H-InvDB
Annotated Human Gene Database

BIRC
Biomedical Information Research Center

2008年12月19日

# トランスクリプトームの一括アノテーションのための自動解析システム SuperTACTの紹介


山崎　千里

chisato-yamasaki@aist.go.jp

産業技術総合研究所　バイオメディシナル情報研究センター
分子システム情報統合チーム

H-InvDB
Annotated Human Gene Database

# 講習内容

- H-InvDBアノテーション概要紹介
- superTACTシステム紹介
  - システム概要
  - 解析内容
  - 画面での実行操作
- ＴＡＣＴ公開ツール紹介

# superTACT

H-InvDB
Annotated Human Gene Database

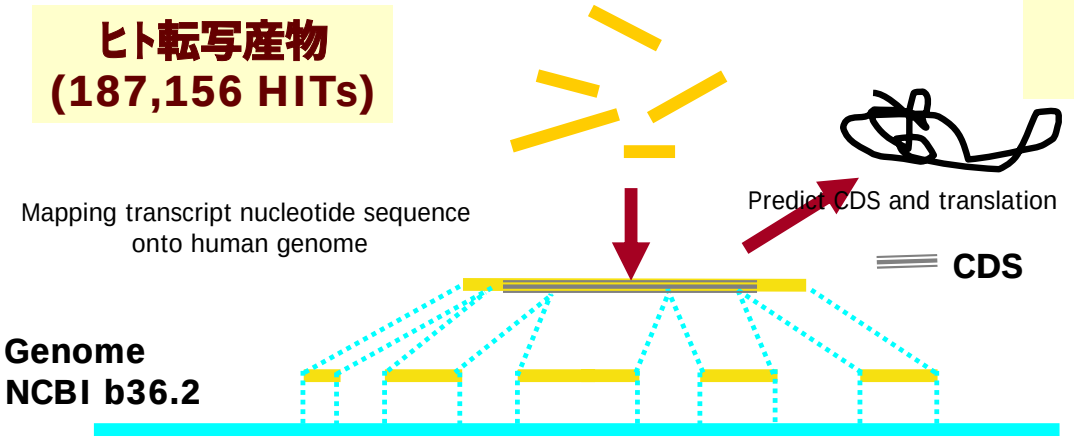
## トランスクリプトームの一括アノテーションのための自動解析システム


- *H-InvDBでの配列解析を自動的に実行するシステム*

# H-InvDBアノテーション概要

H-InvDB
Annotated Human Gene Database

