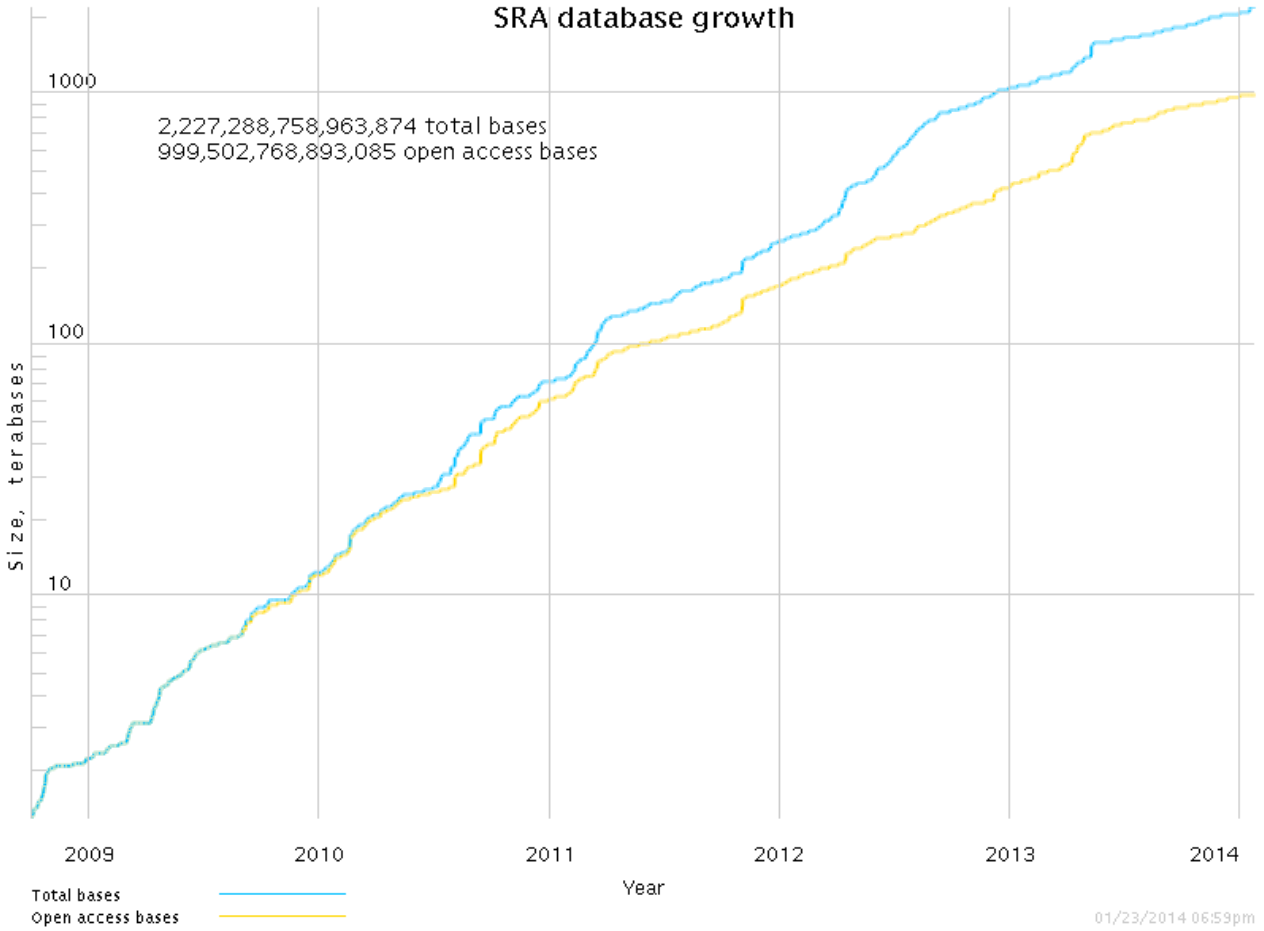
配列データ解析とハードウェア

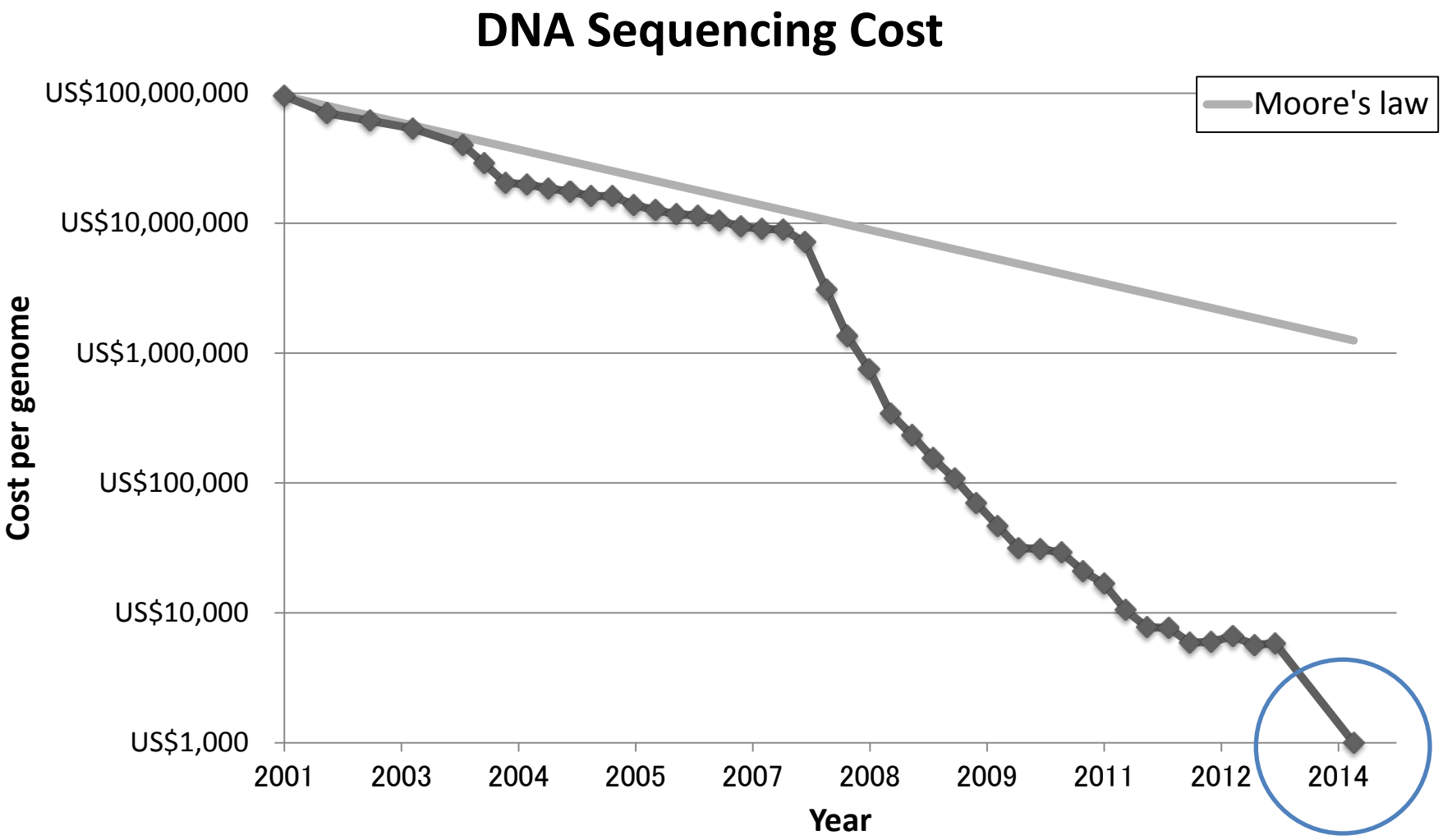
角田将典
東京工業大学　情報理工学研究科

SRAデータベースの規模の推移



NCBI Sequence Read Archive: http://www.ncbi.nlm.nih.gov/Traces/sra/

Accessed Jan 24, 2014



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed Jan 12, 2014.

