

# 次世代シーケンサーによる 遺伝子発現解析

佐藤 健吾

慶應義塾大学理工学部生命情報学科

[satoken@bio.keio.ac.jp](mailto:satoken@bio.keio.ac.jp)

# ねらい

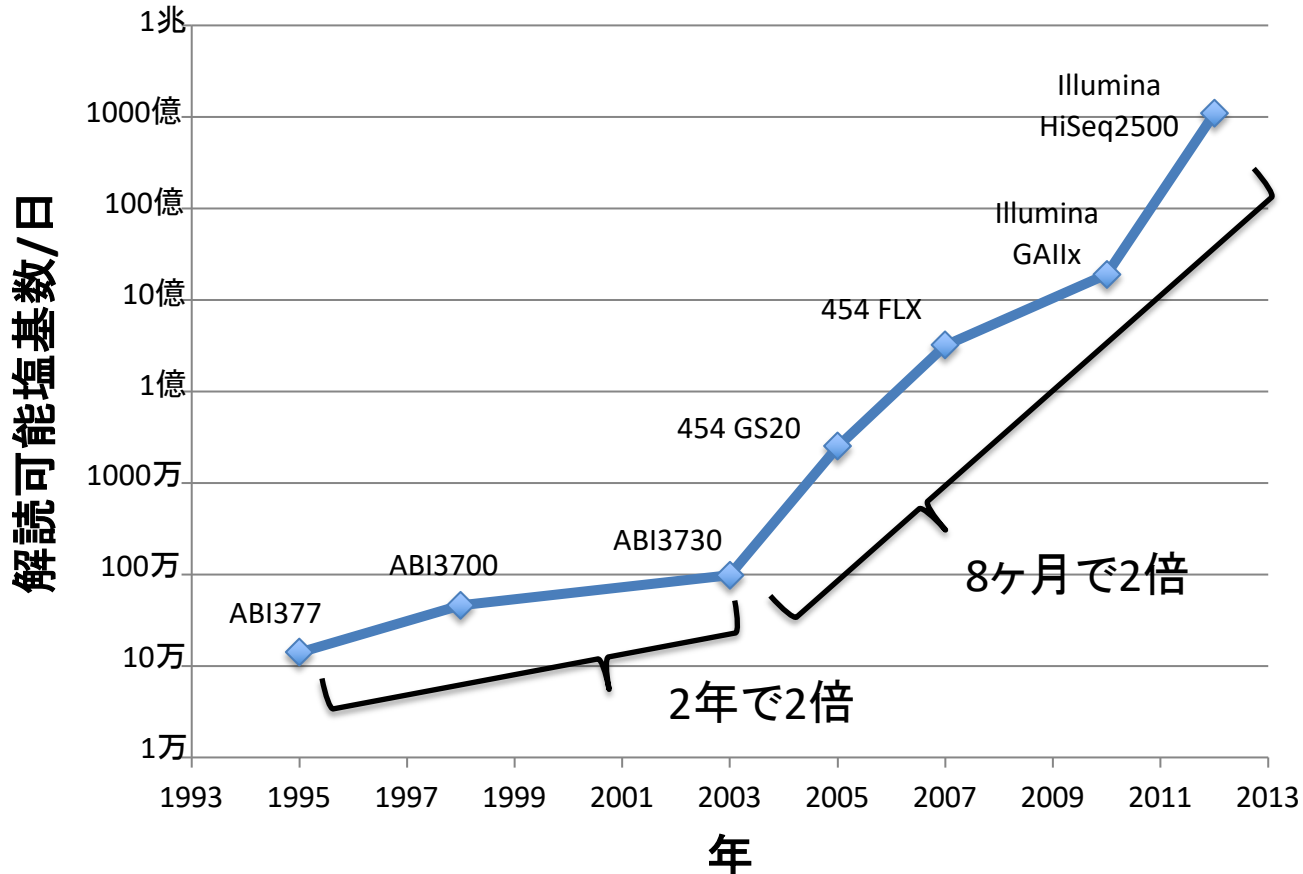
- これからの生命科学において要となるツールである次世代シーケンサー(NGS)が産生するデータを用いたバイオインフォマティクス解析を体験する。
- 次世代シーケンサー
  - 長所: 高速かつ低コスト
  - 短所: 得られる一本一本の配列が短い(= ショートリード)