**ヒト転写産物**
**(187,156 HITs)**




**ヒト遺伝子座**
**(36,073 HIXs)**
**435 UM clusters**

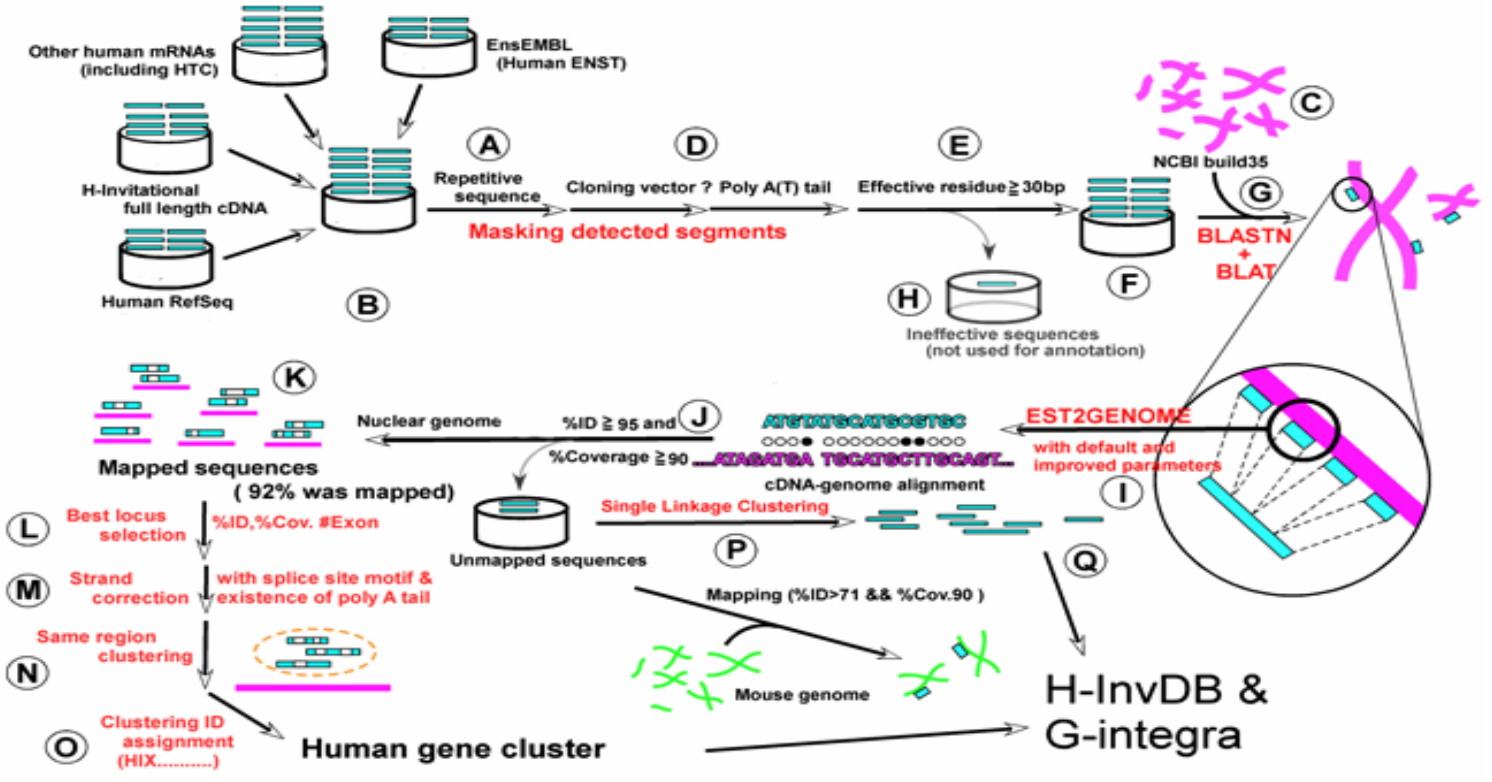
**ヒトタンパク質**
**(124,280 HIPs)**

**相同性検索(ProteinDB)**
**モチーフ予測(InterPro)**

**予測されたタンパク質の**
**遺伝子機能推定**

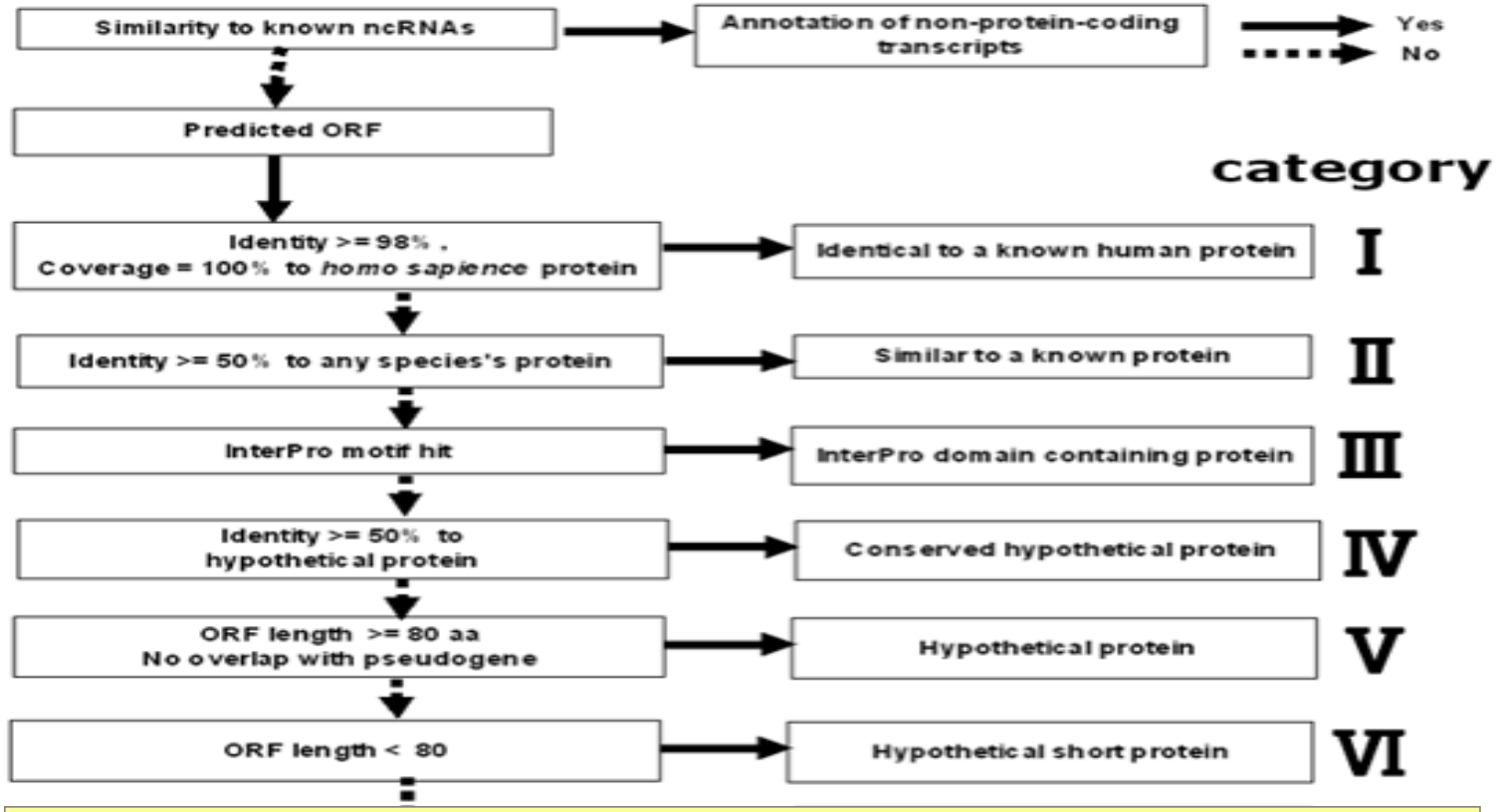
**各種解析を実施し、多数の**
**アノテーション情報を付与**

# マッピングによる遺伝子領域の同定




**転写産物（トランスクリプト）の塩基配列の反復配列をマスクしてからヒトゲノム配列にマップし、遺伝子構造（exon-intron構造）を決定します。**

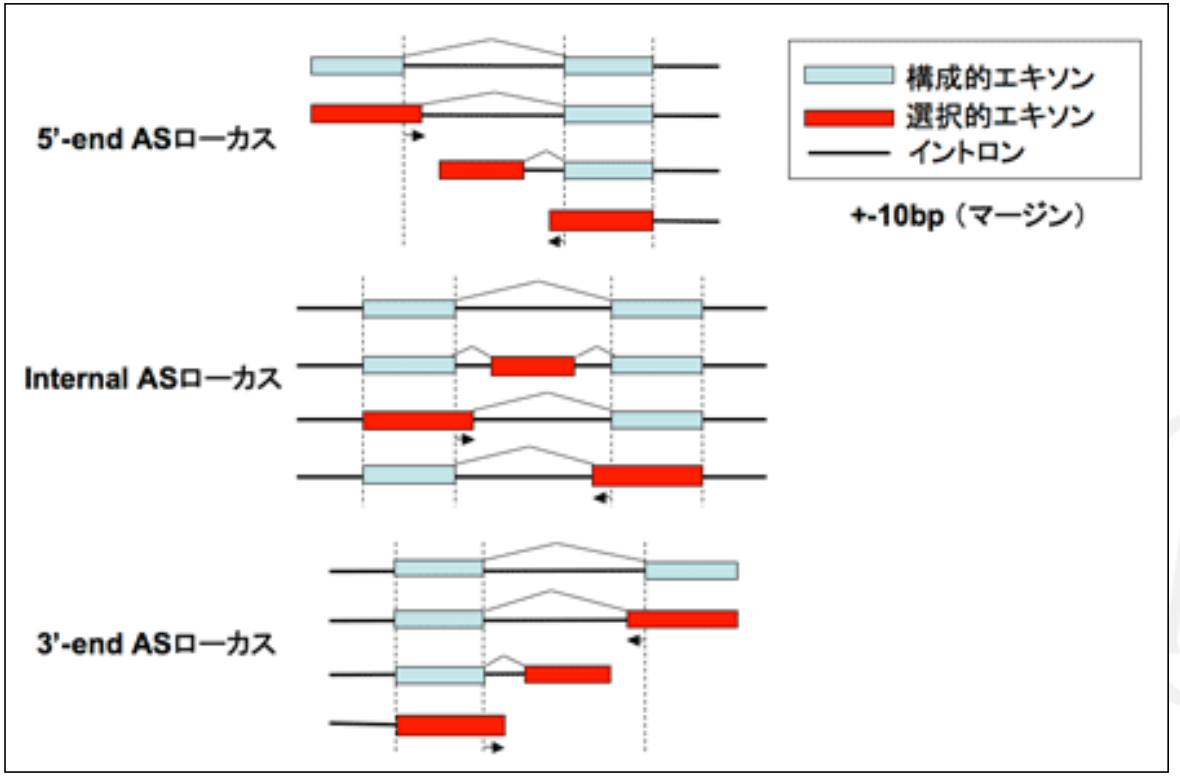
# タンパク質コード遺伝子の機能アノテーション




転写産物のタンパク質をコードする領域を予測し、相同性解析、モチーフ予測により遺伝子の機能を定義するアノテーションを行います。

タンパク質をコードする遺伝子を7つのカテゴリーに分類します。

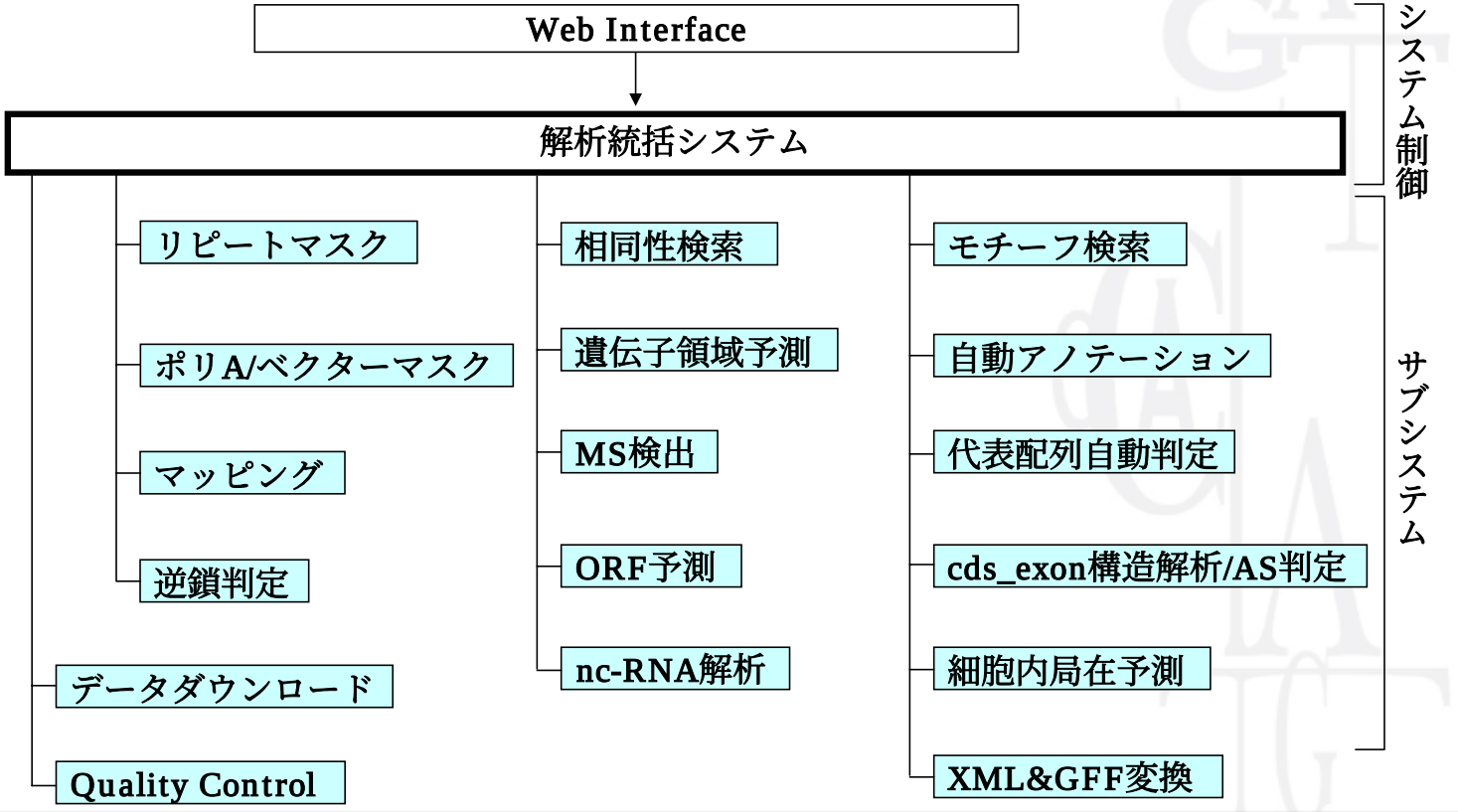