Illumina Hiseq X Ten Sequencing System

http://www.illumina.com/systems/hiseq-x-sequencing-system.ilmn

Accessed Jan 21, 2014

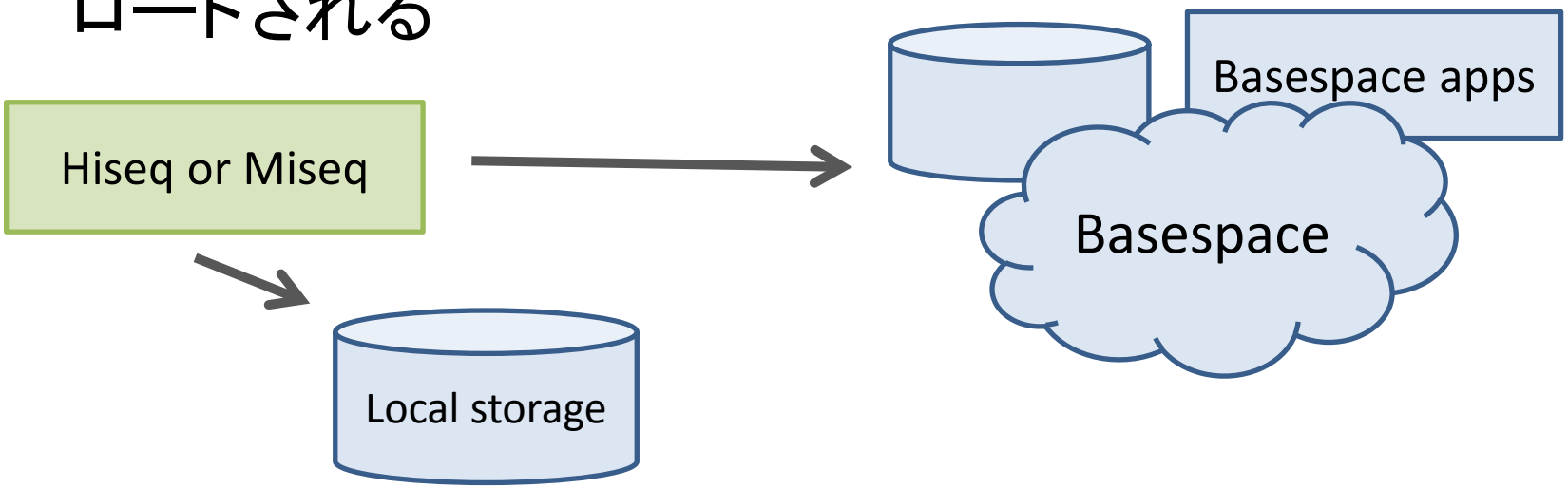
Illumina Hiseqシリーズ

Platform	Hiseq X	Hiseq 2500	
		High-Output Run	Rapid-Run
Output/Run	1.6-1.8 Tb	900 Gb-1 Tb	150-180 Gb
Run time	3 days	6 days	40 hours
Output/Day	530-600 Gb	150-160 Gb	90-110 Gb
Read length	2 x 150	up to 2 x 125	up to 2 x 150
Read passing filter	≤6 billion	≤4 billion	≤600 million

http://res.illumina.com/documents/products/datasheets
Accessed Jan 21, 2014

Basespace

- Illuminaの提供するAmazon Web Services上に構築されたクラウド環境
  - シーケンシングの監視、データの保存や共有、解析が可能
  - データはシーケンシングの進行に伴い逐次アップロードされる