Illumina GAIIx

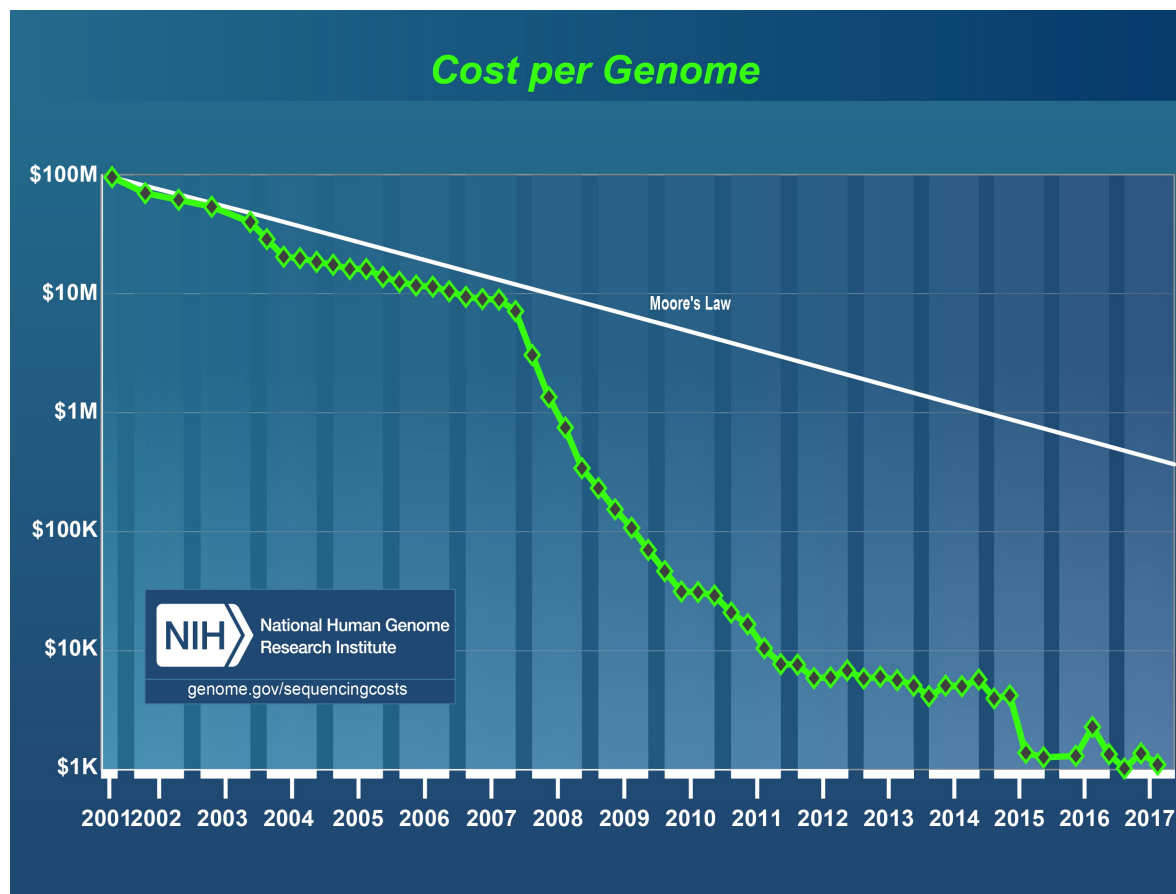
# ゲノム情報ビッグバン

- ゲノムを読むスピードがものすごい勢いで上がっている。



# ゲノム情報ビッグバン

- ゲノムを読むコストもものすごい勢いで下がっている。



# ゲノム情報ビッグバン

- 例: ヒトゲノムの解読

2003年  
キャピラリーシーケンサー  
13年、3000億円以上



2017年  
次世代シーケンサー  
10日間、10万円



ABI 3730



Illumina HiSeq 2500

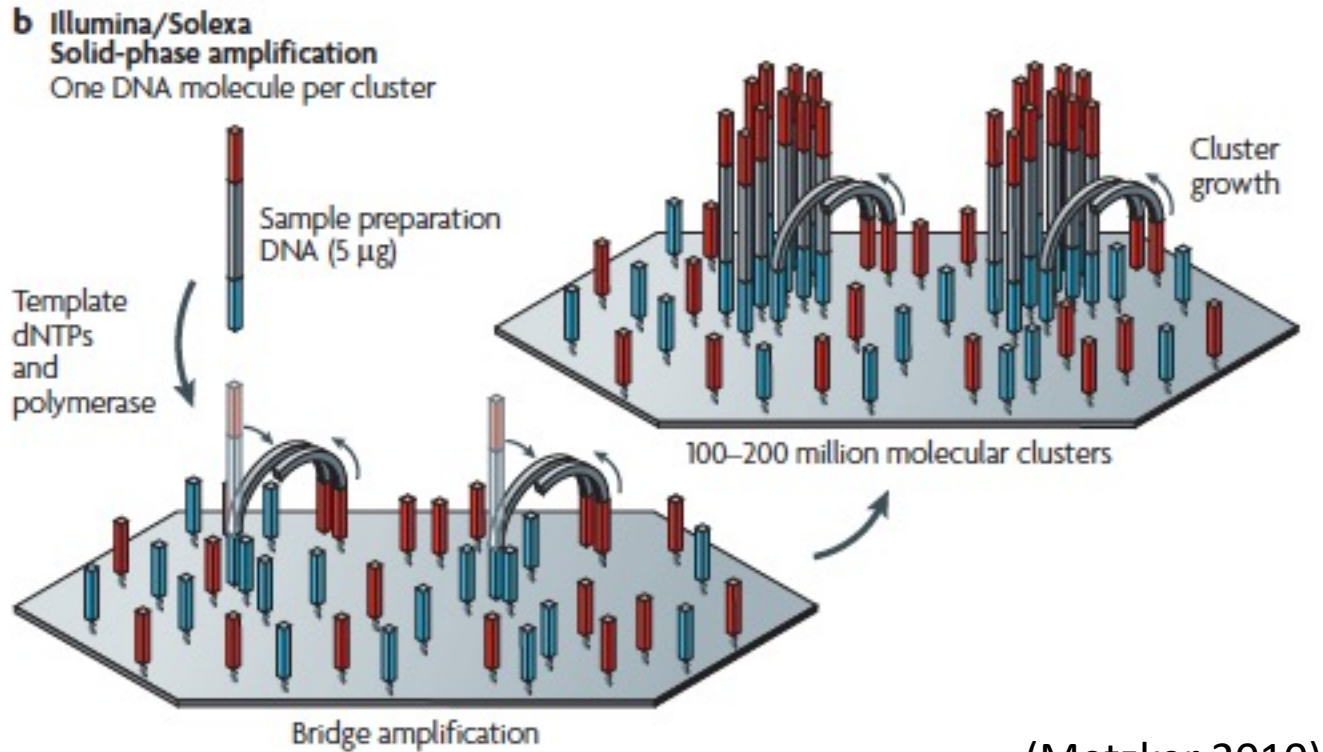
# 次世代シーケンサー

- 超並列シーケンシング
  - Illumina/GA
  - Roche/454
  - ABI/SOLiD
- 特徴
  - リード長が短い。(25～400塩基)
  - 1塩基を読むのに1時間かかる。
  - 5000万～10億個を並列に読む。

# 性能比較

	ABI 3730xl	Illumina GAIIx	Illumina HiSeq 2500	Illumina HiSeq X
年	2002	2009	2011	2014
平均リード長(塩基)	750	75	100	150
リード数	96	1億	30億	60億
塩基数/ラン	72kb	5Gb	600Gb	1800Gb

