# PCクラスタを用いた大量ゲノムマッピング

理研ゲノム科学総合研究センター
遺伝子構造・機能研究グループ
長谷川　哲

# 1）遺伝子構造･機能研究チーム

- 「マウス遺伝子」の研究がメイン

- 「FANTOM3 (Functional Annotation of Mouse)」（September 2005, Science）

- 全長cDNA - 10,000クローン以上
- CAGE –合計 20,000,000タグ以上

# 2）技術者の仕事

- シーケンスされた塩基配列を、チェックし、加工する（アセンブルチーム）
- 研究者のサポート
- 依頼された計算を、期限内に提出する。
- 研究者は結果を元に、研究の改良、論文の執筆を行っている。

- 例：プライマーの設計、マッピング、簡単な統計。

# ３）マッピングの仕事

- チームそれぞれが別々の分野を担当する。
- マッピングを任されている。
- シーケンスされた塩基配列が、マウスの染色体のどの領域に当たるかを調べる。

- 塩基配列のQuality Control
- 既知の遺伝子との関連性。
- エクソン、イントロンのチェック。

# 4）BLAST

- Basic Local Alignment Search Tool
- 一番良く使われているアライメントツール

- 例：

```
Query              GATCTAGGCTACC
                   |||||||||||||  |
Genome ATCTGATCTAGGCTAGCCCAAGTT
```

# ５）コンピューター言語

- perlが一番よく使われている。
- テキスト処理が多い
- 細かい作業、手直しが多い。

- C言語（プログラム）＋perl＋shellの組み合わせが一般的。

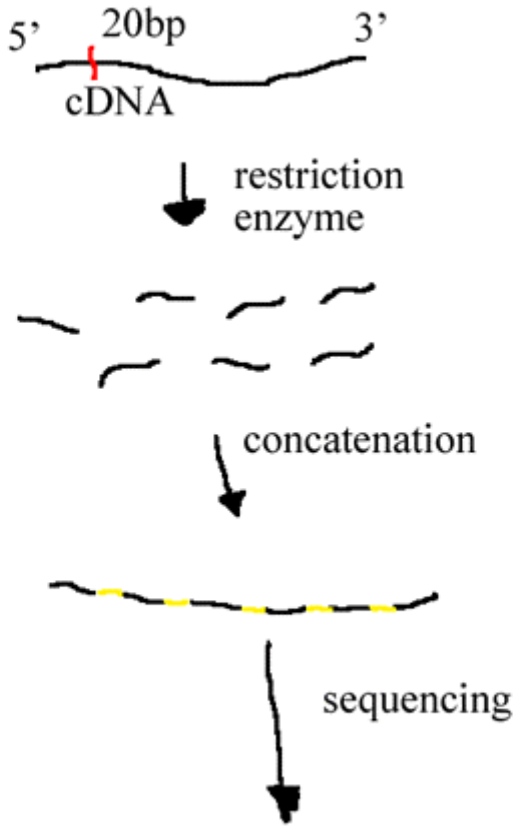
# ６）コンピューター環境

- rlgsw1 - SPARC 1200MHz x 6
- rlgsw11 - インテル 2.7GHz x 4
- rlgs9 - Beowulf 800 MHz x 32
- 個人のコンピューター - DELL 1.1GHz

- 複数バイオインフォマティシャンで使う。
- 計算が間に合わない、大量の場合のみ、RSCCを使う。

# ７）CAGE

- CAGE (Cap Analysis Gene Expression)
- ５‘から20bpを切り出したのシーケンス。 タグと呼ぶ。
- １０以上のタグをつなげ、シーケンサーで読む。
- ESTに比べ、タグの回収の効率が良い。


5’
20bp
3’
cDNA
restriction
enzyme
concatenation
sequencing

# 8）CAGE

- 生体組織や発生段階ごとの遺伝子の発現パターン
- 転写開始位置（TSS: Transcription Start Site）

# 9）CAGE

- 2004年の２月から7月まで、約一日に10万タグ近く読まれていた。
- マウスCAGEタグ数：1200万タグ
- ヒューマンCAGEタグ数：1800万タグ（現在）
- 大量のタグをマップするのにRSCCを使用する。

# 10）マッピングに必要な日数

- マウスCAGE　1200万タグ
- 全染色体（1～19、X,Y）にBLAST

| | | |
|---|---|---|
| 個人のコンピューター | 1.1GHz x1 | 1,947.9日 |
| 旧クラスター<br>32ノード | 800MHz x32 | 83.7日 |
| RSCC<br>50ノード | 3GHz x50 | 14.3日 |