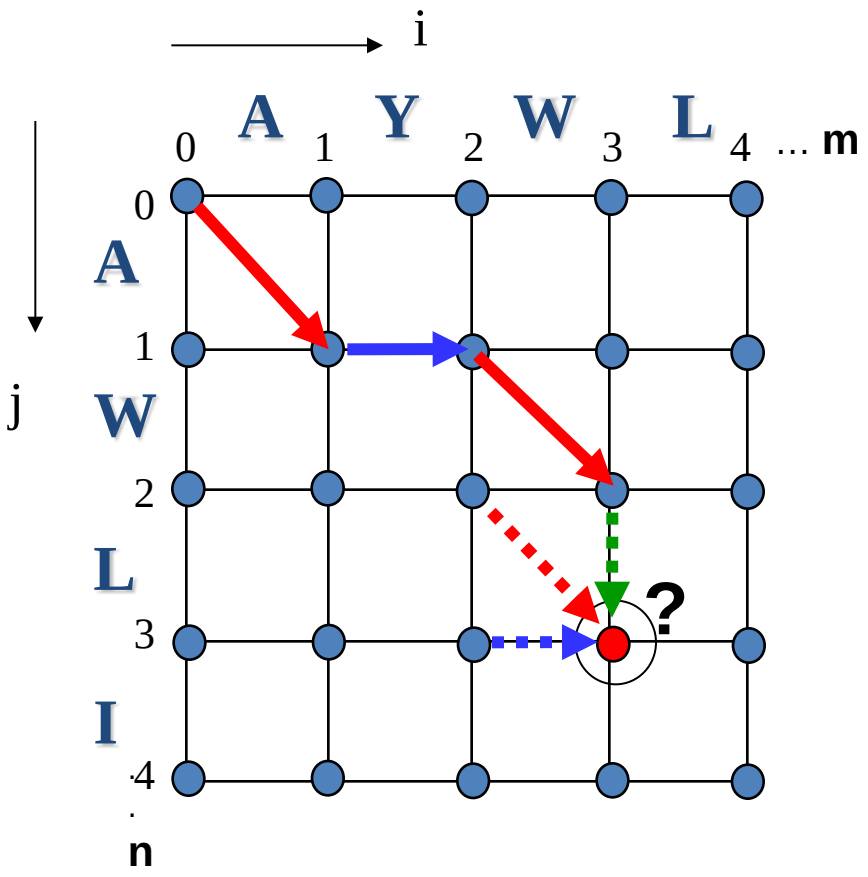
次世代シーケンサーで得られるデータ

アラインメントとアセンブリ

- 次世代シーケンサーによってシーケンシングされた短い断片配列から有用な情報を得るために
  - 参照配列（シークエンス）へのアラインメント
  - de novo assembly

動的計画法による配列アラインメント

Local Sequence Alignment by Dynamic Programming



for i>0, j>0,

$$M[i,j] \leftarrow \max \begin{cases} M[i-1,j] + w \\ M[i,j-1] + w \\ M[i-1,j-1] + S[i,j] \\ 0 \ \text{(discard \& reset)} \end{cases}$$

ただし S[i, j] は alignment score **実数／整数**
w は gap penalty constant (<= 0),
M[i, j] はその点までの**累積スコア値**