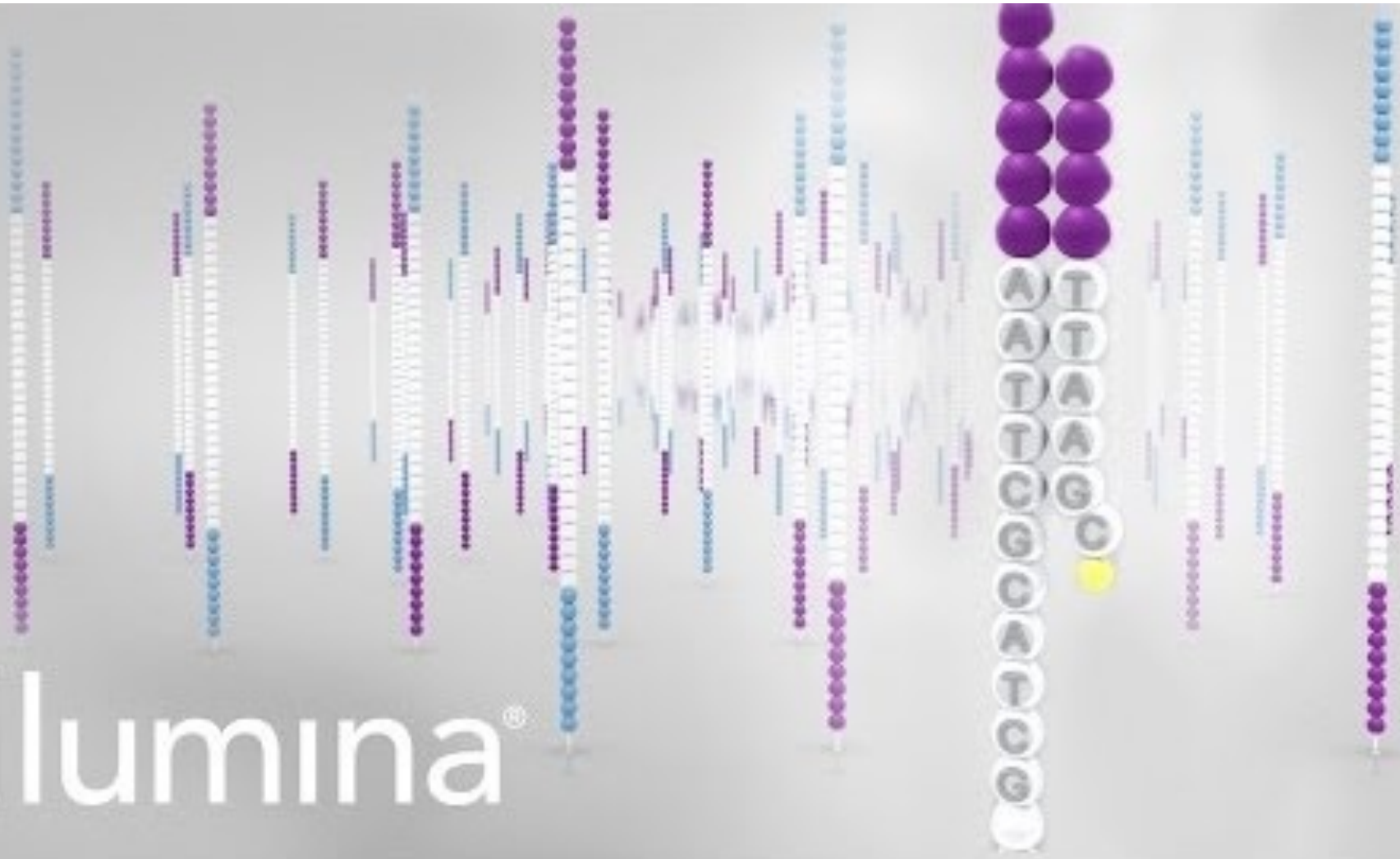
# 次世代シーケンサー



(Metzker 2010)



