サンプル （ full genome, enriched, amplicon etc.)
- リードをリファレンスゲノム配列に対してアライメント(マップ)
- リファレンスとリードで異なる箇所を検出



ショートリードマッピングには多くのコンピュータ資源が必要

- もともとリードがあった、ゲノムなど長配列上の位置を探し出す処理
  - 最もリード配列にマッチしたリファレンス配列上の位置にアラインすることになる

- アライメントプログラムは以下を扱わねばならない
  - 多数のリード配列に対し1つの長いリファレンス（ターゲット）配列
  - マルチプルヒット
  - リードとゲノムとのミスマッチ

ショートリードマッピングには多くのコンピュータ資源が必要

- An illustration

Aligners_might_place_a_short_sequence_in_many_places_in_the_genome_sequence

in (2)

en (3)

リードが短いと多くの位置にヒットしてしまい位置決めが難しい

ショートリードマッピングには多くのコンピュータ資源が必要

- An illustration

Aligners_might_place_a_short_sequence_in_many_places_in_the_genome_sequence

in	(2)	**in_the**	(1)
en	(3)	**enom**	(1)

リードが短いと多くの位置にヒットしてしまい位置決めが難しい

長くすると多少よくなるが…

ショートリードマッピングには多くのコンピュータ資源が必要

- An illustration

Aligners_might_place_a_short_sequence_in_many_places_in_the_genome_sequence

in (2)	**in_the** (1)	**place** (2)
en (3)	**enom** (1)	**sequence** (2)

リードが短いと多くの位置にヒットしてしまい位置決めが難しい

長くすると多少よくなるが…

変わらない場合もある…

ショートリードマッピングには多くのコンピュータ資源が必要

- An illustration

Aligners_might_place_a_short_sequence_in_many_places_in_the_genome_sequence

in (2)	**in_the** (1)	**place** (2)	**placed** (0, 2)
en (3)	**enom** (1)	**sequence** (2)	

リードが短いと多くの位置にヒットしてしまい位置決めが難しい

長くすると多少よくなるが…

変わらない場合もある…

ミスマッチの考慮も重要で必要(SNP検出などに使われる）

ショートリードマッピングには多くのコンピュータ資源が必要

- An illustration

Aligners_might_place_a_short_sequence_in_many_places_in_the_genome_sequence

in (2)	in_the (1)	place (2)	placed (0, 2)
en (3)	enom (1)	sequence (2)	

リードが短いと多くの位置にヒットしてしまい位置決めが難しい

長くすると多少よくなるが…

変わらない場合もある…

ミスマッチの考慮も重要で必要(SNP検出などに使われる）

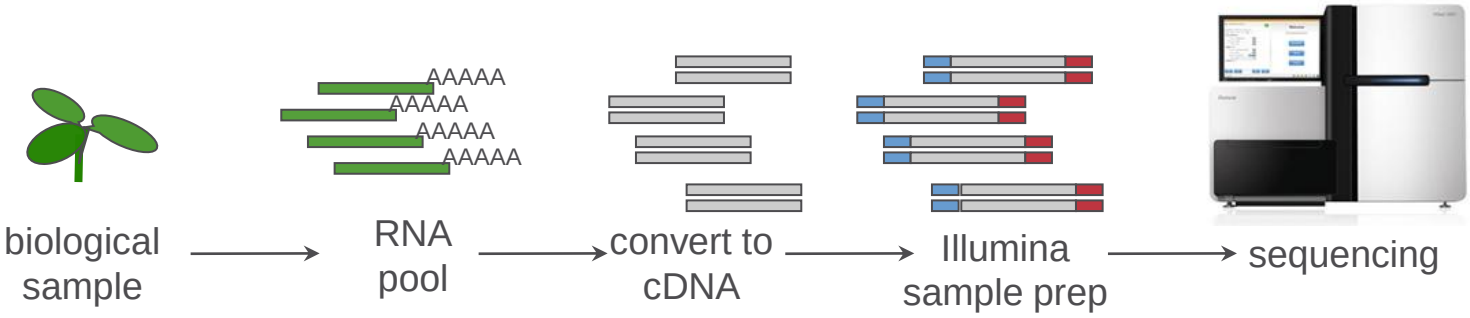
in----------place (1)

