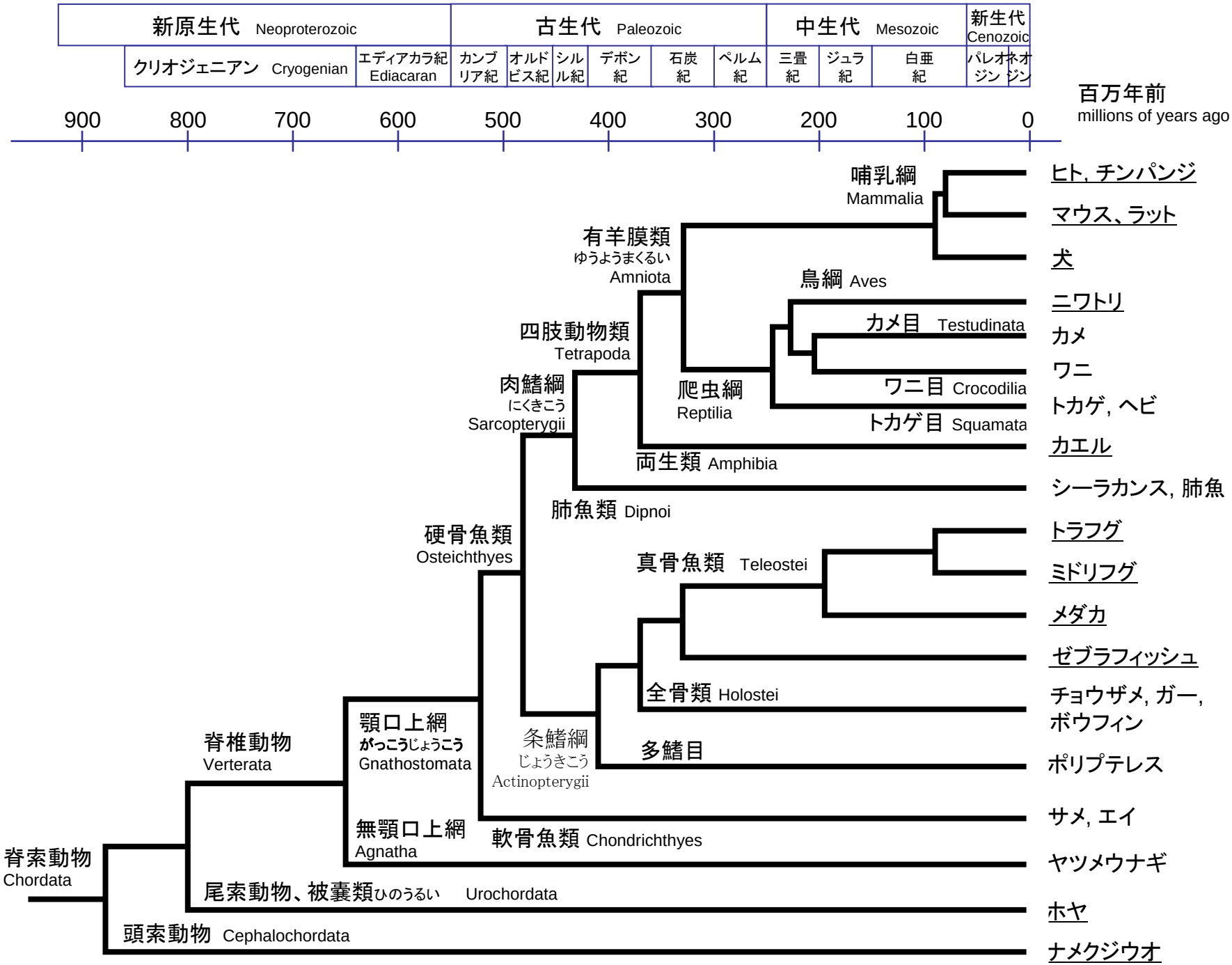
ゲノムアセンブリ


新原生代 Neoproterozoic
古生代 Paleozoic
中生代 Mesozoic
新生代 Cenozoic
クリオジェニアン Cryogenian
エディアカラ紀 Ediacaran
カンブリア紀
オルドビス紀
シルル紀
デボン紀
石炭紀
ペルム紀
三畳紀
ジュラ紀
白亜紀
パレオジン
ネオジン
900
800
700
600
500
400
300
200
100
0
百万年前
millions of years ago
哺乳綱 Mammalia
ヒト, チンパンジ
マウス、ラット
犬
有羊膜類 ゆうようまくるい Amniota
鳥綱 Aves
ニワトリ
カメ目 Testudinata
カメ
ワニ
四肢動物類 Tetrapoda
肉鰭綱 にくきこう Sarcopterygii
爬虫綱 Reptilia
ワニ目 Crocodilia
トカゲ, ヘビ
トカゲ目 Squamata
カエル
両生類 Amphibia
シーラカンス, 肺魚
肺魚類 Dipnoi
硬骨魚類 Osteichthyes
トラフグ
真骨魚類 Teleostei
ミドリフグ
メダカ
ゼブラフィッシュ
全骨類 Holostei
チョウザメ, ガー, ボウフィン
顎口上網 がっこうじょうこう Gnathostomata
条鰭綱 じょうきこう Actinopterygii
多鰭目
ポリプテレス
脊椎動物 Verterata
サメ, エイ
無顎口上網 Agnatha
軟骨魚類 Chondrichthyes
ヤツメウナギ
脊索動物 Chordata
尾索動物、被嚢類 ひのうるい Urochordata
ホヤ
頭索動物 Cephalochordata
ナメクジウオ

大規模ゲノムアセンブリの状況

論文発表	種	総塩基数	アセンブリ方式	
2001 / 2	ヒト	29億	clone-by-clone	国際チーム
			Celera	アメリカ
2002 / 4	イネ(染色体地図なし)	4.7億	RePS (Beijing Genomics)	中国
2002 / 7	トラフグ(染色体地図なし)	3.6億	Jazz (JGI)	アメリカ
2002 / 12	マウス	25億	Arachne (MIT) + clone by clone	アメリカ
2004 / 2	カイコ(染色体地図なし)	5億	Ramen (東大)　農業生物資源研究所	日本
			RePS (Beijing Genomics)	中国
2004 / 4	ラット	25億	Atlas (Baylor College) + clone by clone	アメリカ
2004 / 10	ミドリフグ	3.4億	Arachne (MIT)	フランス, アメリカ
2004 / 12	チキン	10億	PCAP (Wash. U.)	アメリカ
2005 / 8	イネ	3.9億	clone by clone　農業生物資源研究所	日本
2005 / 9	チンパンジ	29億	PCAP, Arachne	アメリカ
2005 / 12	ドッグ	24億	Arachne (MIT)	アメリカ
2006 / 10	ミツバチ	2.3億	Atlas (Baylor College)	アメリカ
2006 / 11	ウニ	8億	Atlas (Baylor College)	
2007 / 4	アカゲザル	29億	Atlas(Baylor), P-CAP(Wash U), Celera	アメリカ
2007 / 5	オポッサム	34億	Arachne (MIT)	アメリカ
2007 / 6	メダカ	7億	Ramen (東大)　国立遺伝学研究所	日本
2008 / ?	ゼブラフィッシュ	16億	Phesion (Sanger Ctr.)	イギリス
2008 / ?	アフリカツメガエル	16億	Jazz (JGI)	アメリカ
2008 / ?	ナメクジウオ	6億?	Jazz (JGI)	アメリカ

大規模なゲノムシークエンシングセンター

米国
- Joint Genome Institute, US Dept. of Energy
- Whitehead Institute / MIT Center for Genome Research
- Washington University Genome Sequencing Center
- Baylor College of Medicine