illumina®

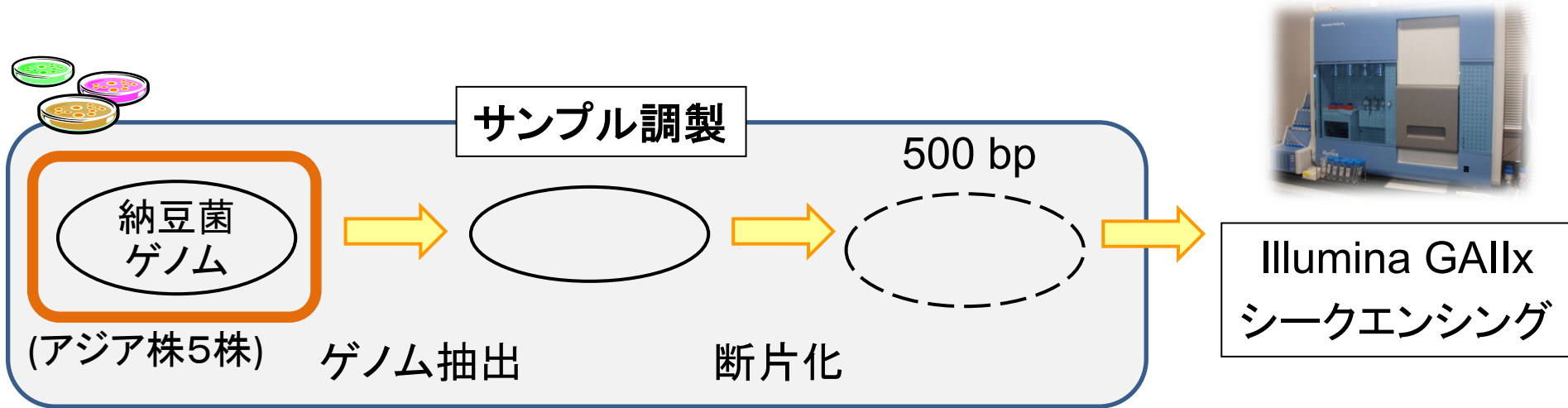


<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

# どのようなデータが得られるか？

- ゲノムデータ
- トランスクリプトームデータ
- エピゲノムデータ
- 相互作用データ

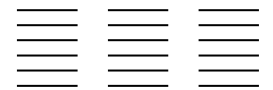
# 解析の流れ - サンプル調製とシーケンシング -



## • ペアエンドリード

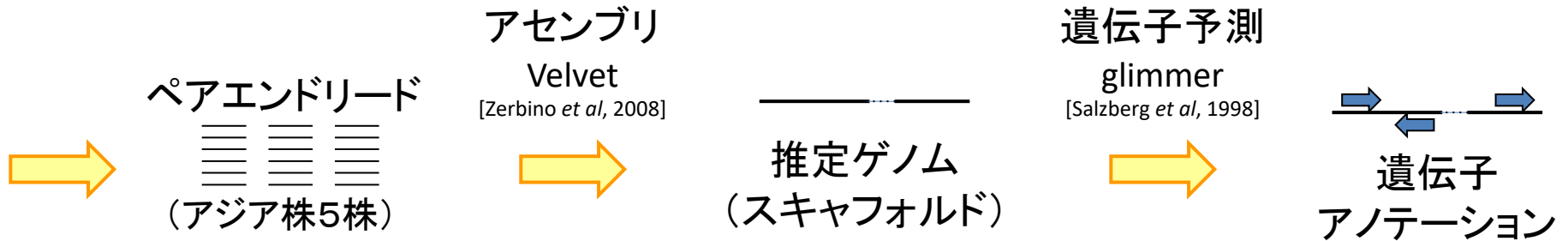
- 断片化されたショートリードの両端を読む

ペアエンドリード

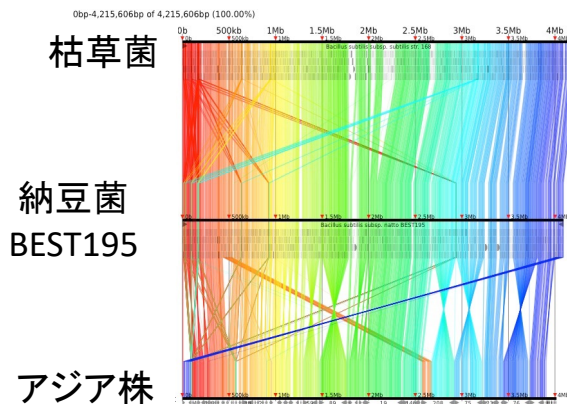