# 選択的スプライシング解析




**エキソンーイントロンのパターンにより、選択的スプライシングの判定を行います。**
**また、各遺伝子座の代表的スプライシングバリアント（RASV）を定義します。**

H-InvDB
Annotated Human Gene Database

# superTACTシステム概要



H-Inv DB
Annotated Human Gene Database

# superTACT:サブシステム定義

**サブシステム ＝ 解析ツールを起動する単位**

| No. | サブシステム | 解析ツール |
|---|---|---|
| 1 | リピートマスク | RepeatMasker |
| 2 | PolyA/vectorマスク | ポリA/ベクターマスクツール (JBIRC original) |
| 3 | 自動マッピング | 自動マッピングツール(JBIRC original:マッピングチーム) |
| 4 | 逆鎖判定 | 逆鎖判定ツール(JBIRC original) |
| 5 | 相同性検索 | Blastx, Fasty |
| 6 | 遺伝子領域予測 | GeneMark |
| 7 | マイクロサテライト検出 | MS検出ツール(JBIRC original) |
| 8 | ORF予測 | ORF予測ツール(JBIRC original) |
| 9 | モチーフ検索 | InterProScan |
| 10 | 自動アノテーション | 自動アノテーションツール(JBIRC original) |
| 11 | 代表配列判定 | 代表配列判定ツール(JBIRC original) |
| 12 | cds_exon/AS判定 | AS判定ツール(JBIRC original:武田さん) |
| 13 | 細胞内局在予測 | TargetP, TMHMM, SOSui |

H-InvDB
Annotated Human Gene Database

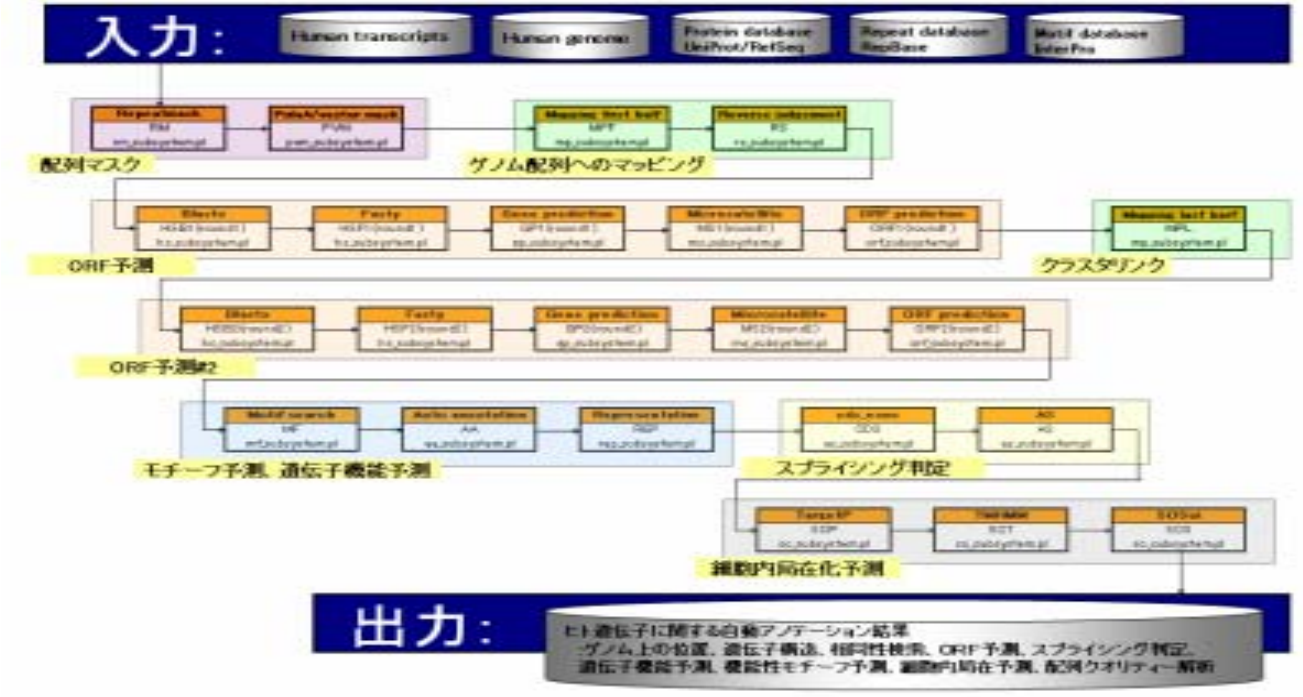
# superTACT:パイプライン定義

**パイプライン　＝　サブシステムから構成される一連の処理を行う単位**

| No. | パイプライン | 処理内容 |
|---|---|---|
| 1 | マスク処理 | 解析時に不必要な領域のマスク処理 |
| 2 | マッピング | ヒトゲノムへのマッピング処理 |
| 3 | ORF予測Round1 | ORF予測処理 |
| 4 | クラスタリング | 遺伝子クラスターの定義 |
| 5 | ORF予測Round2 | ORF予測処理(逆鎖反転差分) |
| 6 | ORF配列を用いた解析 | モチーフ検索、自動アノテーション、代表配列判定<br>AS判定、細胞局在予測など |