PE情報があると助かる

RNA-Seq

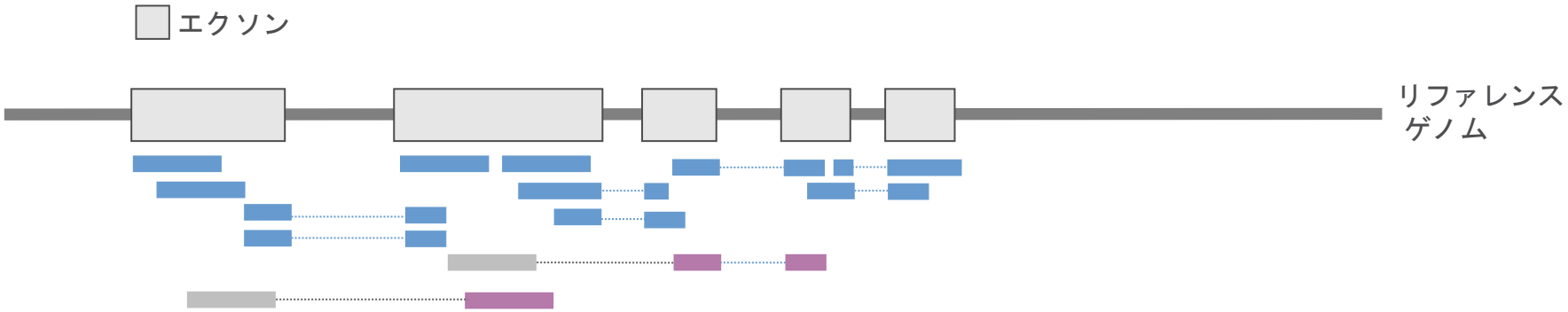
- RNA-Seq ・・遺伝子発現解析等に使われる
- 特定の遺伝子領域のリードの存在量が遺伝子転写産物の存在量を示していると考える
- リード発生量を測定 = カウンティング アプリケーション

<RNA-Seq図>



RNA-Seq

- RNA-Seq アプリケーションは、アライメントから始まる
- RNA-Seqリード はエクソン領域にアラインされる
  - エクソン領域内（エクソンボディー）にアラインする
  - イントロンをまたいだエクソンスプライスジャンクションにアラインする
  - PE では複数のエクソンやスプライスジャンクションにわたるものにも対応

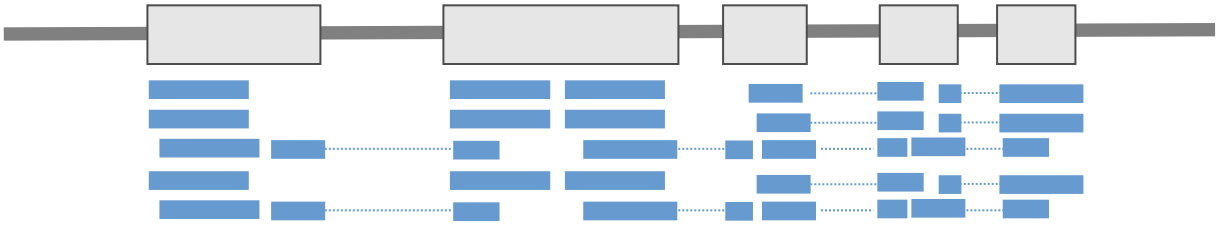


- リードカウント数がRNA 転写産物量に対応すると考える

RNA Seq
遺伝子発現レベルを比較するための正規化

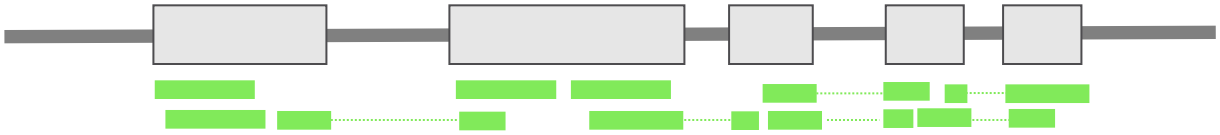
- 発現量の計算はそのサンプルがマップされたリード数、総リード数
  (coverage depth) に影響される

サンプルA　　depth = 5　(50 Million 総リード)



25リード

サンプルB (コントロール)　　depth = 2　(10 Million 総リード)



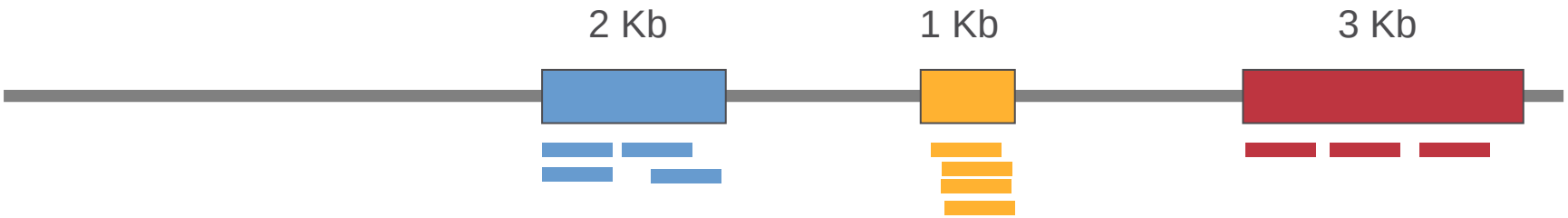
10リード

**サンプル は
コントロールに対し
1/2の遺伝子発現量**

RNA Seq
遺伝子発現レベルを比較するための正規化

- リードカウント数は遺伝子の長さ(全exonの長さ)にも影響される
- 長ければ長いほどリードがマップされる数が多くなり易い

< 異なる３つの遺伝子を想定 >



Raw リードカウント	4	4	3
転写産物の長さで正規化後のリード数	**2**	**4**	**1**

De Novo アセンブリ

- リファレンスゲノムを使わず、

  いちからリード配列をつなぎ合わせて元のゲノムを再構築することがゴール

- 大量のリードを使う（Millions ～ Billions ）
- リード配列のオーバラップを利用しコンティグを作成する
- 非常に計算リソースを消費する

TGACGCTAGCGTCTCAGTCTGATCATACAGTACGTTGACGATAGCGTCT
ACGTTGACGCTAGCGTCTCAGTCTGATCATACAGTACGTTGACGAT
GACGCTAGCGTCTCAGTCTGATCATACAGTACGTTGACGATAGCGTC
GCTAGCGTCTCAGTCTGATCATACAGTACGTTGACGATAGCGTCTC