# 解析の流れ - de novo アセンブリ -

## • 参照するゲノムがない場合の解析方法



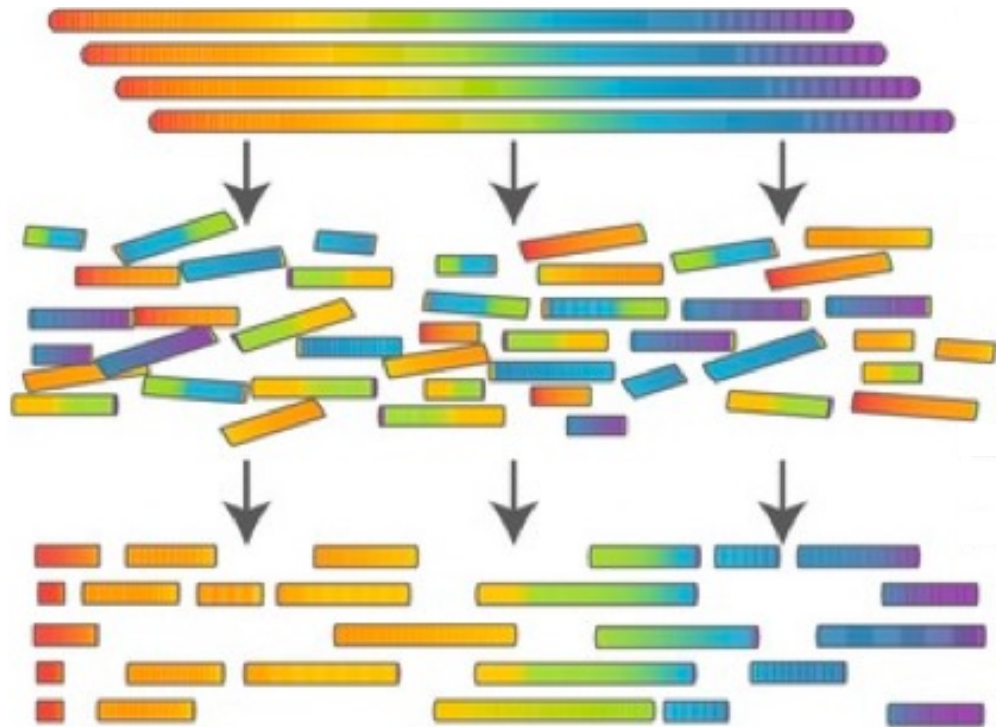
比較ゲノム  
Murasaki  
[Popendorf et al, 2010]



- オーソログ遺伝子
- リピート領域の解析

# ショットガンゲノムアセンブリ

- ゲノムを細切れにしてから解読し、つなぎ合わせて復元する。



①ゲノム抽出

②断片化・シーケンシング

③ゲノムアセンブリ

ATGTTCCGATTAGGAAACCTATCTGTAACGTGTTTCATTTCAGTAAAAGGAGGAAATATAA

# 解析の流れ - 変異解析 -

## • 参照するゲノムがある場合の解析方法



変異影響度  
アノテーション  
snpEff  
[Cingolani et al, 2012]

- 同義置換
- 非同義置換
- 非コード領域の置換
- ...