H-InvDB
Annotated Human Gene Database

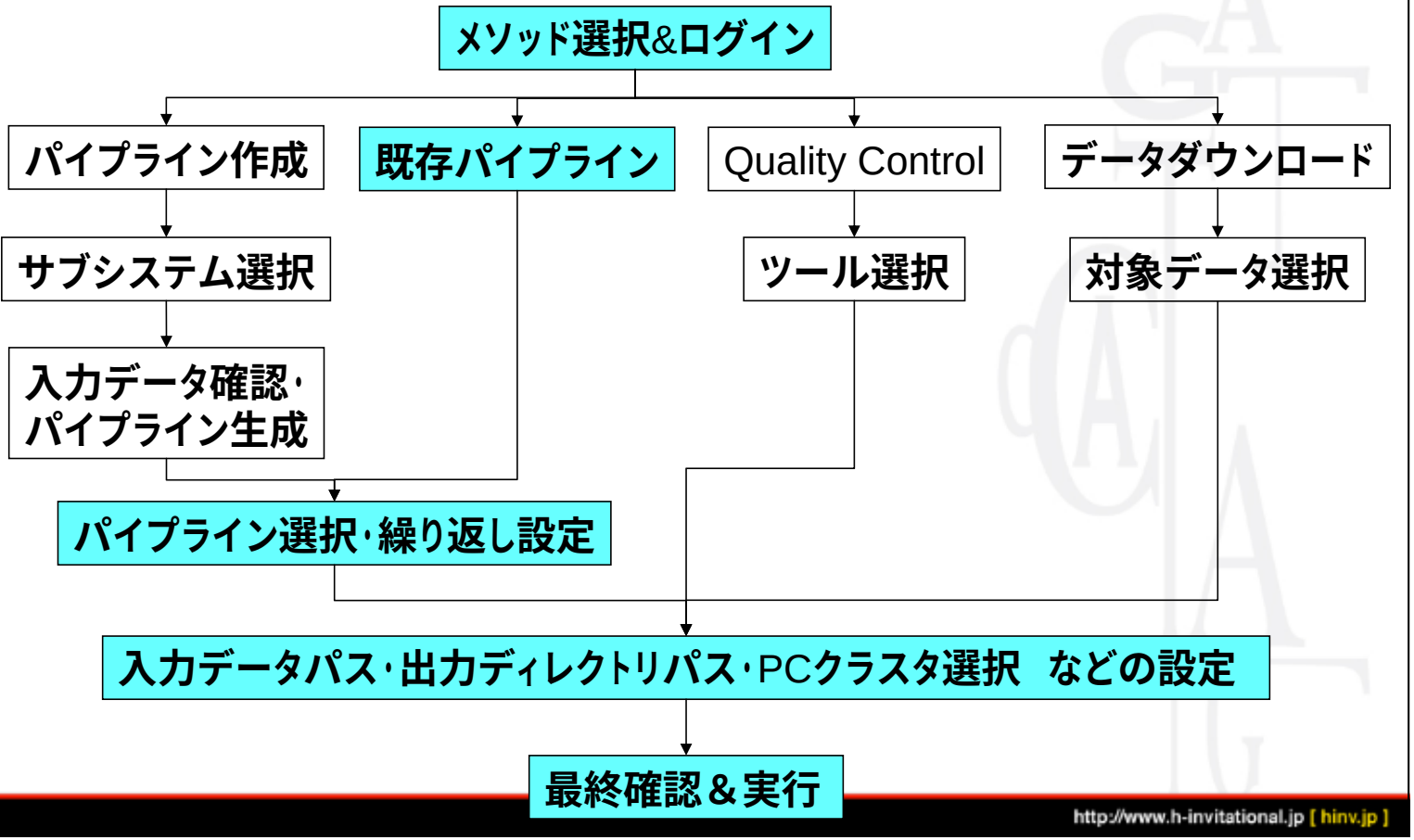
# H-InvDB自動配列解析：superTACT解析フロー




**合計17のサブシステムで構成される6つのパイプラインのうち、任意に選択した解析を実行可能**

H-InvDB
Annotated Human Gene Database

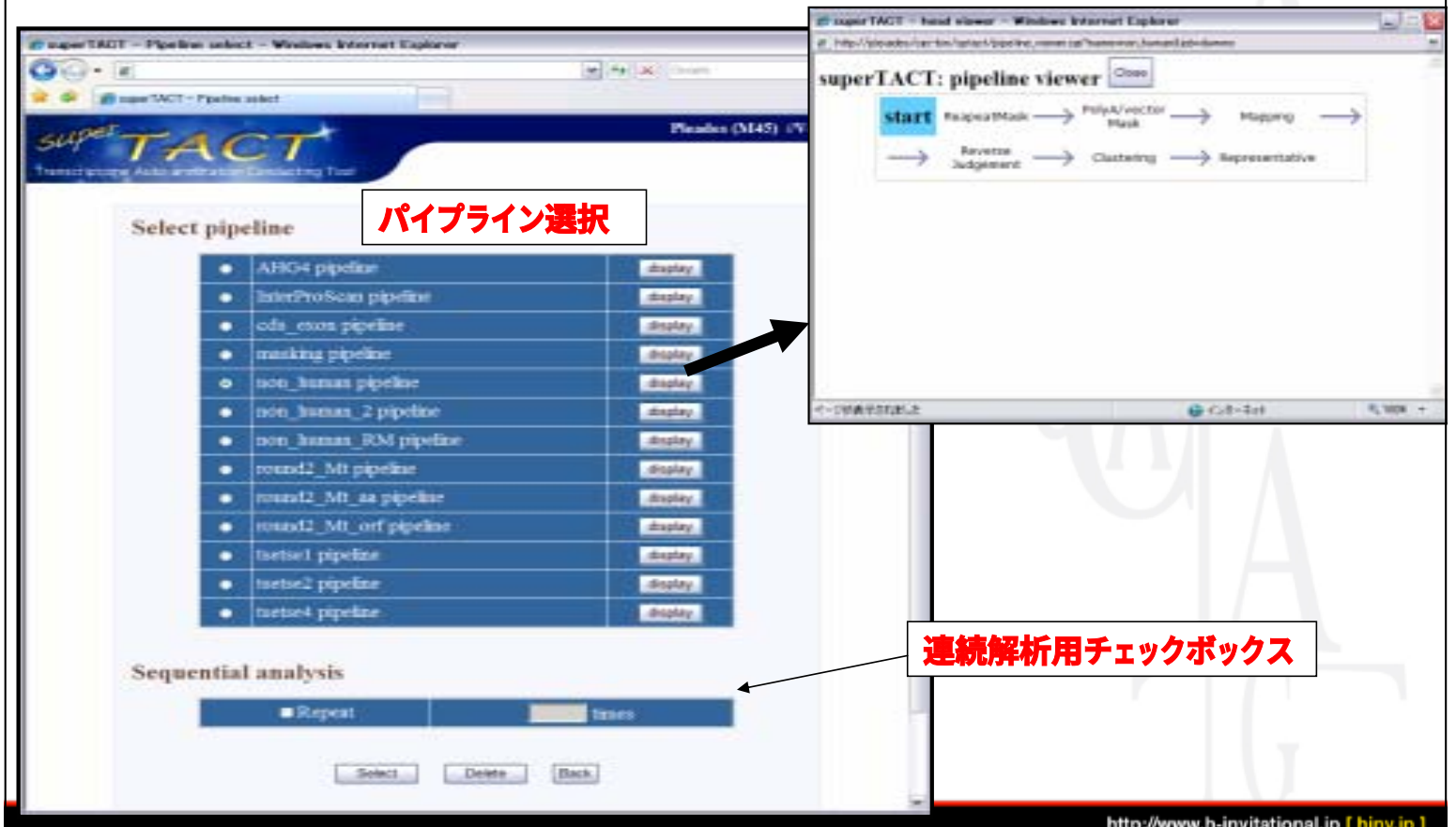
# superTACT操作紹介



H-InvDB
Annotated Human Gene Database