# 11)改良点

- 長さ＋塩基が重複しているタグは、再計算する必要はない。
- ユニークなシーケンスだけをマップする。

- ほぼ毎日タグがシーケンスされる。
- 過去にシーケンスされたタグを（ハッシュ）データベース化しておき、新しいタグだけをマッピングにまわす。

# 12)マッピングに必要な日数(改)

- マウスCAGE　3,822,196 タグ
- 全染色体にBLAST

| | | |
|---|---|---|
| 個人のコンピューター | 1.1GHz x1 | 617.7 日 |
| 旧クラスター<br>32ノード | 800MHz x32 | 26.5 日 |
| RSCC<br>50ノード | 3GHz x50 | 4.5 日 |

# 13)タグのセレクション

- BLASTの出力から、一番きれいにマップされた箇所を選ぶ。
- 20bp中16bp以上連続してミスマッチなしにマップされた箇所。
- 一番長くマップされたものを選択する。

# 14)マップの種類

- Single map - 全染色体で一カ所だけマップされたタグ（約60%）
- Multi map - 全染色体で複数マップされたタグ（約10%）
- Unmapped - マップされなかったタグ（約30%）

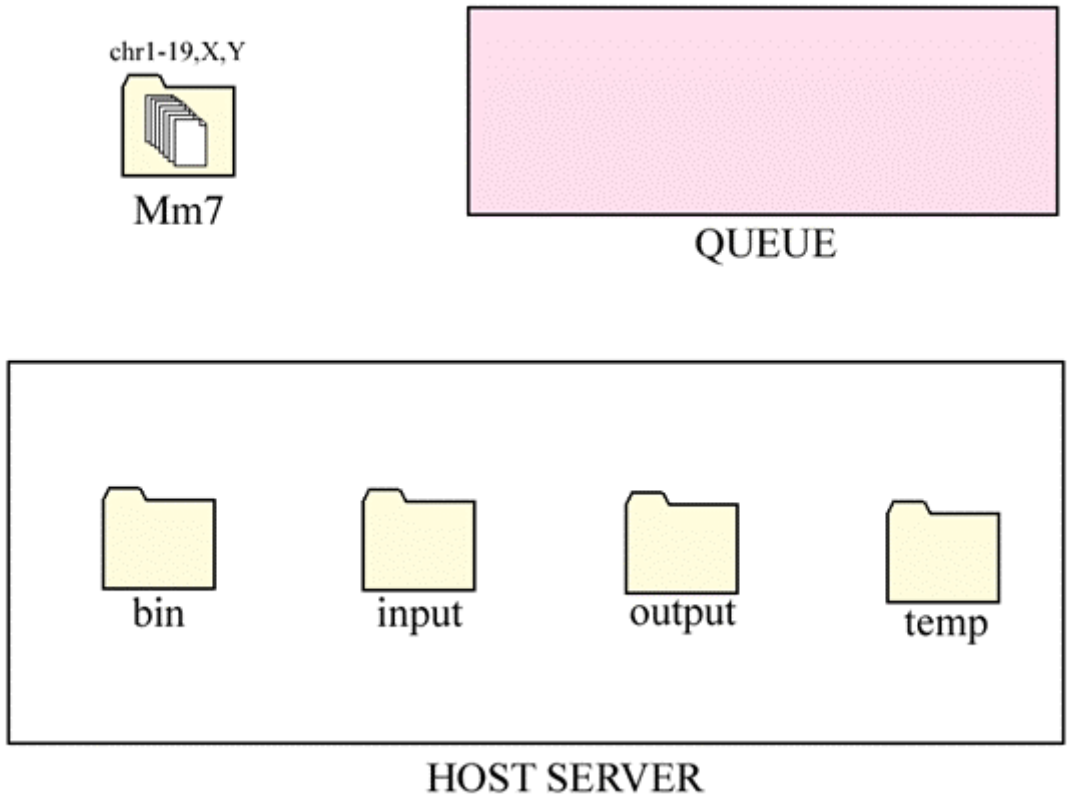
# 15) マッピングの自動化

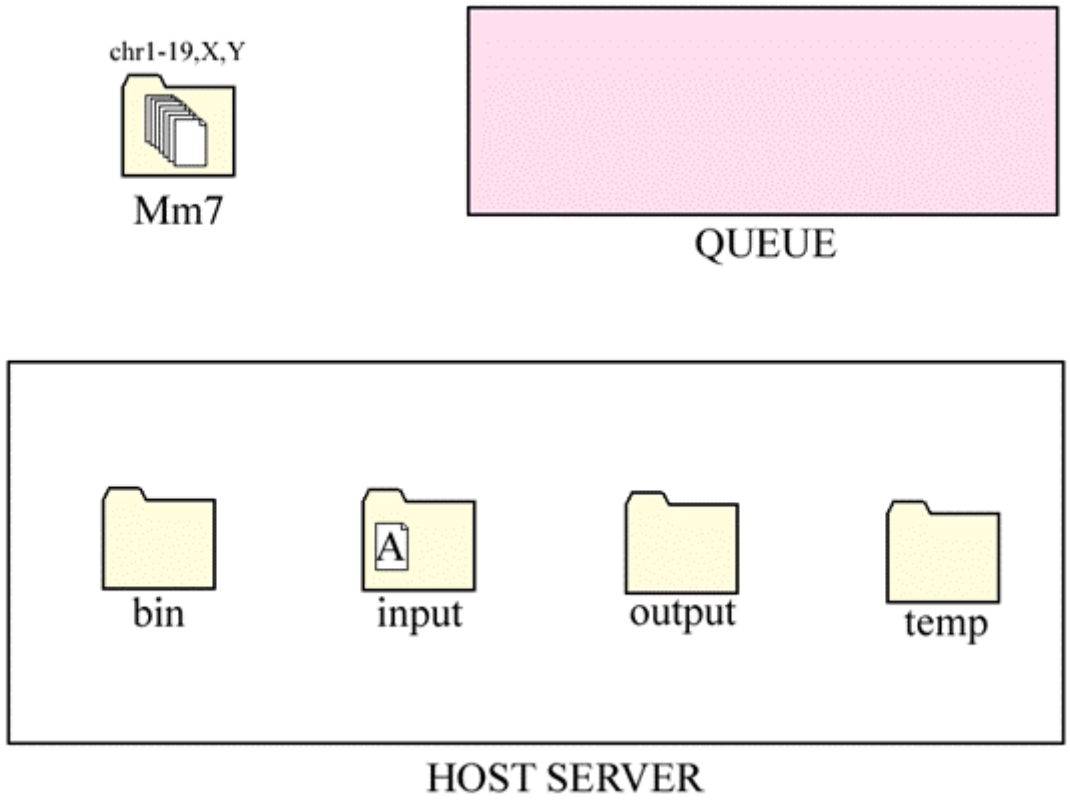
- 毎日、塩基配列が大量にシーケンスされる。
- マウスのゲノム以外にもマップする必要がある。
- 複数の染色体に当てるため、タイプするコマンドがあまりにも多い。