- 遺伝子の機能
- 表現型の解析

# 変異解析

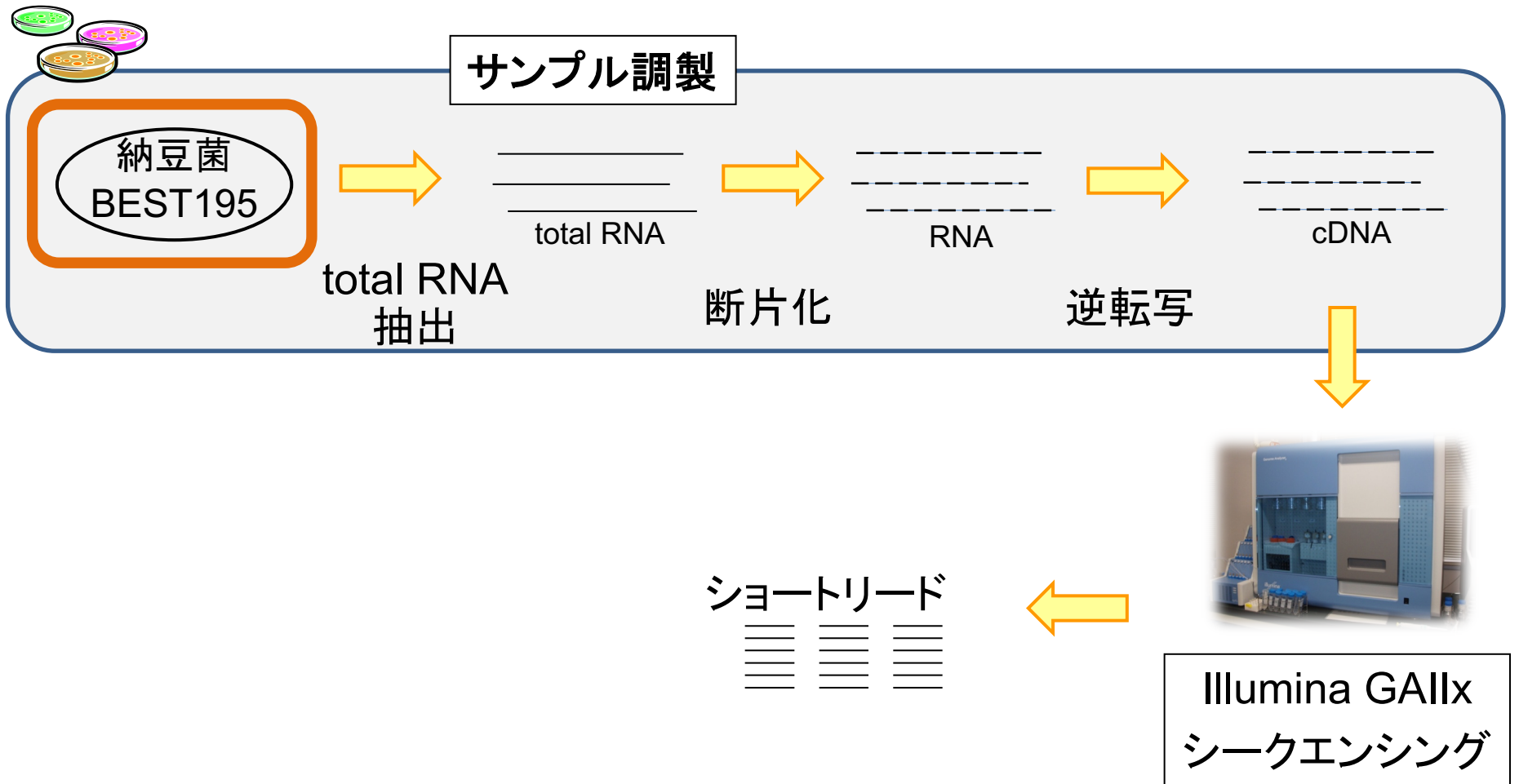
- シークエンサーで読んだリードをリファレンスゲノムにマッピングすることによって、どこに変異が入っているかを網羅的に計測できる。
- 最近はrare variant (<1%) が注目されている。  
→ パーソナルゲノム

# トランスクリプトーム解析

- RNA-seqという技術で転写物を網羅的に計測できる。
  - mRNAの発現量
  - 選択的スプライシングの検出
  - 遺伝子のフュージョン(ガン細胞など)
  - 新規non-coding RNAの発見



# 解析の流れ - RNA-Seq -



# 解析の流れ - RNA-Seq -

- 異なる条件におけるmRNAの発現量を比較する

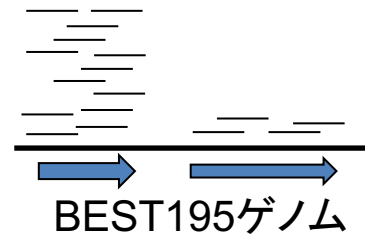
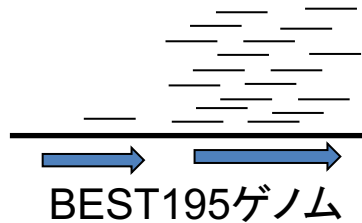
ねばねばする  
培養条件(BN)



ねばねばしない  
培養条件(PN)