英国
- Wellcome Trust Sanger Institute

日本
- 国立遺伝学研究所
- 理化学研究所　ゲノム科学総合研究センター
- かずさ DNA 研究所
- 農業生物資源研究所

シークエンシング技術の高速化

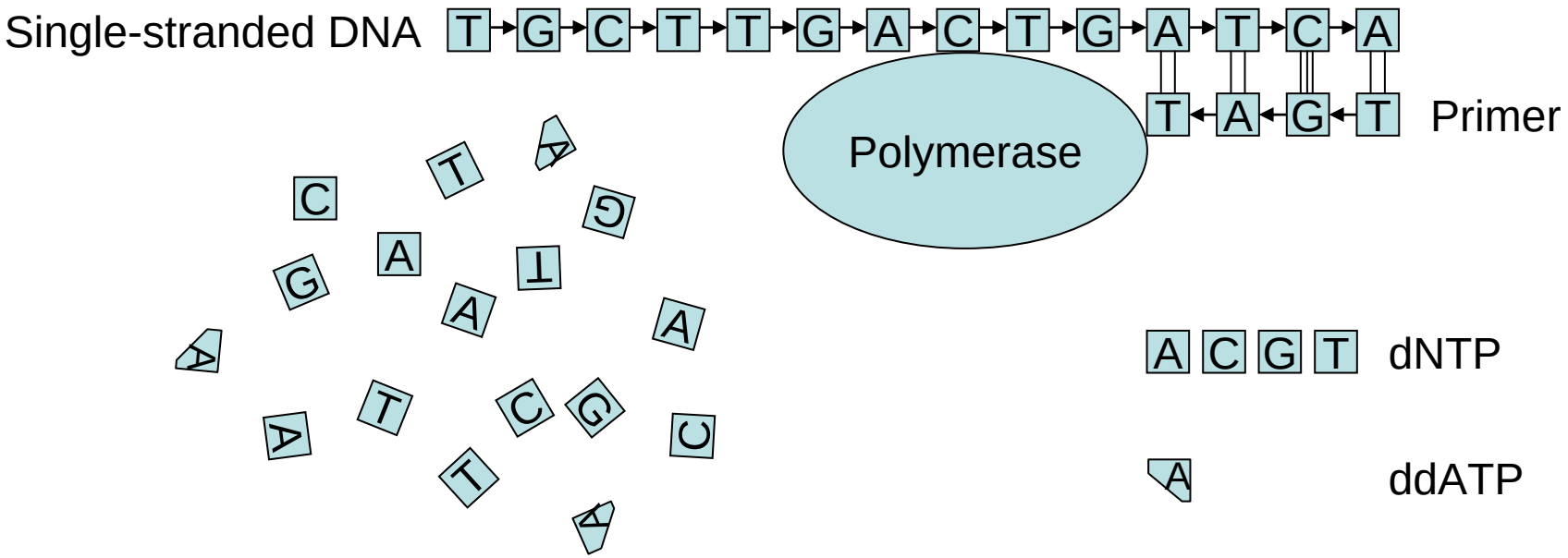
- ヒトゲノムプロジェクト $2.7 billion, 17年
- 2004年の段階　哺乳類ゲノム (3G塩基) の解読 $10-50 million
- NIHファンド　“$1000 genome project” Feb.2004
- 2005年夏から驚異的な高速化

	配列（リード）長	収集可能タグ数*	総塩基数
SOLEXA	25 – 50 nt	40,000,000/実験	10 – 20 億/実験
454	100 – 250 nt	300,000/実験	0.3 – 0.75 億/実験
ABI 3730xl	500 – 800 nt	2304/　日	0.012 – 0.02 億/　日

*SOLEXAは1回の実験に 3-4 日, 454 は 7-8 時間
註 SOLEXA の方式は illumina の HP を参照してください

新型シークエンサーの応用例
- あたらしいSNP･挿入･削除の発見
- 免疫沈降法と組み合わせた エピゲノム解析および転写因子結合部位解析
- 遺伝子発現のプロファイリング

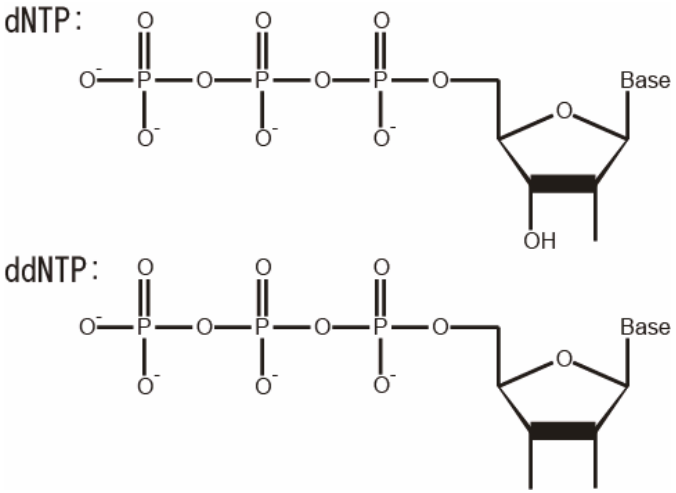
Sanger Method (1975)

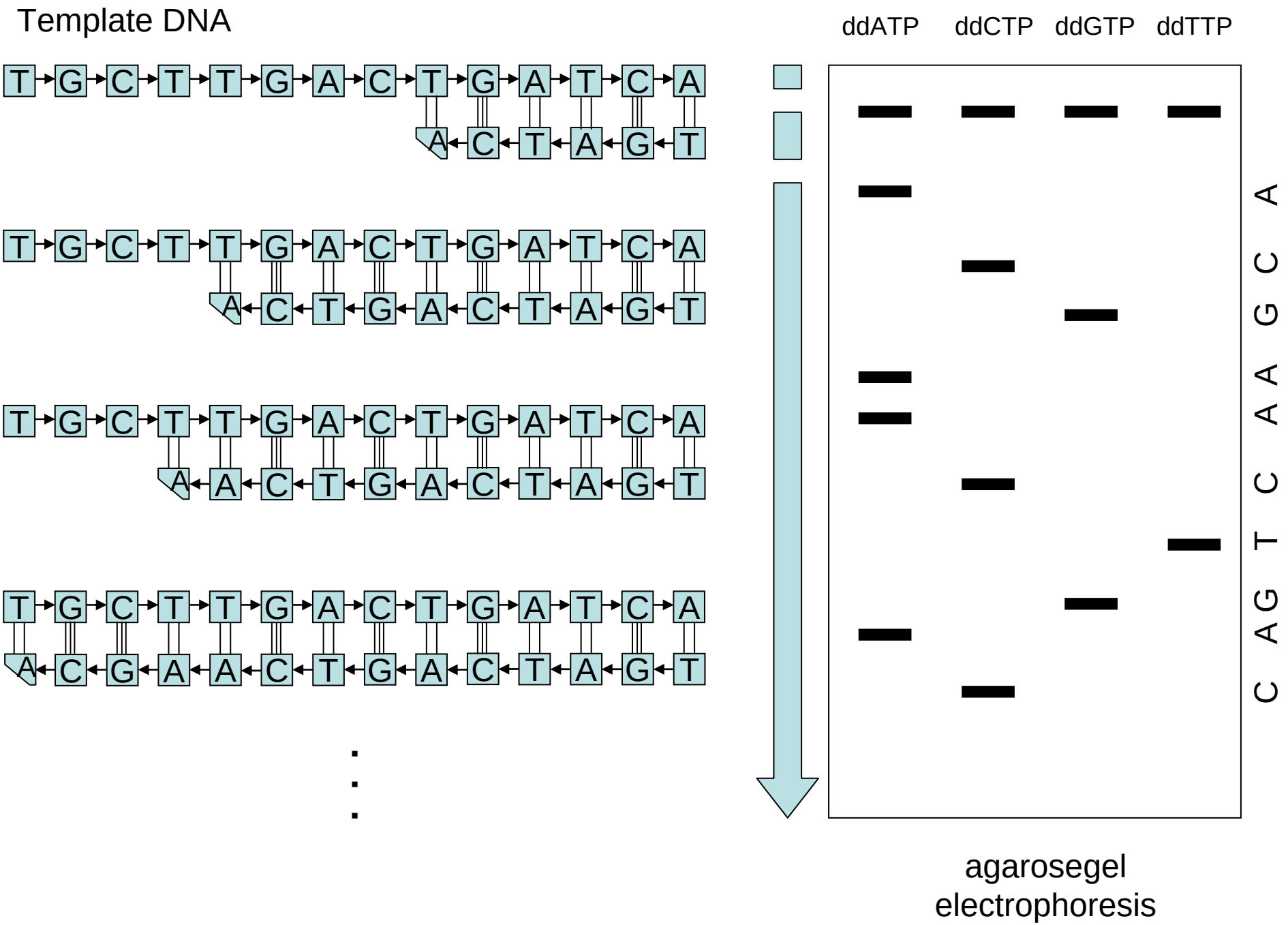


ddNTP (deoxy_____ triphosphate)

adenosine
cytosine
guanosine
tyrosine

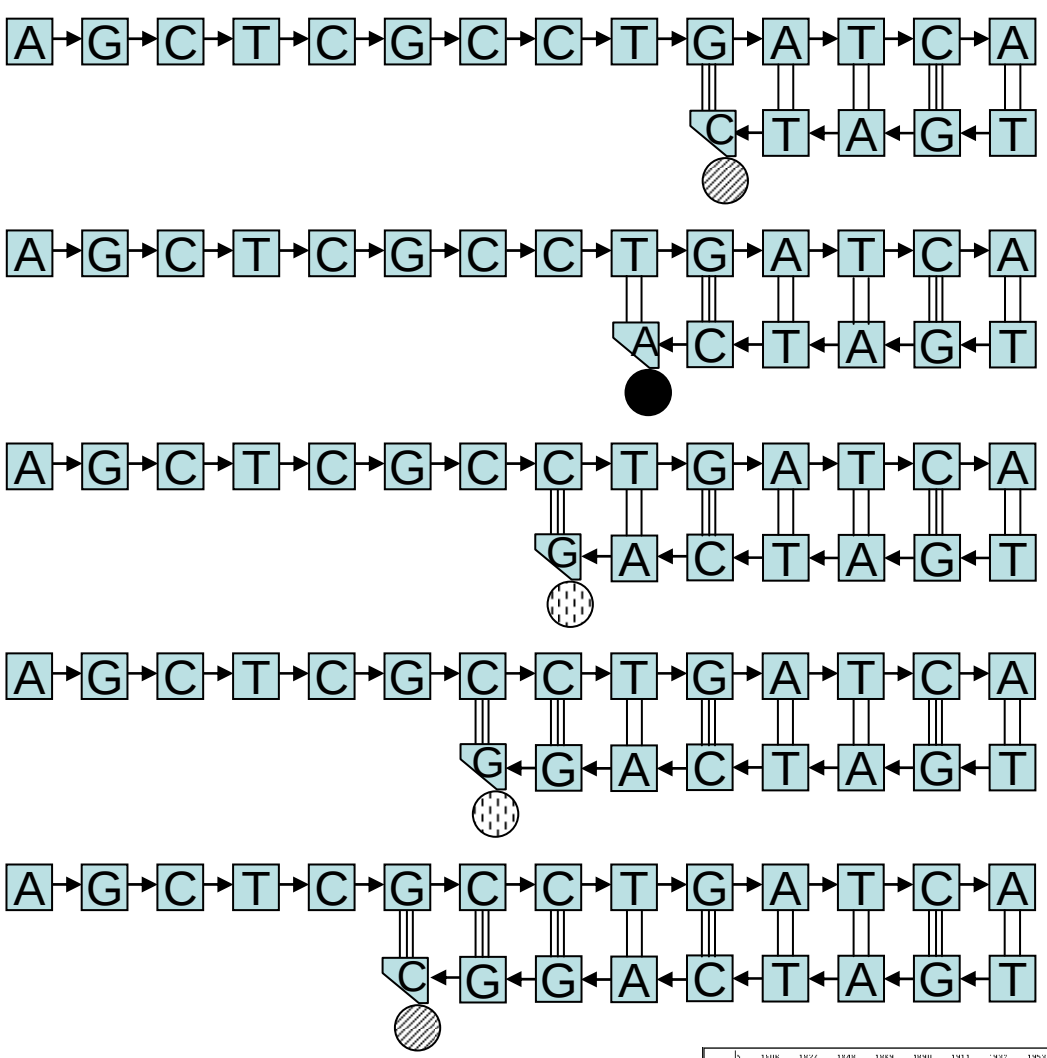
ddNTP(dideoxy _____ triphosphate)