- De Novo アセンブリ はアライメントのアプローチとは全く異なる
  - 参考 : M Baker (2012) De Novo genome assembly: what every biologist should know., Nature Methods 9:333-337

    http://www.nature.com/nmeth/journal/v9/n4/full/nmeth.1935.html

De Novo アセンブリ (de Brujin graph)

- De Novo アセンブリでは全リード を k-mer に分解する（特定の長さのサブシーケンスに分解）

kmer = 10を適用した場合のイメージ;

TGACGCTAGCGTCTCAGTCTGATCATACAGTACGTTGACGATAGCGTCT

TGACGCTAGC
GACGCTAGCG
ACGCTAGCGT
CGCTAGCGTC

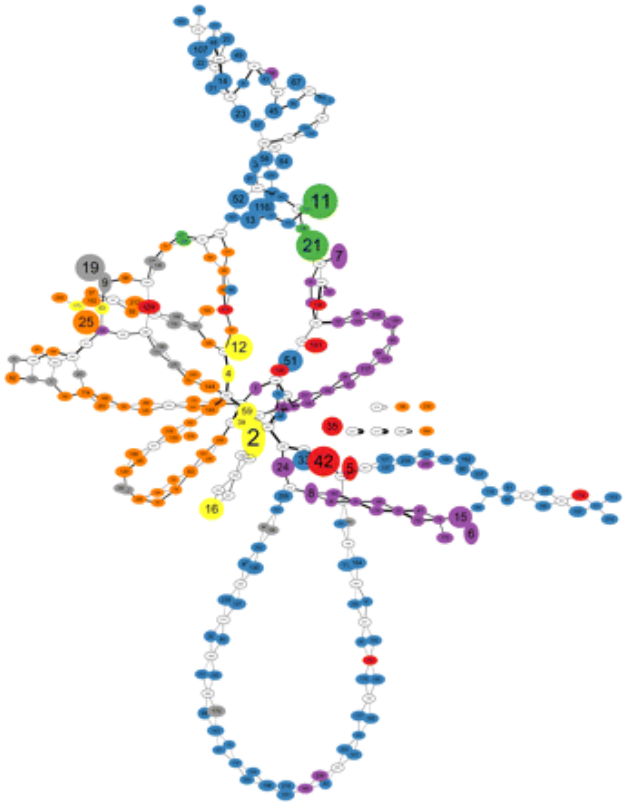
⟶ etc.

- リード全長にわたり行われる
- 全リードに対して繰り返し行われる

- 各 kmer の出現は頻度とともに記録される
- これを全リードに対して行う事で、存在する全てのk-merとその頻度の詳述を作成
- これらの情報を使ってde Brujin グラフを構築する

De Novo アセンブリ (de Brujin graph)

- リード中にある全ての k-mer 間を通る路をみつけることで、ゲノム配列をその路として再構築



- 小ゲノムサイズの、リピートの少ないゲノムで上手くいきやすい
- 概ね50x 以上のカバレッジは必要
- de novoアセンブル はアライメントよりずっと多くの計算リソースを消費する
  （計算がクラッシュすることなどは良くある）
- de Brujinグラフについては論文、Wikipedia、blogなどweb上に多数の情報あり

McComish, Hill, Biggs and Penny. Genome Biol Evol 2010: 2

まとめ

- イルミナシーケンシング計画
  - リード数 / リード長 / ペアード(PE) かシングル(SR) か
  - カバレッジ

アプリケーション毎のScientific community標準と、
装置や試薬の対応範囲を参考

- データ解析の目的と前提
  - 2つの解析タイプ(アライメント、De Novo アセンブリ)
  - サンプルタイプにその解析ソフト選択があっているか
  - 解析により得られる結果と意味

バイオインフォマティクスによる結果は仮説に対する
計算上の実験結果であり、これを踏まえた上での解釈が必要

Appendix

- 弊社英語ホームページ www.illumina.com
  - Coverage Calculator Tech Note
  - onlineコンテンツ: Illumina Technology
  - onlineコンテンツ: CASAVA 1.8

- 文献
  - Bentley et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59

- 弊社日本語ホームページ www.illuminakk.co.jp
  - webinar series
  - 日本語版Tech Note (*全ての日本語版があるわけではありません)

ご清聴ありがとうございました。

ご質問はtechsupport@illumina.comでも承ります。