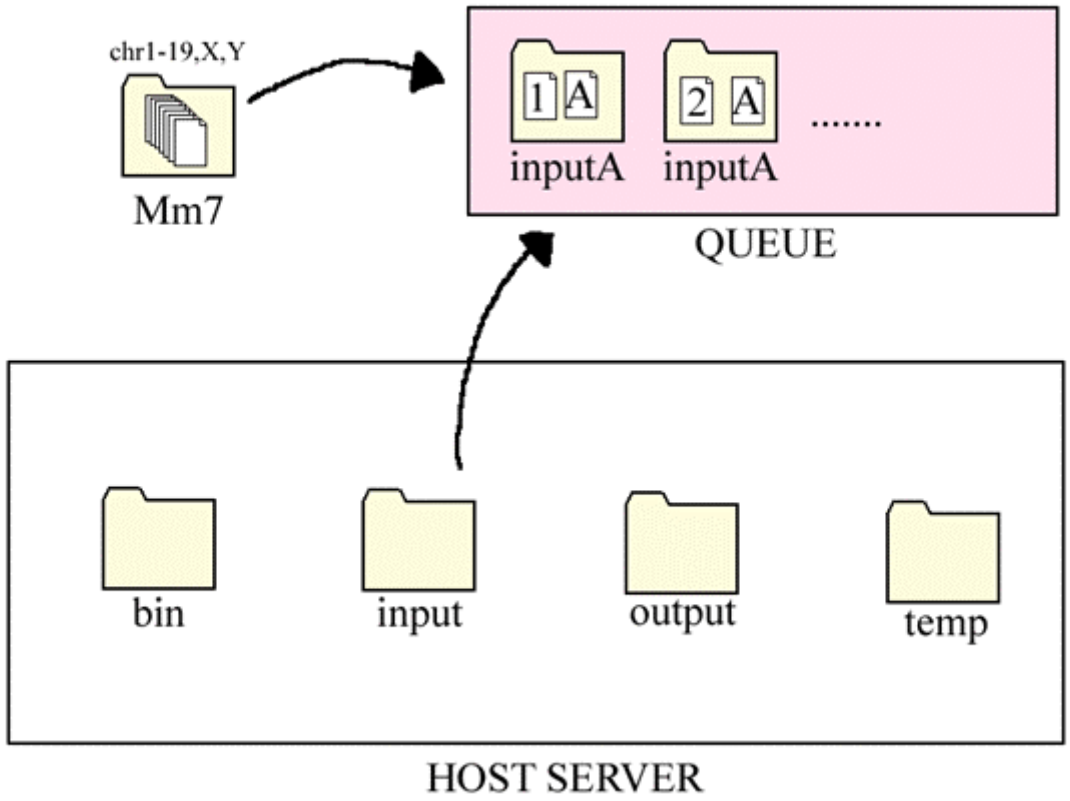
マッピンングの自動化が必要

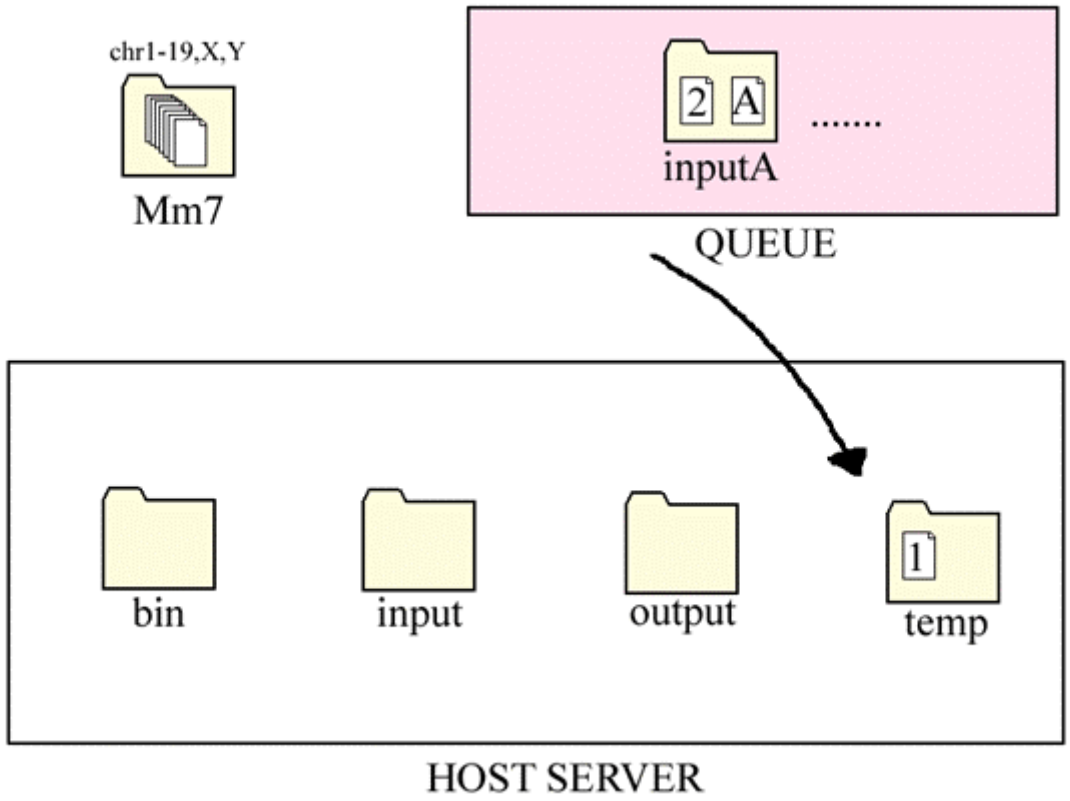
# 16)マップ自動化スクリプト

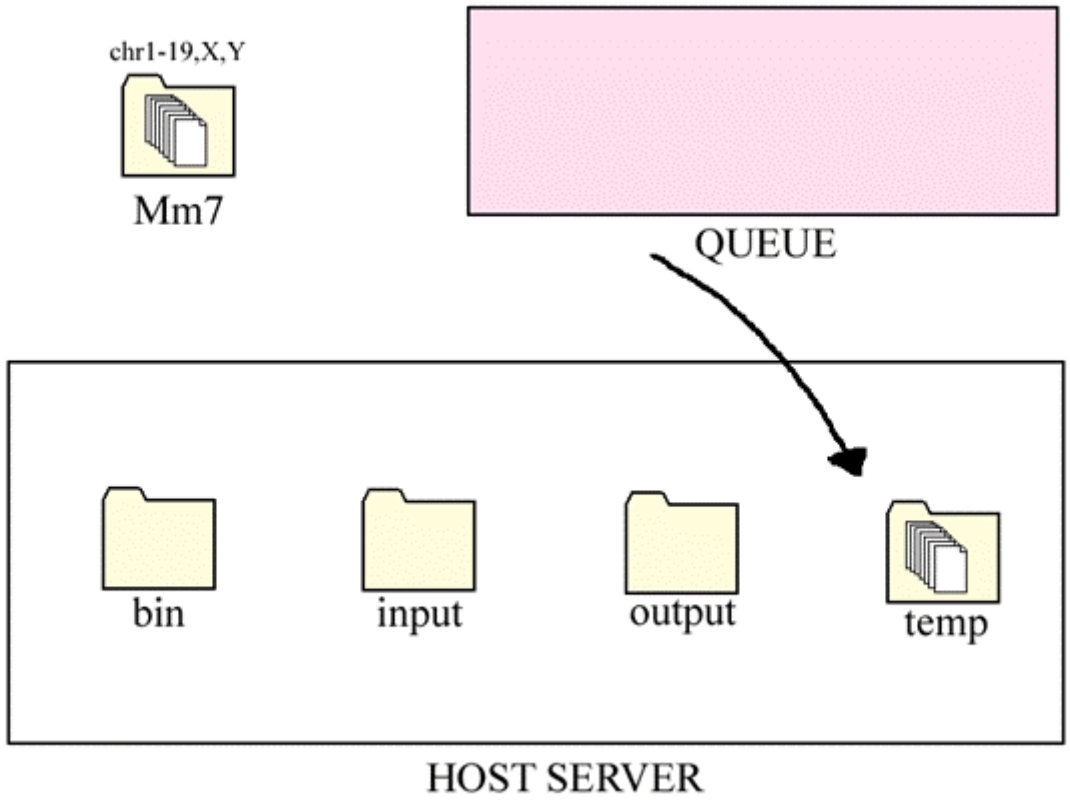
- 10分おきにinputフォルダーをチェックする(background)。
- 新しいファイルがあり、ノードが空いている場合、ブラストを開始する(qsubで投げる)。
- 全染色体にマップし終わった後、セレクションにかけ、一つのファイルにまとめ、outputフォルダーに移動させる(background)。
- 全データベースで繰り返す。