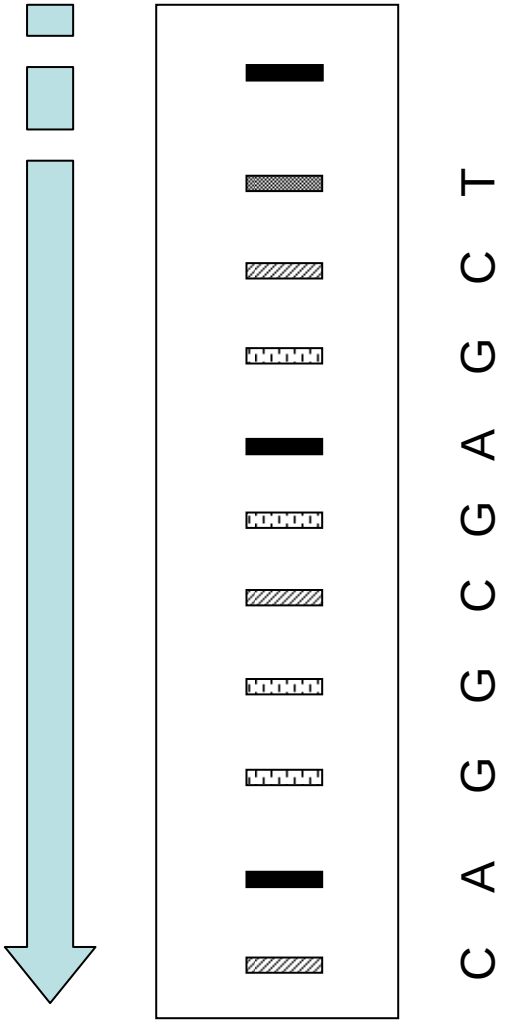


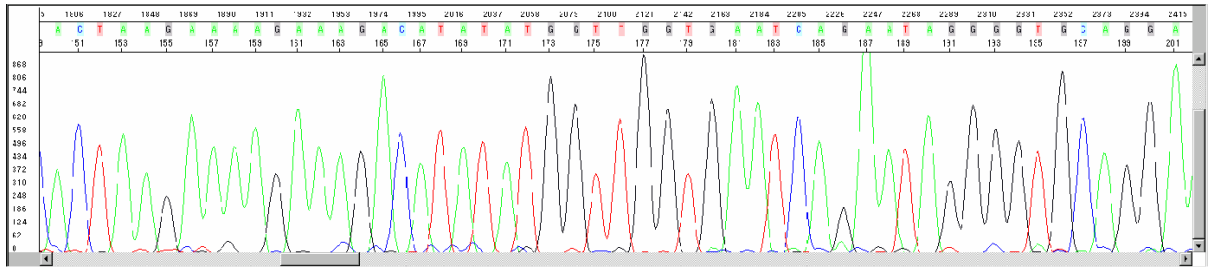

Template DNA
T G C T T G A C T G A T C A
A C T A G T
T G C T T G A C T G A T C A
A C T G A C T A G T
T G C T T G A C T G A T C A
A A C T G A C T A G T
T G C T T G A C T G A T C A
A C G A A C T G A C T A G T
ddATP
ddCTP
ddGTP
ddTTP
C A G T C A A G C A
agarosegel
electrophoresis

four-color fluorescent dye method (1986)

Template DNA





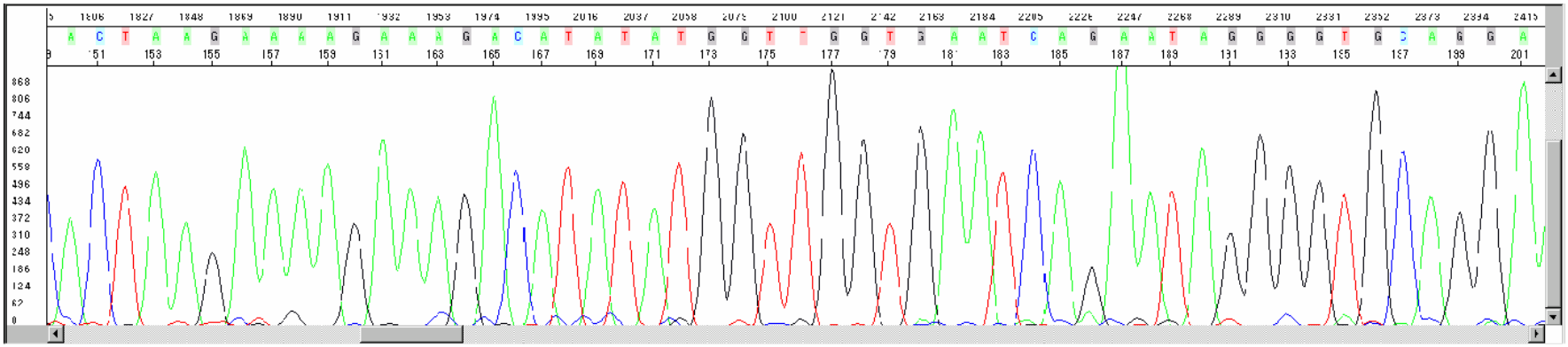


which is often accomplished using the computer program *phred* [32]. Phred detects a series of peaks in input electropherograms, and outputs nucleotide sequences and their *quality values* (QVs), which use a logarithmic scale and satisfy the following equation:

$$\mathrm{QV} = -10 \log_{10} P,$$

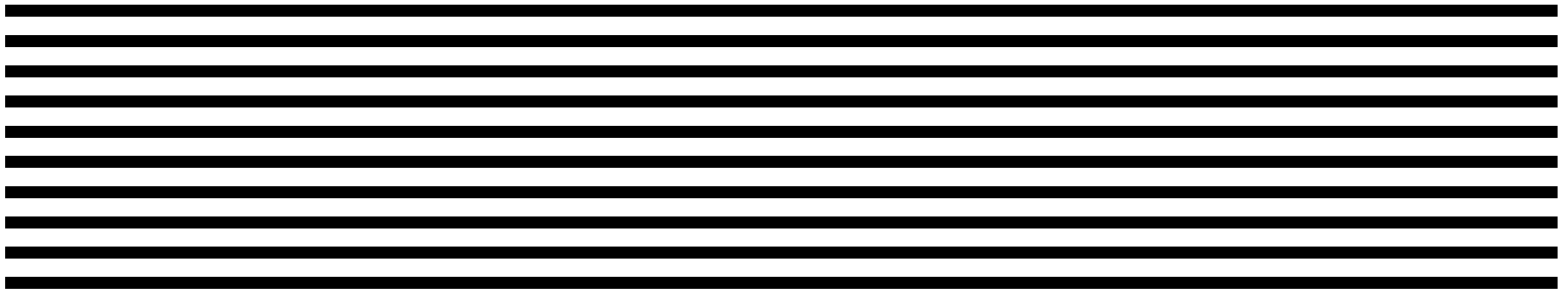
where P is the probability of an error at a base. A QV is assigned to each

QV	accuracy
4	60%
7	80%
10	90%
14	96.0%
20	99.0%
24	99.6%
30	99.9%
40	99.99%
50	99.999%