境界のInitialization

$M[i,0] \leftarrow 0$

$M[0,j] \leftarrow 0$

最もナイーブな解法では、
O (m n) の時間・空間複雑度

配列アラインメントの高速化

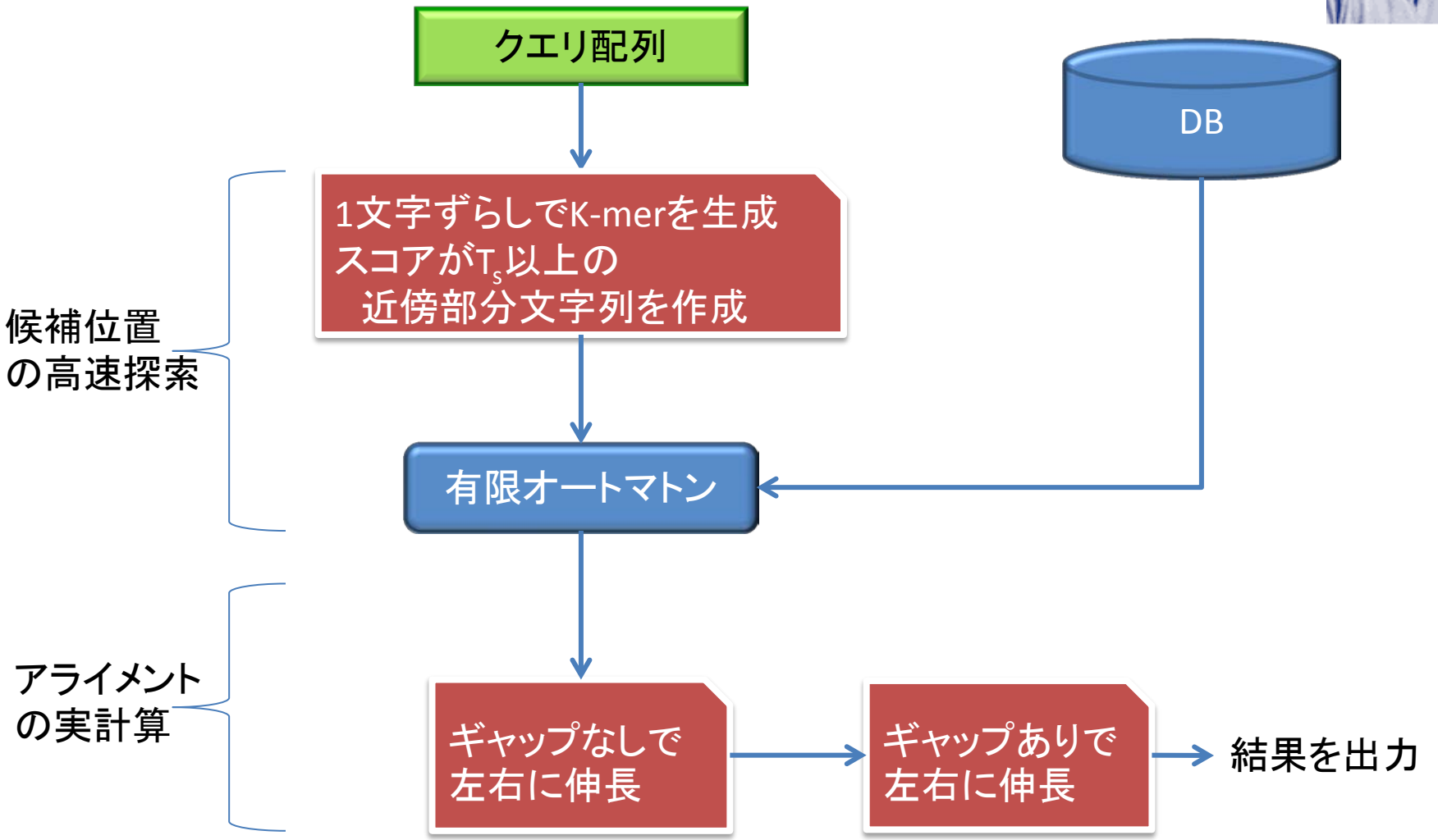
- SIMD
- ヒューリスティクス
- 効率的なデータ構造
  - 索引の利用
  - FM-index, Suffix Array

SIMDによる高速化

- Wozniak Comput. Appl. Biosci. 1997,13:145-150
- Rognes and Seeberg, Bioinformatics 2000, 16:699-706
- Farrar, Bioinformatics 2007,23:156-161

ヒューリスティクスによる高速化

BLAST (Altschul, *et al.*, JMB, 1990)