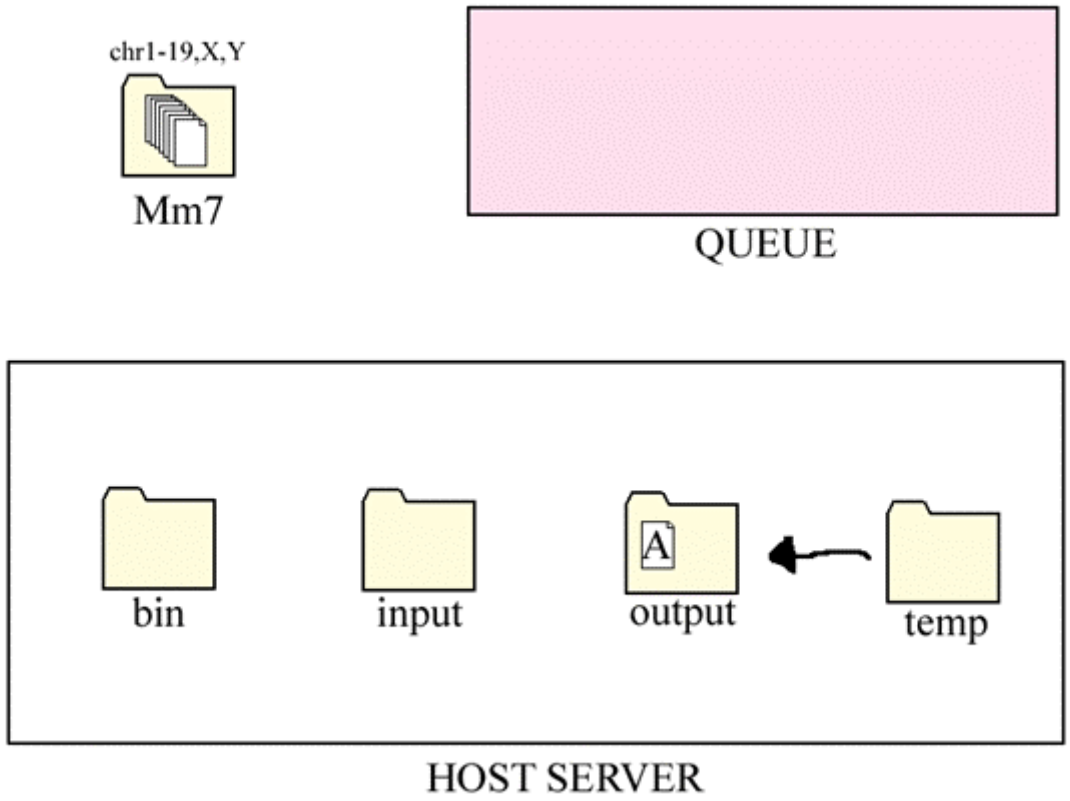

chr1-19,X,Y
Mm7
QUEUE
bin
input
output
temp
HOST SERVER

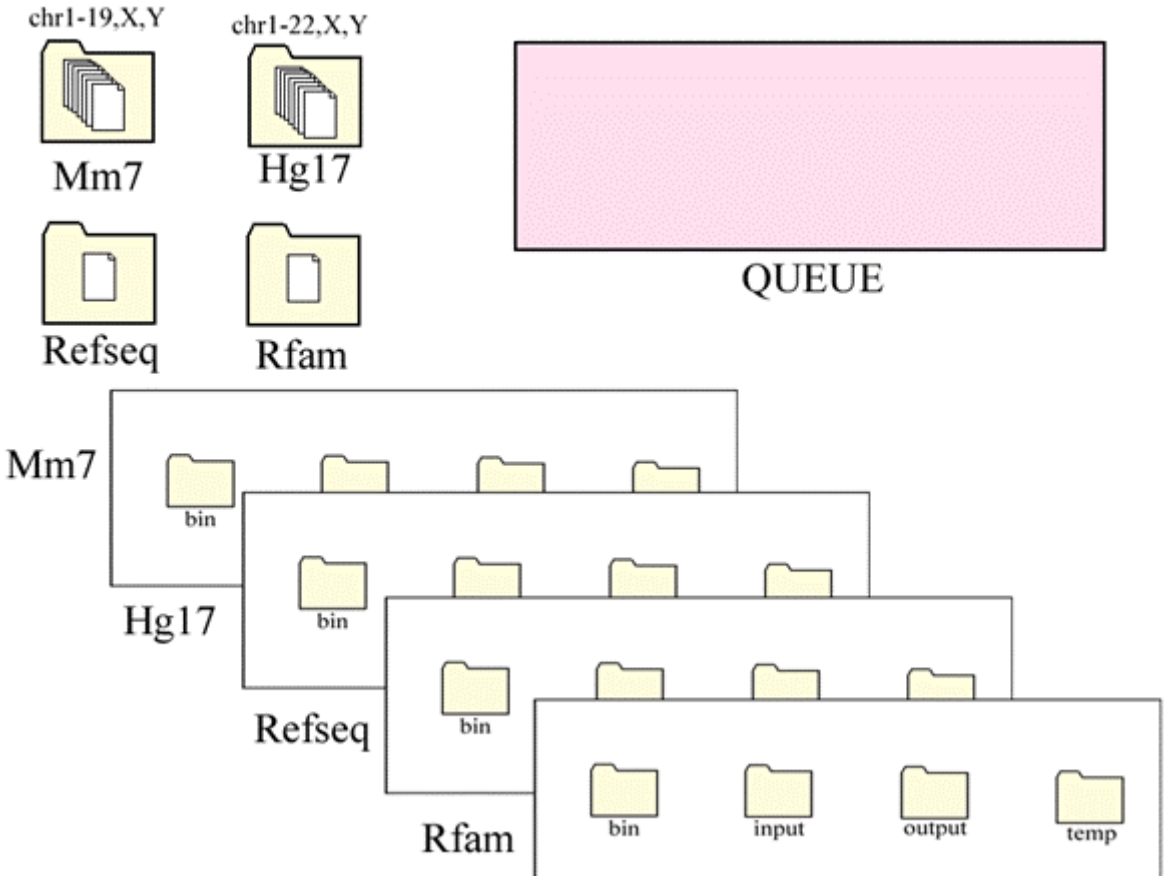

chr1-19,X,Y
Mm7
QUEUE