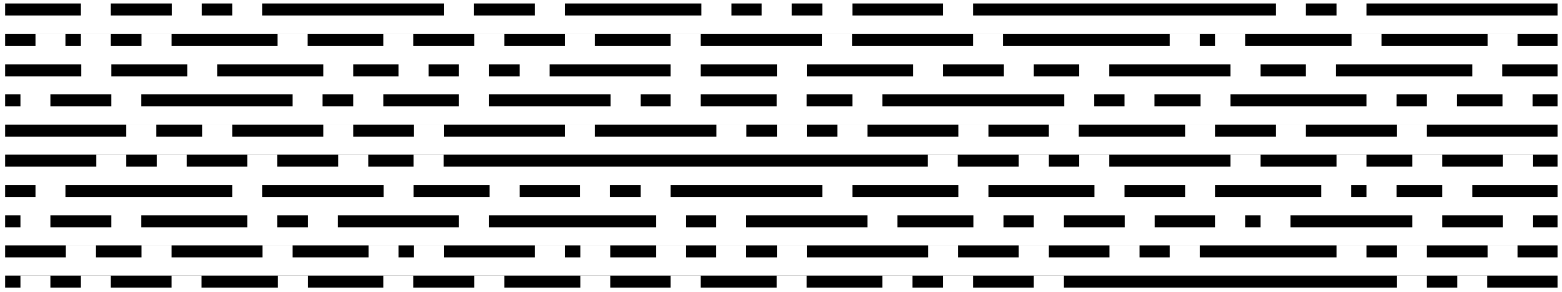


ゲノムアセンブリの手順

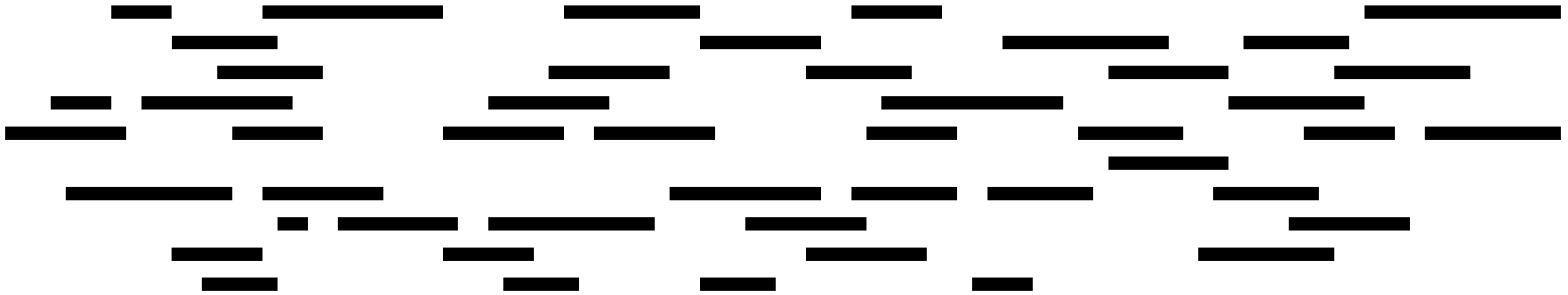
a) Multiple copies of genome

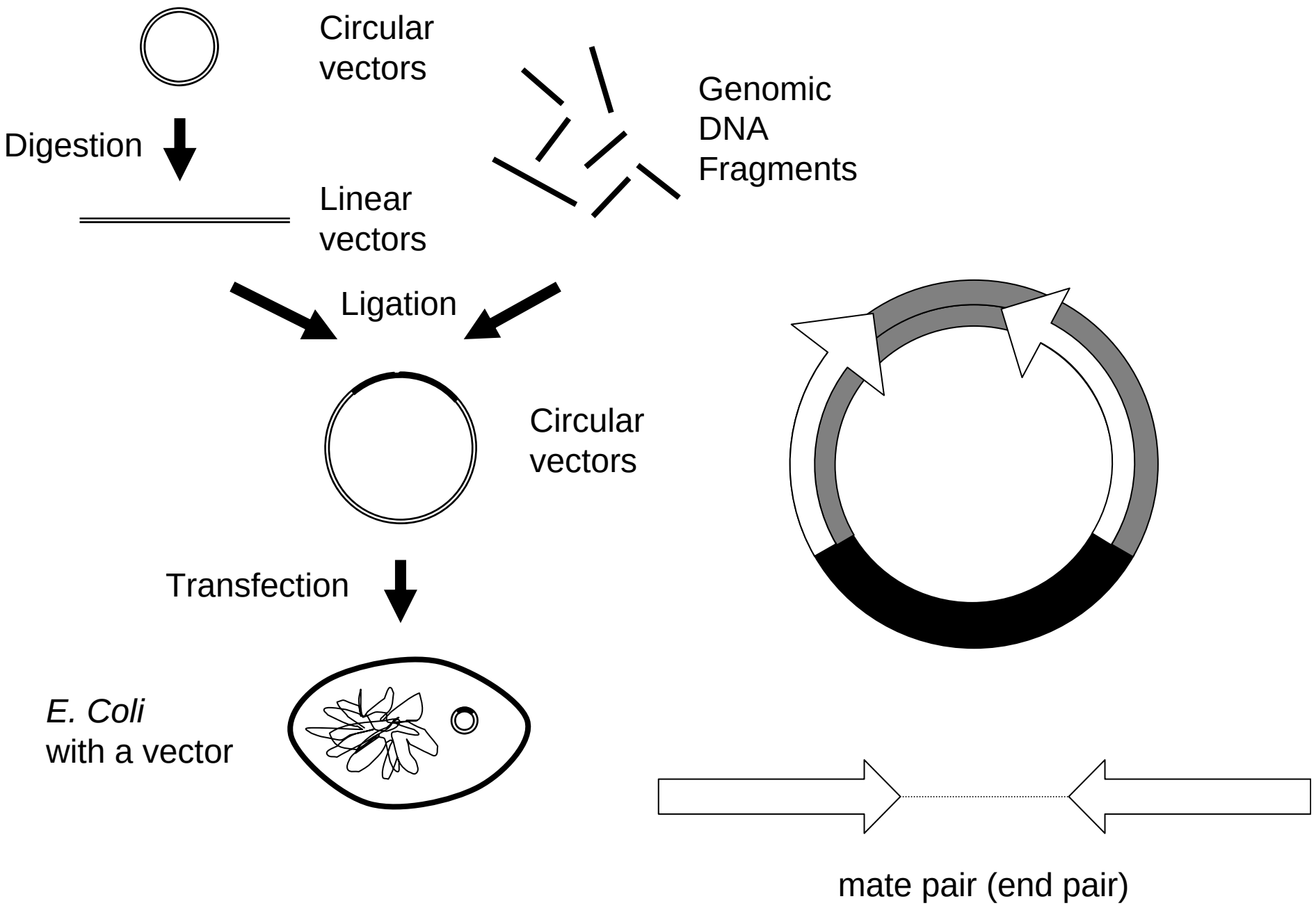


b) Sheared random fragments by fast water flow



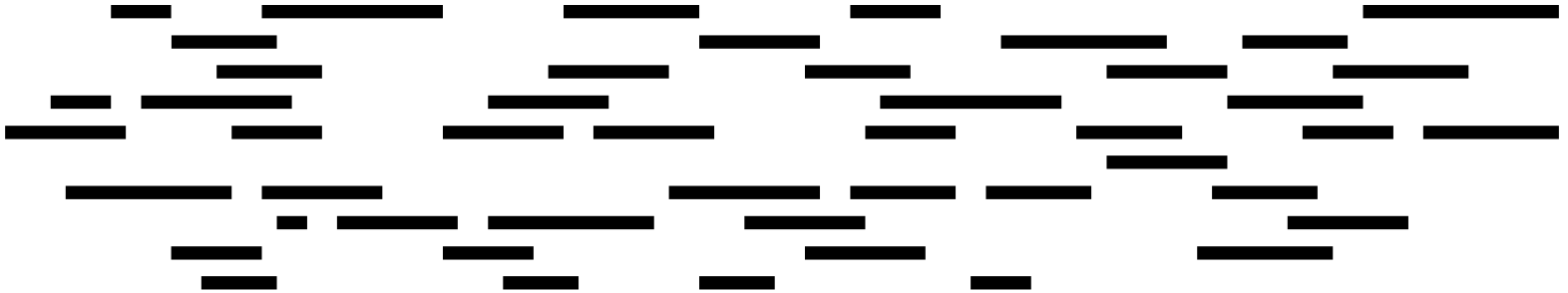
c) Size fractionated fragments



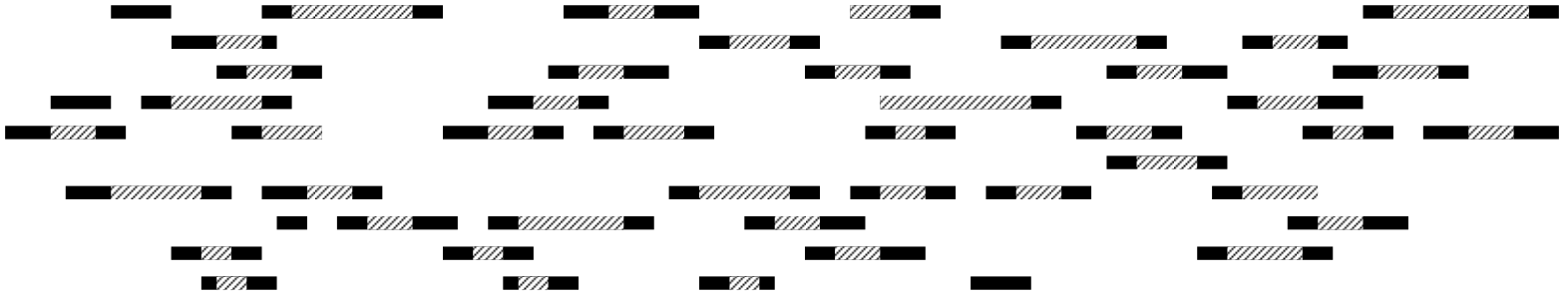

Circular
vectors
Digestion
Linear
vectors
Genomic
DNA
Fragments
Ligation
Circular
vectors
Transfection
E. Coli
with a vector
mate pair (end pair)