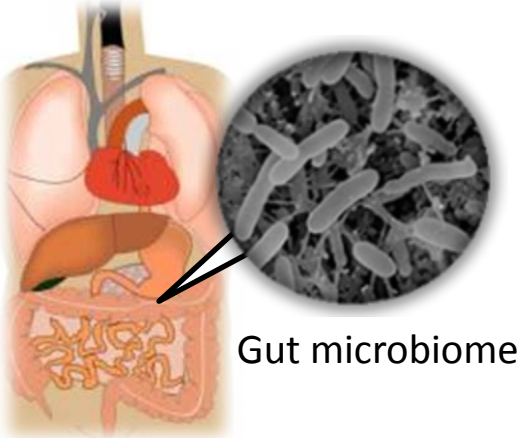
注： BLASTXは、上記にDNA配列→アミノ酸配列（6フレーム）の変換機能を加えたもの

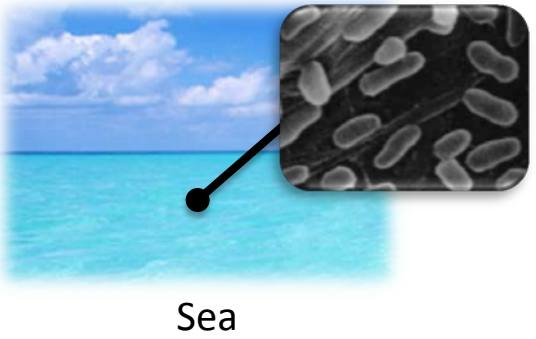
索引の利用

- ハッシュテーブル
  - MAQ, BFAST, soap
- FM-index
  - BWA, bowtie, soap2

メタゲノム解析

- メタゲノム
  環境中に存在する微生物から分離・培養を行わず直接DNAを分離し、解析を行う
  - 難培養性微生物からの新規遺伝子の発見
  - 微生物種・遺伝子組成と環境との相互作用
  - サンプル中存在する微生物の塩基配列そのものが、アノテーション済みの配列データベース中に含まれていることは稀であるため、類縁微生物の配列とのアラインメントが必要になる