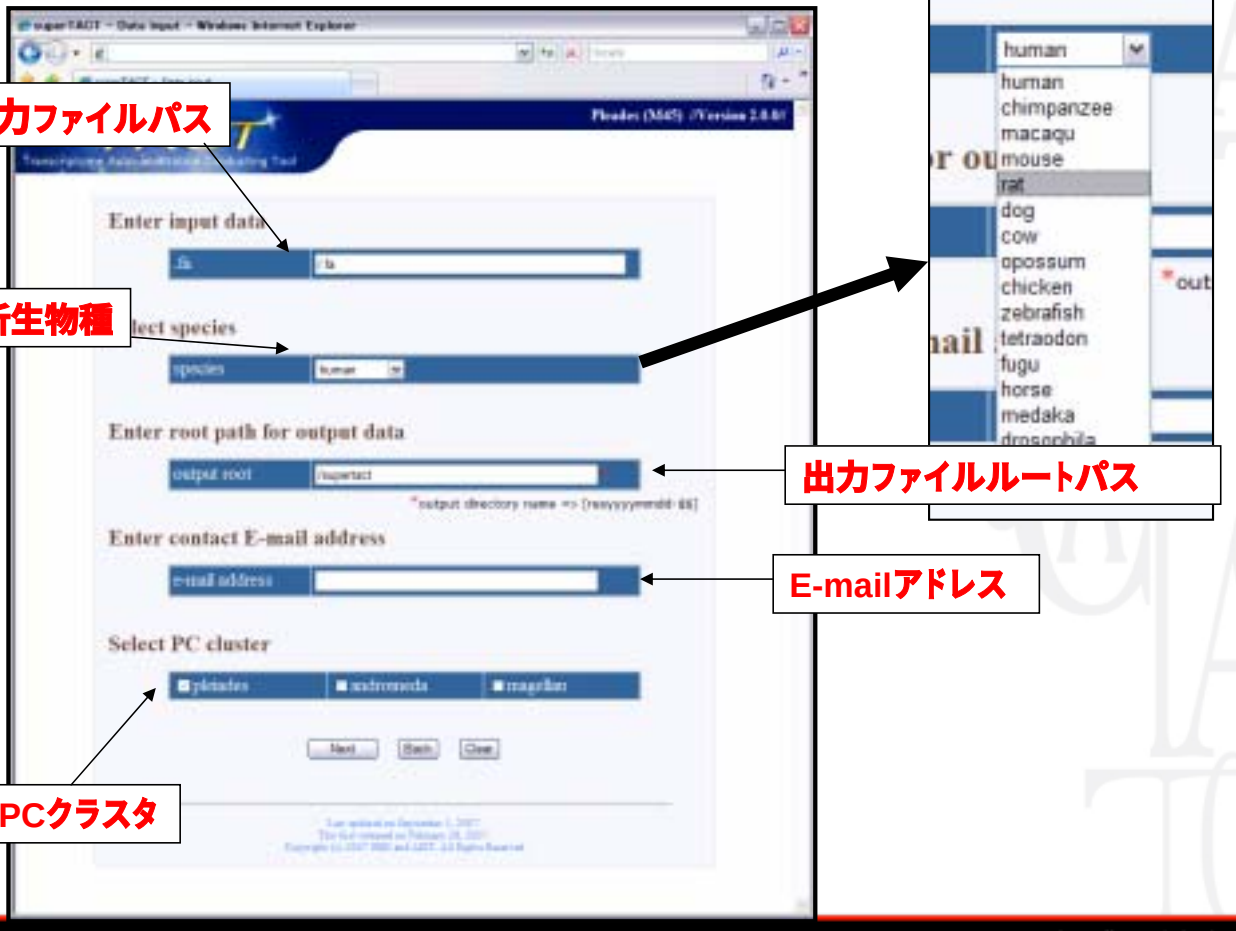
# superTACT画面操作①

## ～パイプライン選択～



H-InvDB
Annotated Human Gene Database

# superTACT画面操作②



H-InvDB
Annotated Human Gene Database

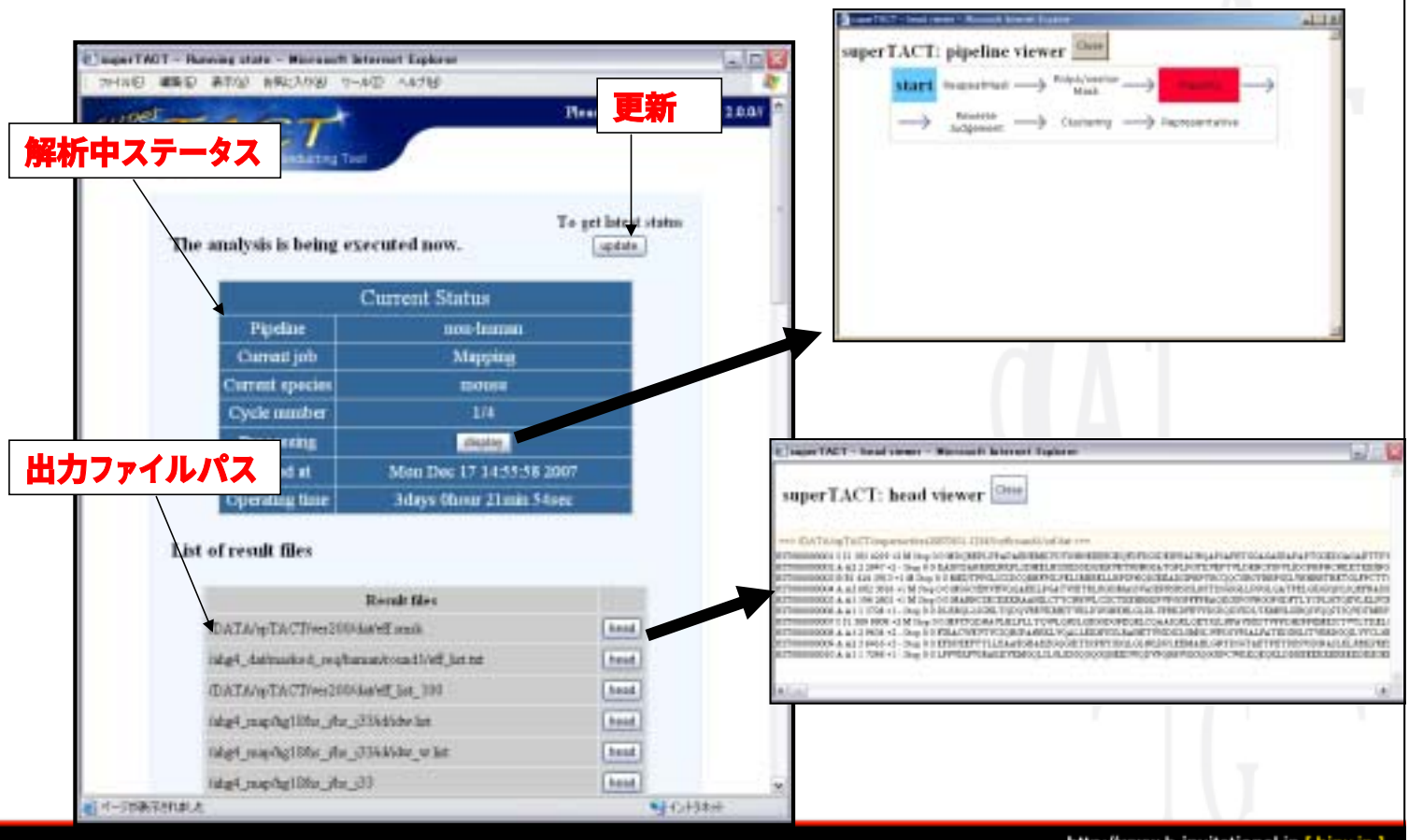
# superTACT画面操作③：解析実行


解析条件確認

解析実行ボタンをクリック

H-InvDB
Annotated Human Gene Database

# superTACT画面：実行中確認



H-InvDB
Annotated Human Gene Database

# 進捗連絡メール

## 開始時

```
== superTACT-pleiades-:repeatmask START ==

[SUBSYSTEM]
  repeatmask

[START TIME]
  Mon May 12 23:18:42 2008
[OUTOUT DIR]
  /ahg4_dat/tsetse_analysis/gm_res/supertact/test_20080512/res20080512-5717

-----
"superTACT System" 2006-2007 JBIC and AIST.
```