ゲノムアセンブリの手順

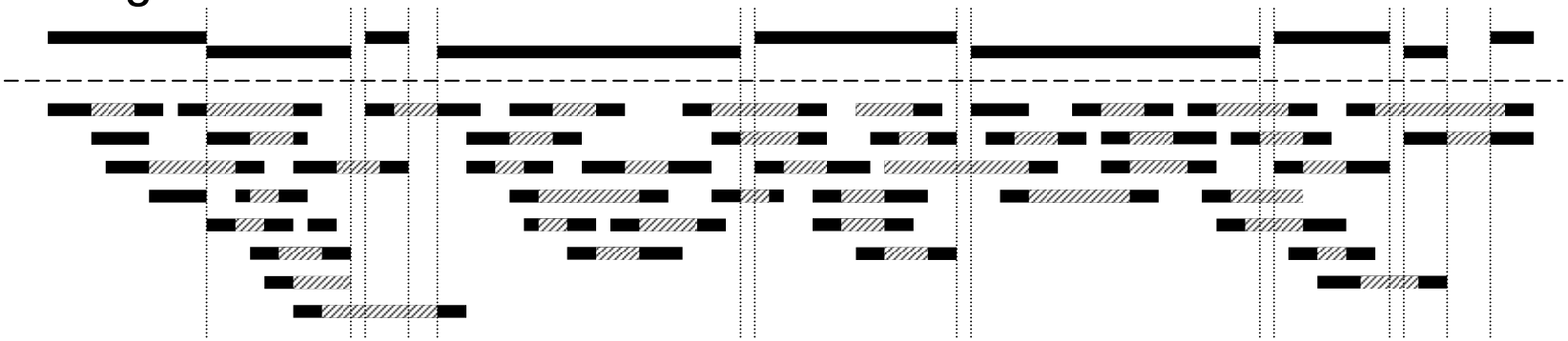
c) Size fractionated fragments



d) Reads



e) Contigs



Contig 生成の詳細

Original Reads

```
1:CCTATGCTAGTCA
2:CGACTGACTAGCAT
3:GCTAGTCAGTCGATCTACC
4:ACCGGTAGATCGACTG
```

Double Stranded Reads

```
1:CCTATGCTAGTCA
1̄:TGACTAGCATAGG
2:CGACTGACTAGCAT
2̄:ATGCTAGTCAGTCG
3:GCTAGTCAGTCGATCTACC
3̄:GGTAGATCGACTGACTAGC
4:ACCGGTAGATCGACTG
4̄:CAGTCGATCTACCGGT
```

Assembly

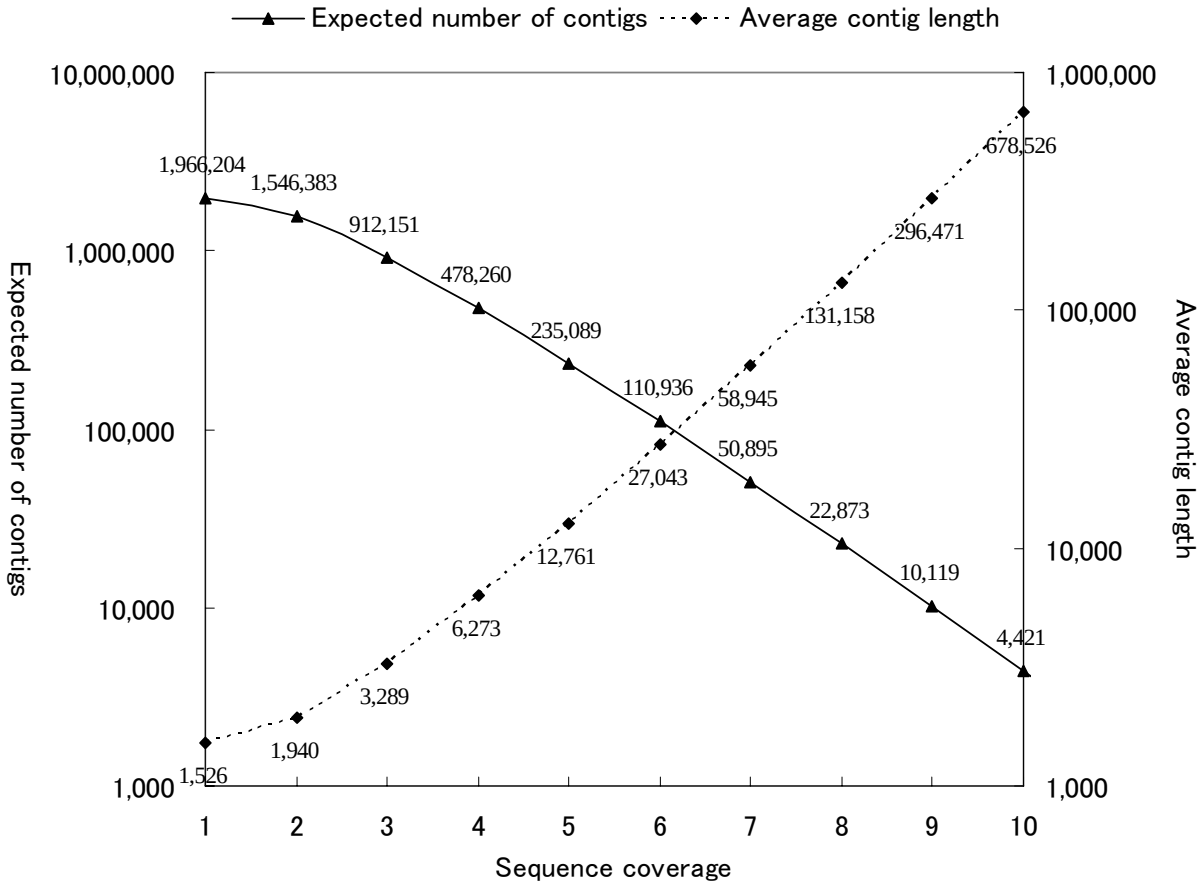
```
1:CCTATGCTAGTCA
    2̄:ATGCTAGTCAGTCG
       3:GCTAGTCAGTCGATCTACC
             4̄:CAGTCGATCTACCGGT
```

```
1:CCTATGCTAGTCA
    2̄:ATGCTAGTCAGTCG
       3:GCTAGTCAGTCGATCTACC
             4̄:CAGTCGATCTACCGGT

4:ACCGGTAGATCGACTG
    3̄:GGTAGATCGACTGACTAGC
          2:CGACTGACTAGCAT
             1̄:TGACTAGCATAGG
```

Lander-Waterman Statistics （Contig 平均長の推定）

Genome size $G = 3*10^9$. Given a random collection of N fragments of size $L = 600$.
Sequence coverage = NL / G, *e.g.*, = 10 if $N = 5* 10^7$.
Join two fragments that share $L\theta$ nucleotides ($\theta = 0.1$).

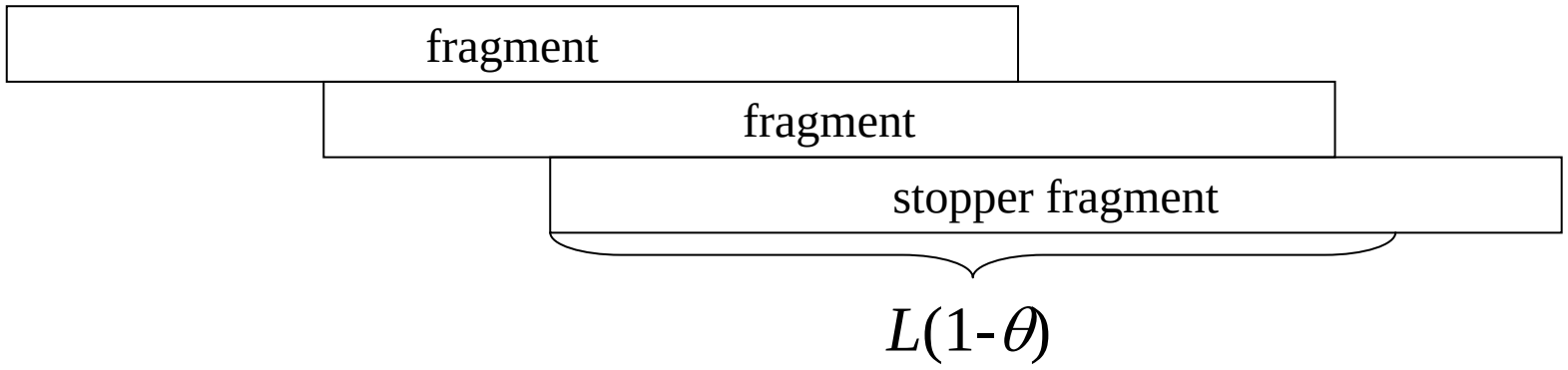


$$\textit{The expected number of contigs is } Ne^{-(1-\theta)\frac{LN}{G}}.$$

Lander-Waterman Statistics （Contig 平均長の推定）

$$\textit{The expected number of contigs is } Ne^{-(1-\theta)\frac{LN}{G}}.$$

A contig stops at the “stopper” fragment.