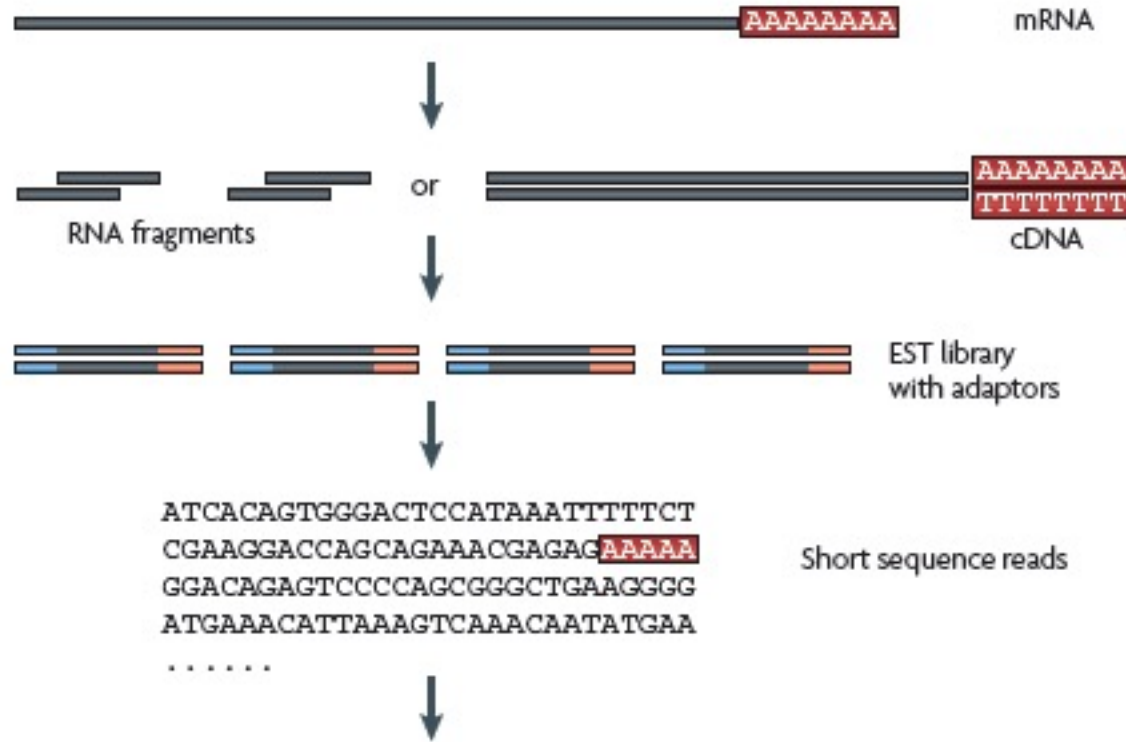
マッピング  
Bowtie2  
[Langmead et al, 2012]



発現量比較  
cuffdiff  
[Trapnell et al, 2010]

各遺伝子の発現差  
があるかないか

# RNA-seq

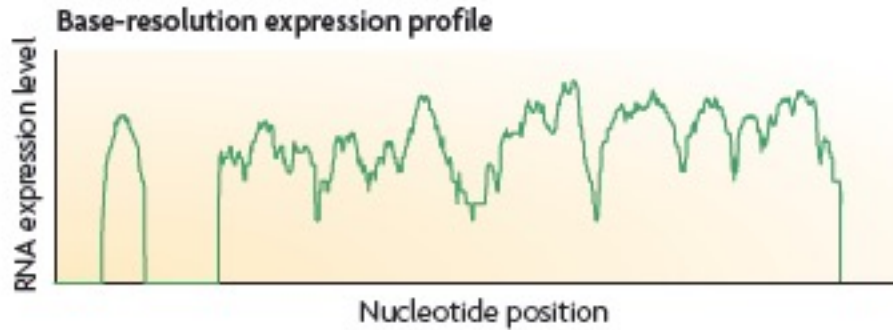
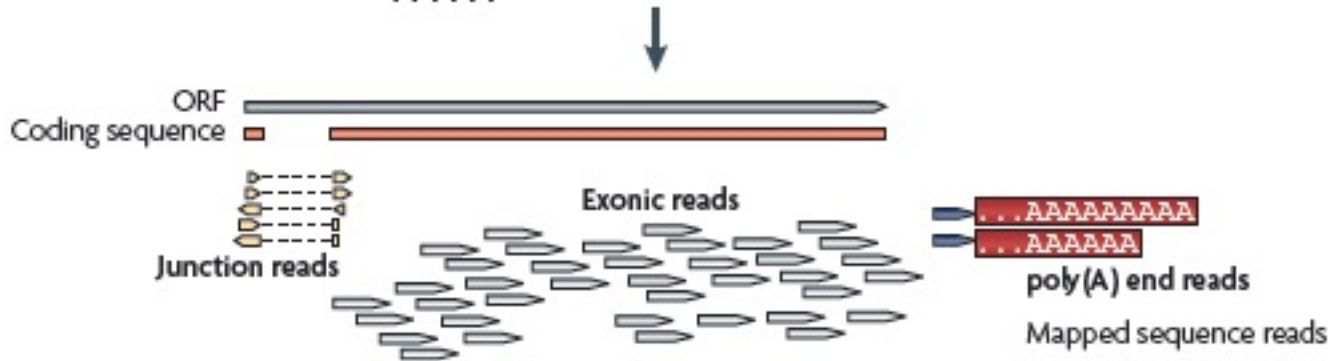