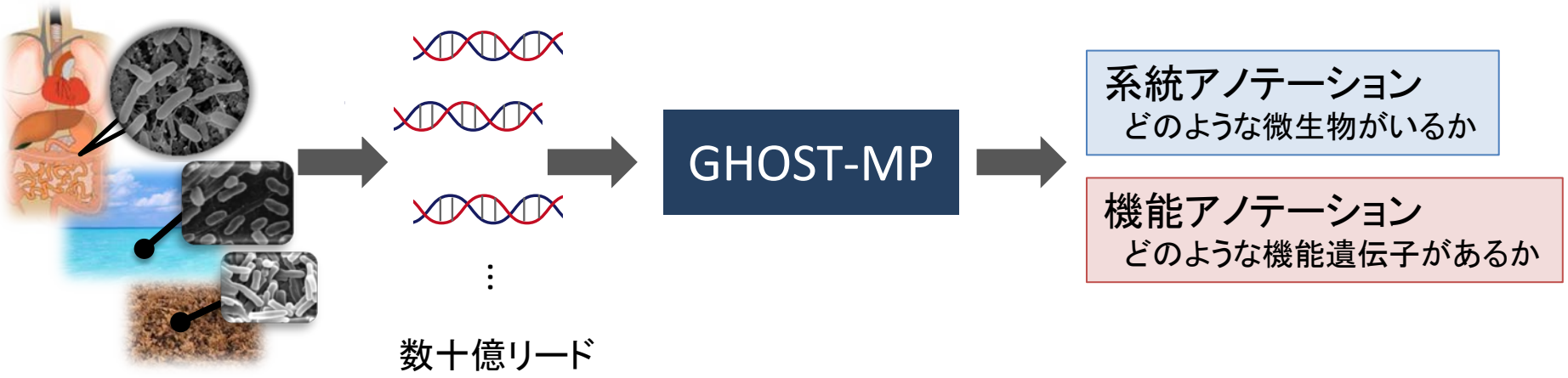




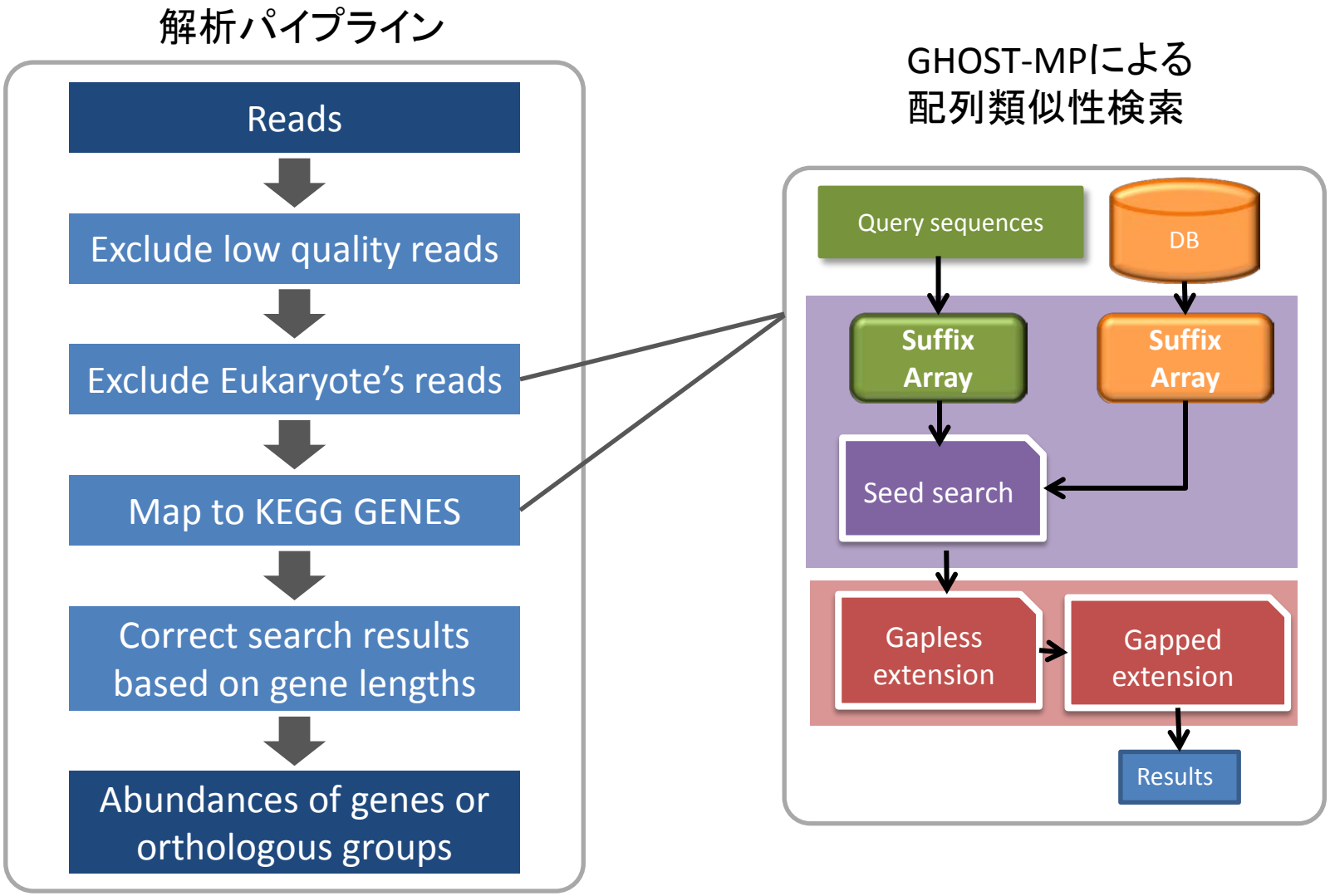
H24成果

メタゲノム解析パイプラインの開発

- 次世代シーケンサによる大量のメタゲノムデータの超高速解析のためのパイプライン（GHOST-MP v1.2.3, 2013/03, H24末公開予定）
  - 相同性検索（アノテーション済みの類似配列から機能推定を行う）によってメタゲノムデータに対しアノテーションを行う
  - メタゲノム解析では、遠縁の相同配列を検出できる高感度な検索が要求される
- 従来の相同性検索では対応できない高感度な大量データの処理を、検索アルゴリズムと「京」での大規模並列化によって解決する



メタゲノム解析パイプライン



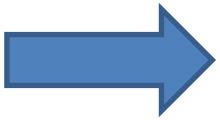
Suffix Array（接尾辞配列）

- 辞書順にソートされた接尾辞の開始位置の配列（Manber, 1989）

T = abracadabra$

0:	abracadabra$
1:	bracadabra$