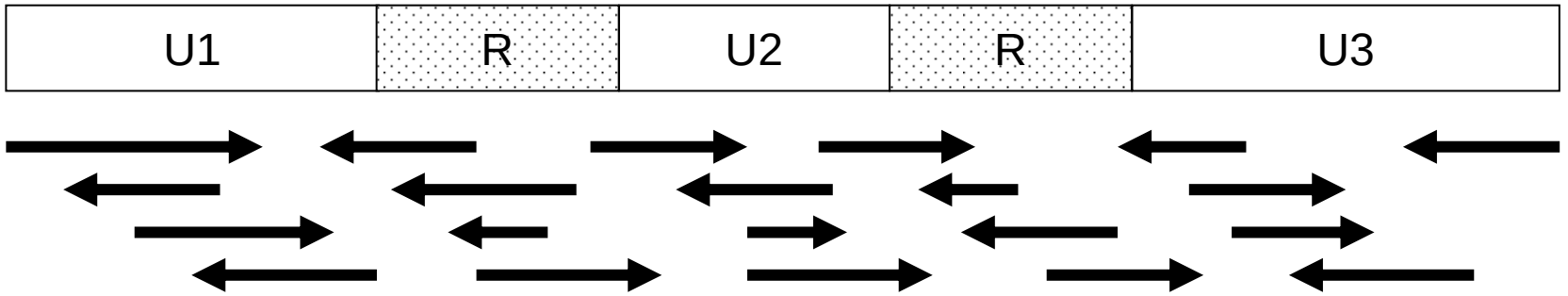


No fragments appear at any of the first $L(1-\theta)$ base pairs.
N/G: Probability that some fragments appear at an arbitrary position.

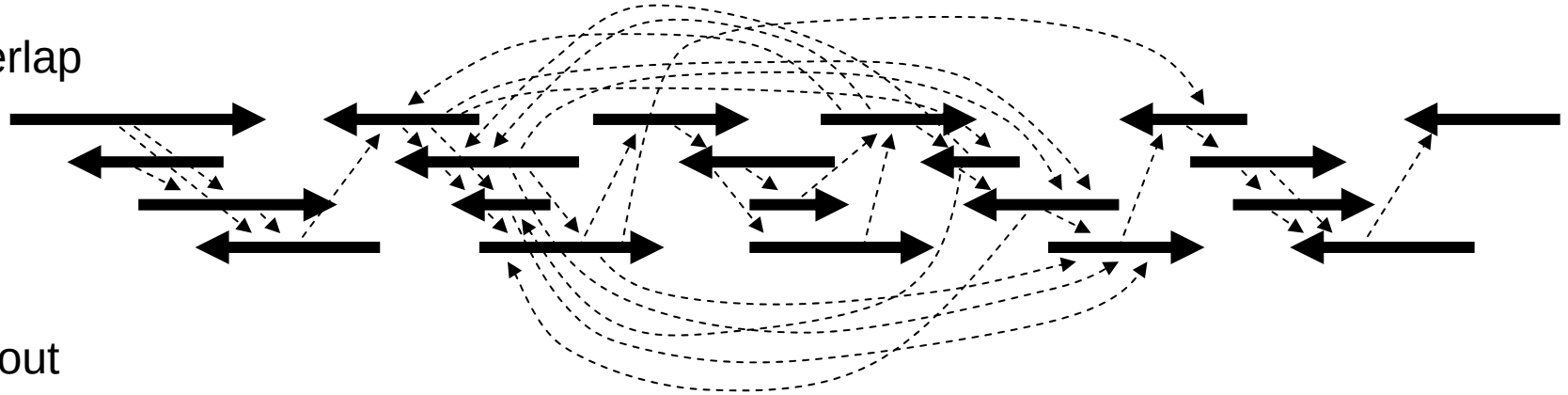
Probability of having a stopper fragment:

$$\left(1-\frac{N}{G}\right)^{L(1-\theta)} = \left\{1+\left(-\frac{N}{G}\right)\right\}^{\frac{1}{\left(-\frac{N}{G}\right)}-\frac{LN}{G}(1-\theta)} \approx e^{-\frac{LN}{G}(1-\theta)}$$

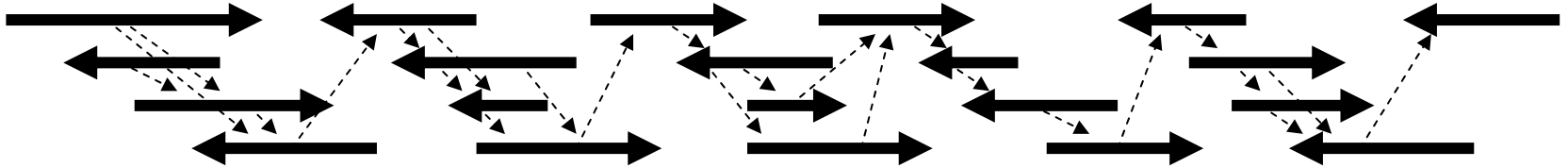
Contig 生成の困難な点



1) Overlap



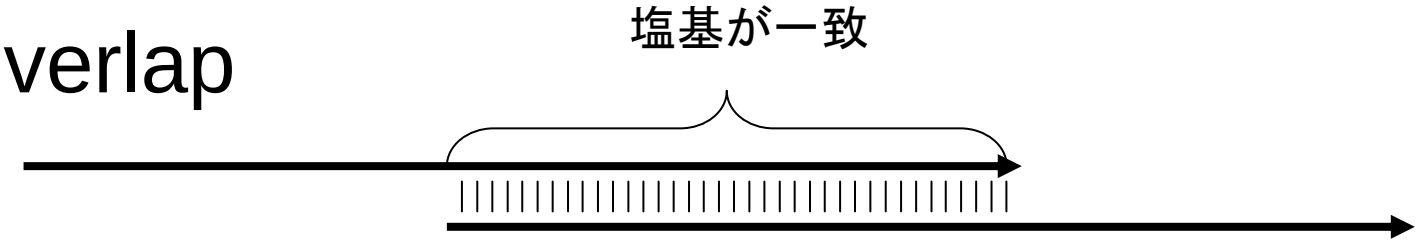
2) Layout



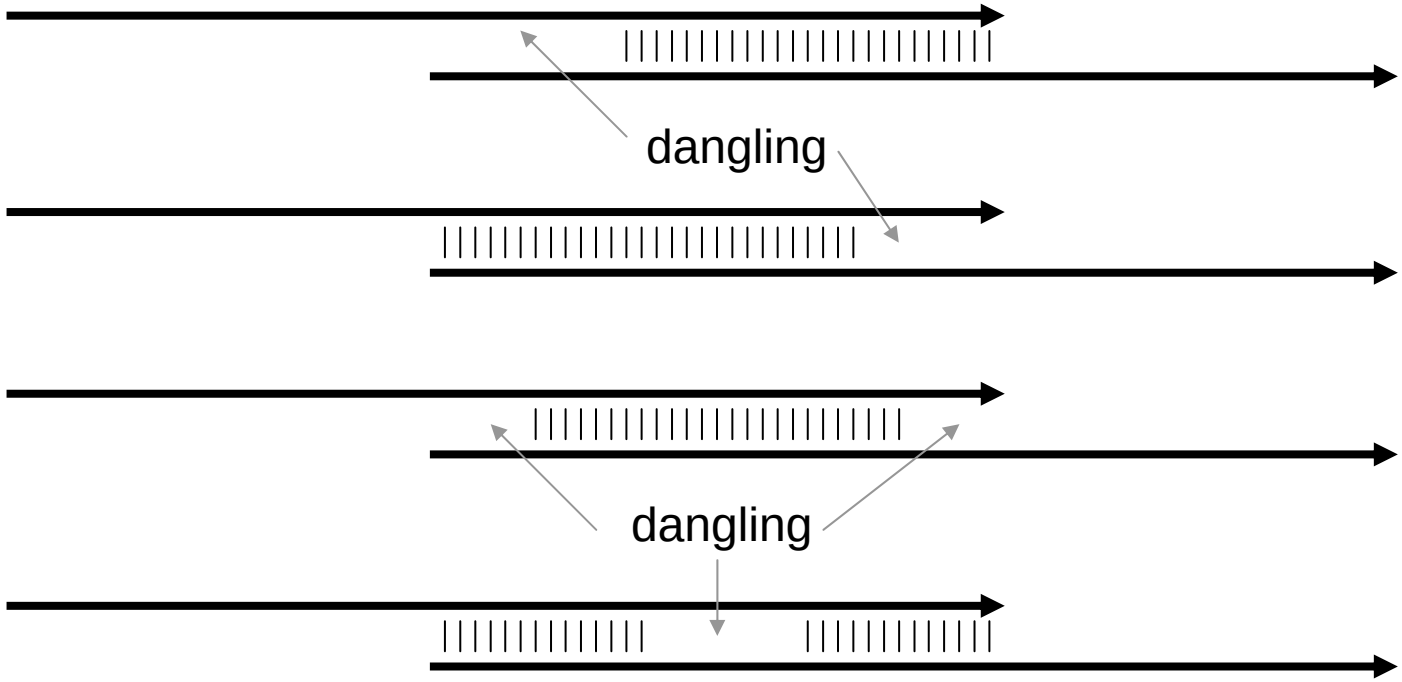
3) Consensus

```
CCTATG-TAGTCAGTCG
    ATGCTAGTCAG
      GCTAGTCGGTCGATCTACC
          CAGTCGATCTGCCGGT
            GTCAGTC-ATCTAC-GGTTAGCATTGC
Consensus CCTATGCTAGTCAGTCGATCTACCGGTTAGCATTGC
```