(Wang+ 2009)

# RNA-seq

```
ATCACAGTGGGACTCCATAAATTTTCT  
CGAAGGACCAGCAGAAACGAGACAAAA  
GGACAGAGTCCCAGCGGGCTGAAGGGG  
ATGAAACATTAAAGTCAAACAATATGAA  
.....
```

Short sequence reads



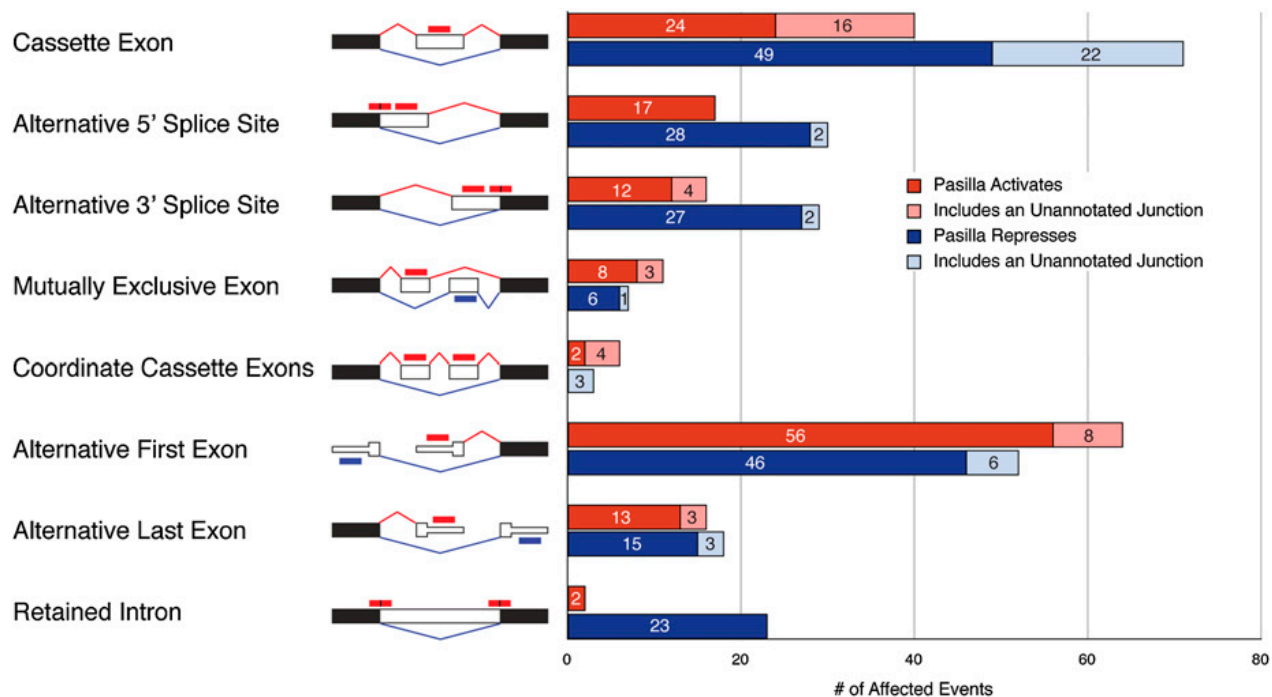
(Wang+ 2009)

# 今週の課題

- キイロショウジョウバエ *D. melanogaster* のトランスクリプトーム解析を Galaxy 上で行う。
- [Brooks *et al.*, 2011] において、スプライシング制御因子である *Pasilla* 遺伝子 (*ps*) を RNAi でノックダウンしたサンプルとコントロールサンプルの RNA-seq を行なった。
- 本課題ではこの論文のデータを使って、
  - 二つの条件で得られた RNA-seq リードを参照ゲノムにマッピングし、各遺伝子の発現量を推定する。
  - 二つの条件の間で発現に差異がある遺伝子を同定する。
  - これらの解析に基づき遺伝子機能解析を行い、二つの条件の生物学的な差異を考察する。

# 元の論文[Brooks *et al.*, 2011]の概要

- 二つの条件でスプライシングがどのように変化するかを調べ、*ps*に制御されている遺伝子を同定した。



# 元の論文[Brooks *et al.*, 2011]の概要

- 哺乳類におけるオーソログ遺伝子であるNOVA1およびNOVA2が制御している遺伝子とpsが制御している遺伝子を比較した。
- その結果、制御因子(NOVA1/2, ps)の配列はよく保存されているものの、その制御対象に共通する遺伝子は多くないことがわかった。