2:	racadabra$
3:	acadabra$
4:	cadabra$
5:	adabra$
6:	dabra$
7:	abra$
8:	bra$
9:	ra$
10:	a$
11:	$

sort

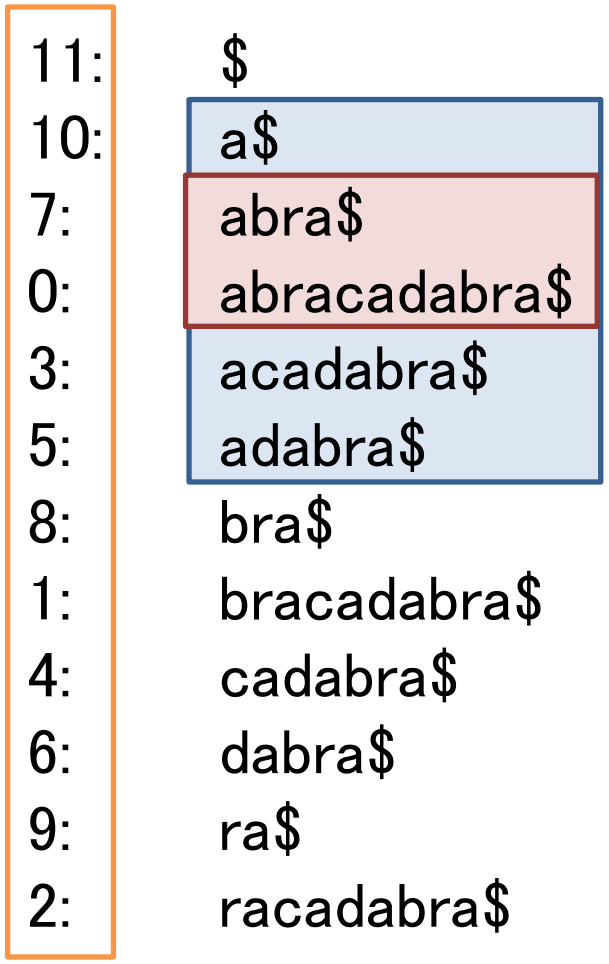


Suffix Array

11:	$
10:	a$
7:	abra$
0:	abracadabra$
3:	acadabra$
5:	adabra$
8:	bra$
1:	bracadabra$
4:	cadabra$
6:	dabra$
9:	ra$
2:	racadabra$

Suffix Array（接尾辞配列）

- 特定の文字列で始まる範囲（全ての開始位置）を2分探索によって効率的に探索可能
  - abrで始まる文字列を検索
- DB側のSAは常に同じであるため、SA構築時に先頭数文字分の結果を記録しておくことで、さらなる高速化が可能



まとめ

- 次世代シーケンサと産出されるデータ
- 配列アラインメントの高速化の紹介
- メタゲノム解析パイプライン
  - 口腔内メタゲノムへの応用