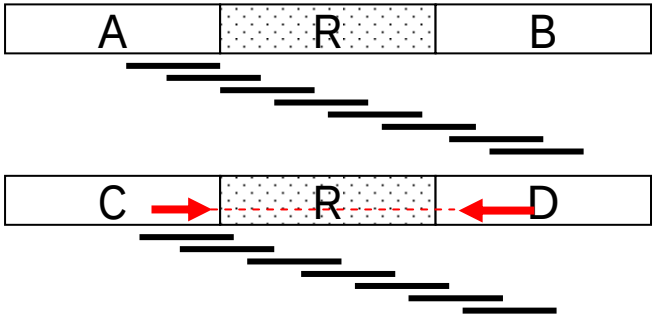
Overlap



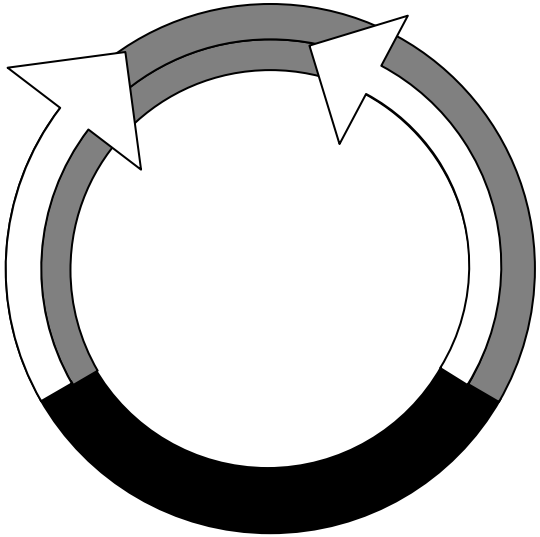
Non-Overlap



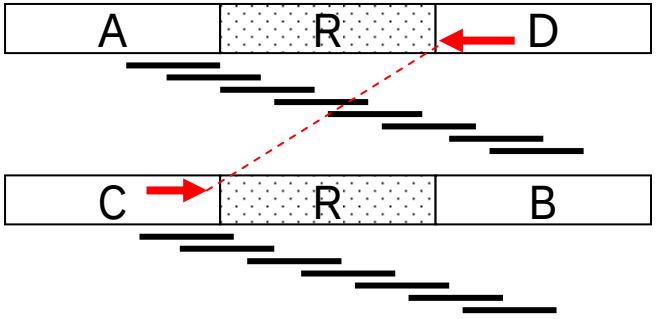
Contig 生成エラーの検出: mate-pair 情報の利用1

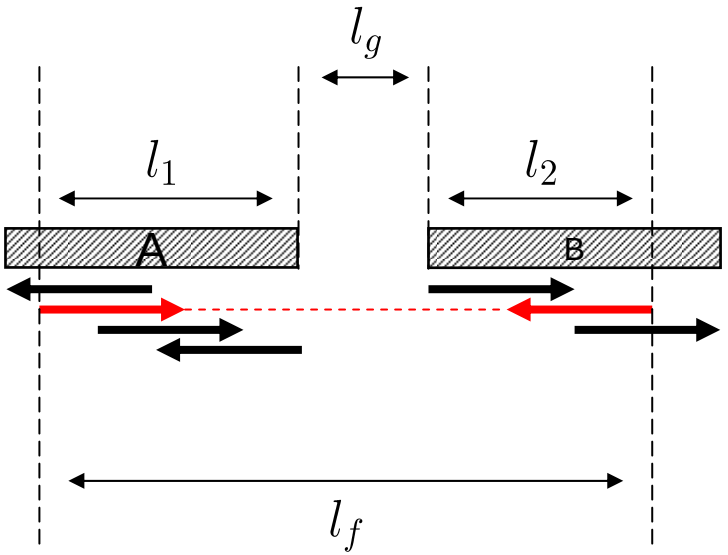


Correct



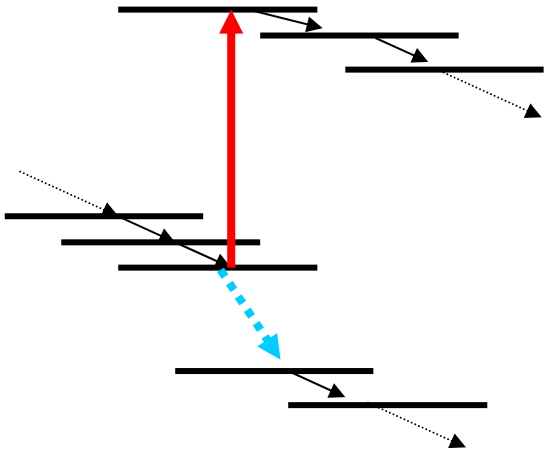
Which is the correct layout?
Are A and B linked?