A
bin
input
output
temp
HOST SERVER

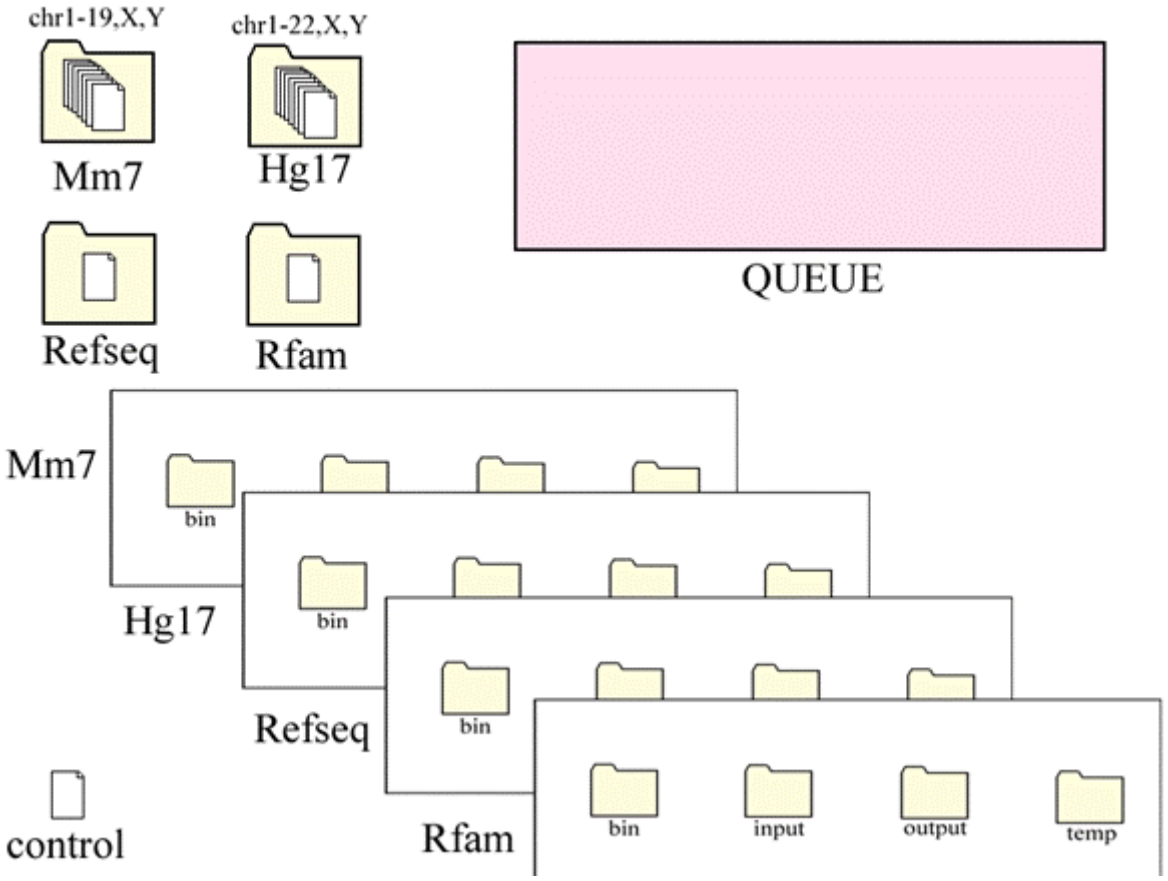

chr1-19,X,Y
Mm7
1 A
2 A
.....
inputA
inputA
QUEUE
bin
input
output
temp
HOST SERVER