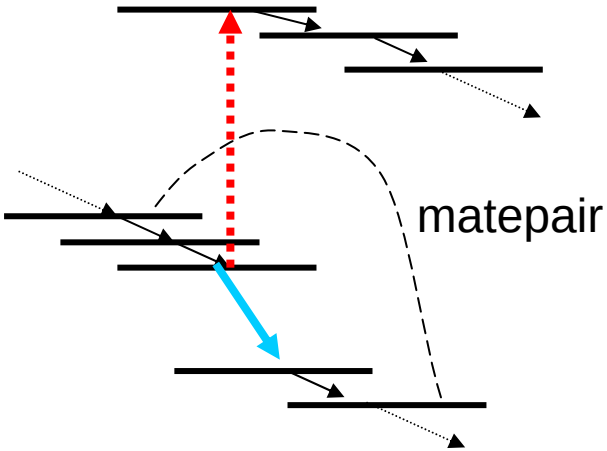


Contig 生成エラーの検出 : mate-pair 情報の利用2

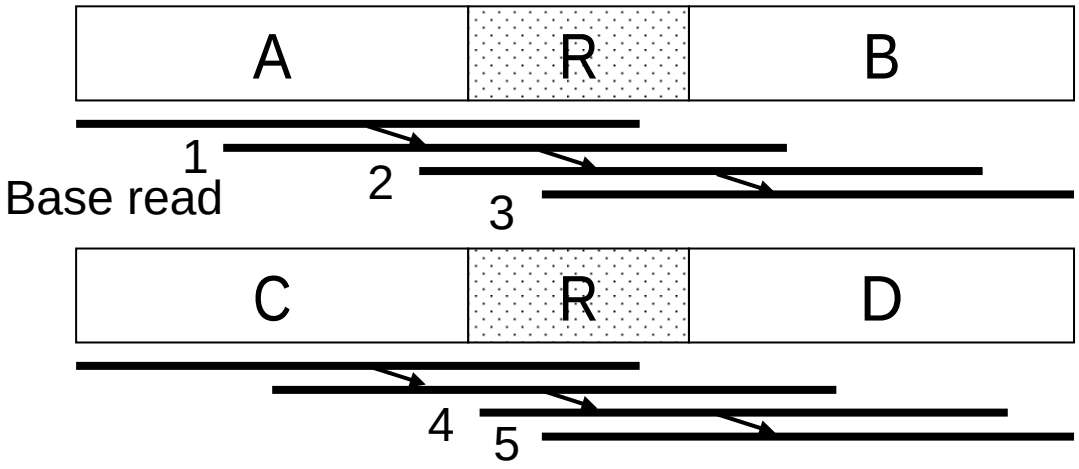
Misjoin by better
alignment scores



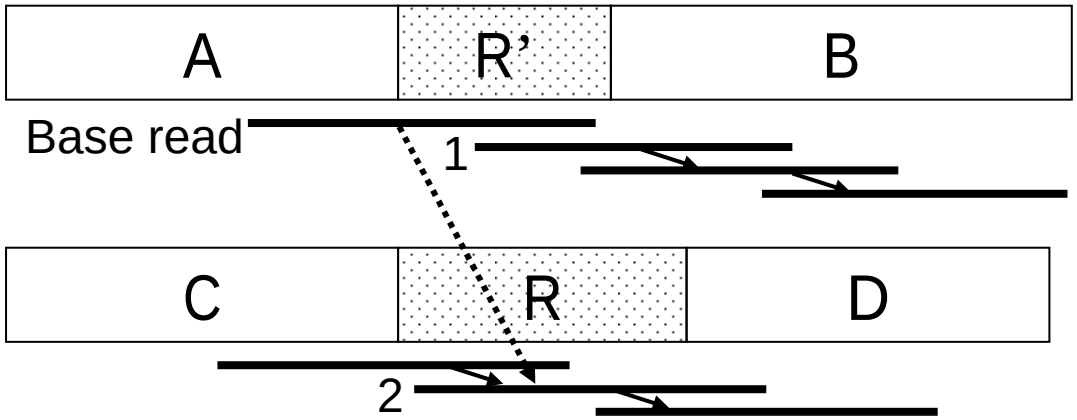
Revision by matepair



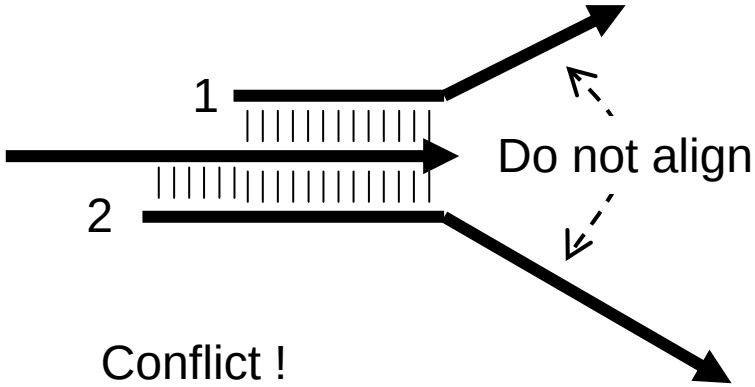
Contig 生成エラーの検出: 矛盾の検出



Small repeat sequences



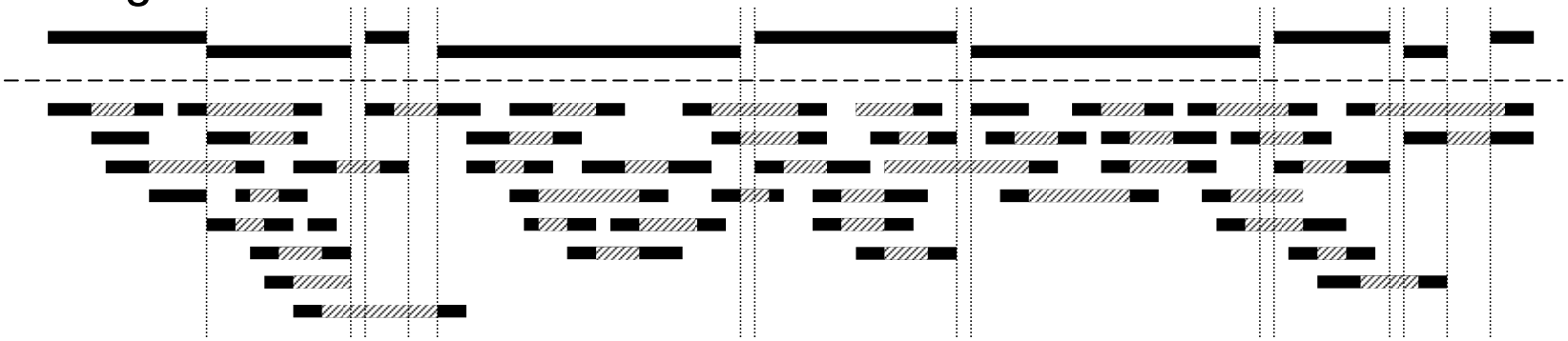
A repeat R and a truncated repeat R’,
e.g. incompletely retro-transposed elements



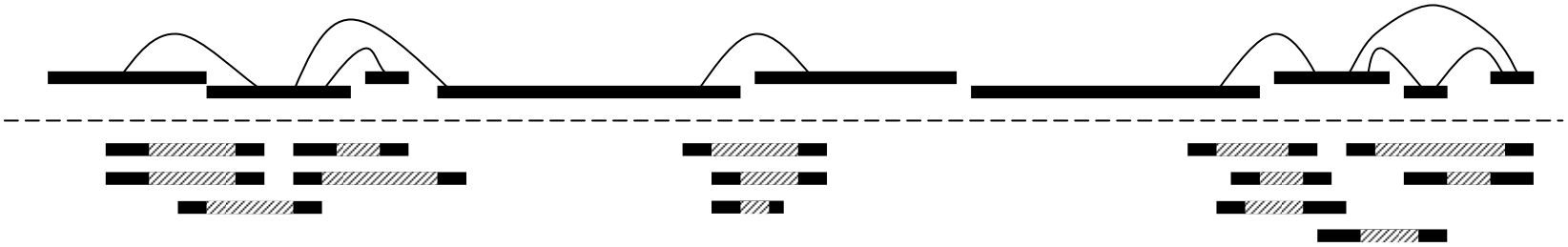
Conflict !
安全をみて Contig を
これ以上伸ばさない

Scaffold の生成

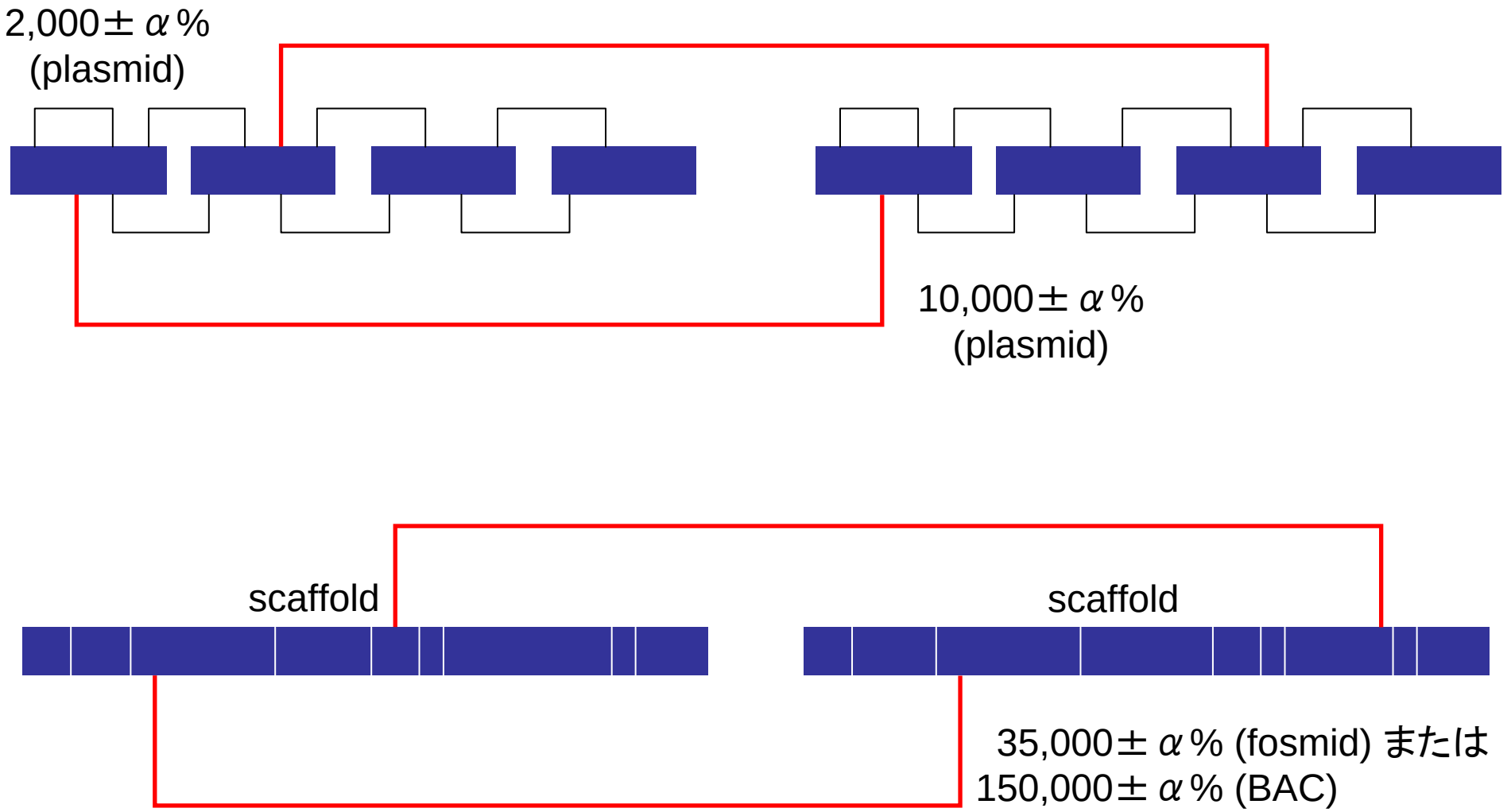
e) Contigs



f) Scaffolds(Super contigs)



Mate-pair を使った scaffold の構築



どのぐらいの量の mate-pair 情報が必要か？

クローンのタイプ	mate-pair 間の平均長 (L)	ゲノムカバー率
プラスミド	5 ~ 10 kb	10~20
フォスミド	40 kb	10
BAC	130 ~ 200 kb	10 ~ 20

* ゲノムカバー率＝ L × (mate-pair の個数) / ゲノムサイズ

ゲノムの完成度を測る指標は？

Scaffold N50値： 50%以上の塩基がN50値以上の長さの scaffold に含まれる。少なくとも1M塩基以上、5M以上が望ましい。

染色体被覆率：染色体の塩基のうち scaffold に含まれることが判明している割合。90% 以上が望ましい。