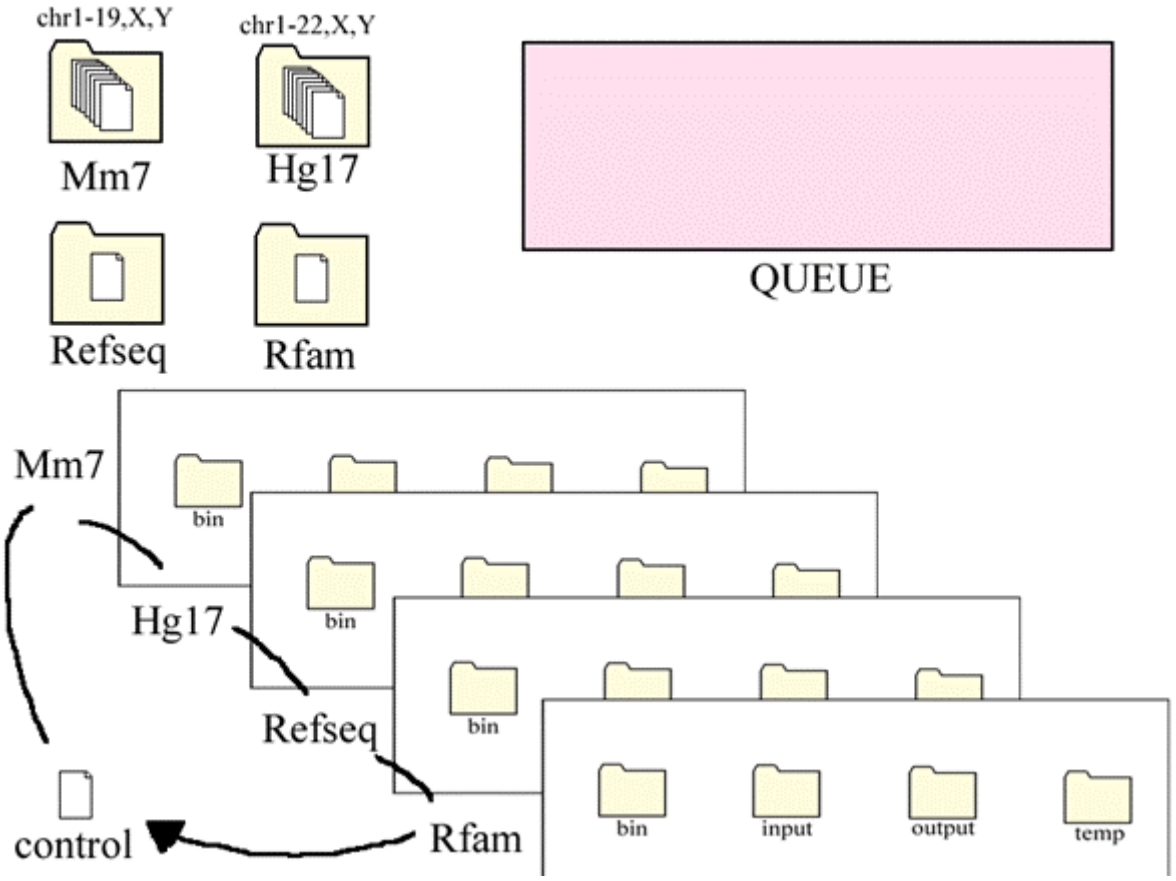

chr1-19,X,Y
Mm7
2
A
.....
inputA
QUEUE
bin
input
output
1
temp
HOST SERVER


chr1-19,X,Y
Mm7
QUEUE
bin
input
output
temp
HOST SERVER


chr1-19,X,Y
Mm7
QUEUE
bin
input
output
temp
HOST SERVER


chr1-19,X,Y
chr1-22,X,Y
Mm7
Hg17
Refseq
Rfam
QUEUE
Mm7
bin
Hg17
bin
Refseq
bin
Rfam
bin
input
output
temp


chr1-19,X,Y
chr1-22,X,Y
Mm7
Hg17
QUEUE
Refseq
Rfam
Mm7
bin
Hg17
bin
Refseq
bin
Rfam
bin
input
output
temp
control


chr1-19,X,Y
chr1-22,X,Y
Mm7
Hg17
Refseq
Rfam
QUEUE
Mm7
bin
Hg17
bin
Refseq
bin
Rfam
bin
input
output
temp
control

# 16)現在

## 問題点

- BLASTを用いたマッピングは、あまりにもコンピューターリソースを食うので、人に多大な迷惑をかける。

## 解決策

- Suffix Arrayを用いてCAGEタグをマッピングする。
- 全CAGEタグのマッピングを普通のサーバーコンピューター(1.2GHz)で、一晩での計算が可能。
- 現在Suffix Arrayを用いての計算を検討中。

# 17)将来

- 来月からは、今まで以上にCAGEタグがシーケンスされる予定。
- 新しいシーケンサーの改良の為、素早いフィードバックが必要。

- ウェッブのフォームからもタグを入力でき、自動的にマッピング結果をHTMLで表示できるようにしたい。

# 18)Acknowledgement


- ゲノムネットワーク
- 理事長ファンド機能性RNA

- アッセンブルチームの皆さん