## 終了時

```
== superTACT-pleiades- ALL JOB FINISHED ==

[START TIME]
  Mon May 12 23:18:42 2008
[END TIME]
  Mon May 12 23:49:37 2008
[SPECIES]
  tsetse
[output.list PATH]
  /data1/commontool/supertact/analysis/20080512-5717/output.list
[OUTOUT DIR]
  /ahg4_dat/tsetse_analysis/gm_res/supertact/test_20080512/res20080512-5717
[LOG FILE]
  /data1/commontool/supertact/analysis/20080512-5717/sptact.log
[FROM]
  RM
[TO]
  ORF1

-----
```

終了時→出力データとログを確認

H-InvDB
Annotated Human Gene Database

# H19 superTACT自動解析実績

## ヒト369,985　転写産物の配列解析にsuperTACTシステムを使用

モデル生物約130万件のについても一部の解析を実施

| 生物種 | トランスクリプト件数 |
|---|---|
| **human (H-InvDB_5.0)** | **369,985** |
| chicken | 69,514 |
| chimp | 95,829 |
| cow | 71,667 |
| dog | 64,543 |
| fugu | 23,515 |
| horse | 19,438 |
| macaqu | 98,112 |
| medaka | 26,143 |
| mouse | 371,310 |
| opossum | 55,118 |
| rat | 165,061 |
| tetraodon | 128,064 |
| zebrafish | 96,057 |
| 合計 | 1,654,356 |

**ヒト解析時間実績**
**＝ 約786hr (33days)**
**[マッピング除く]**

**イネ(RAP), タイレリア、ツェツェバエなどのアノテーションプロジェクトへ技術提供**

**データベース自動更新システムも連携して開発(superSOUP)**

H-InvDB
Annotated Human Gene Database

# superTACTの長所

- 実行
  - 各解析をサブシステム単位で組み合わせて実行可能
  - オリジナルパイプラインを登録可能
  - 同じパイプラインの別条件（他生物等）の連続実行が可能
  - ＰＣクラスター３台で実行可能
- 進捗確認
  - 解析の進捗をメールで連絡
  - 進捗は画面上でも確認可能
  - 出力データをheadコマンドで確認できる
- 途中停止・再開
  - サブシステム単位での解析の途中停止、（必要な入力ファイルを指定すれば）途中からの再実行が可能
- 出力
  - 出力パスは、ルートを指定するとその下に所定のDirectory/ファイル名で自動出力される

H-InvDB
Annotated Human Gene Database

## superTACTのパイロット版システム

# TACT
# 統合的自動アノテーションツール
**Transcriptome Aute  annotation Conducting Tool of**
**H  InvDB**

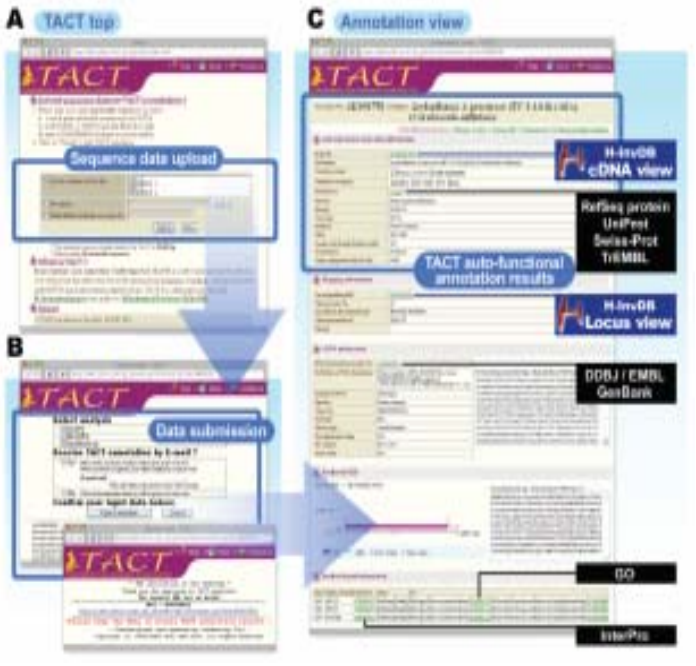

**http://www.h-invitational.jp/tact/index_jp.html**

H-InvDB
Annotated Human Gene Database

superTACTの機能の一部を一般に提供

# 公開TACTシステム v.1.9.0

## Transcriptome Auto-annotation Conducting Tool




2006/07/01論文掲載("TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB"Yamasaki C. et al. NAR 34 Web-server Issue )と同時に一般公開を開始

H-InvDBでの自動解析；相同性検索、ORF予測、モチーフ検索の解析を統合し遺伝子の機能を自動予測できる統合的自動アノテーションシステム

2008年3月3日公開のTACT v.1.9.0

1．42生物種に対応

2．アノテーション結果のダウンロード機能

：FASTA,XMLなど4種の形式で提供

3．定期的に参照データベース更新

H-InvDB
Annotated Human Gene Database

# TACT自動アノテーションフロー

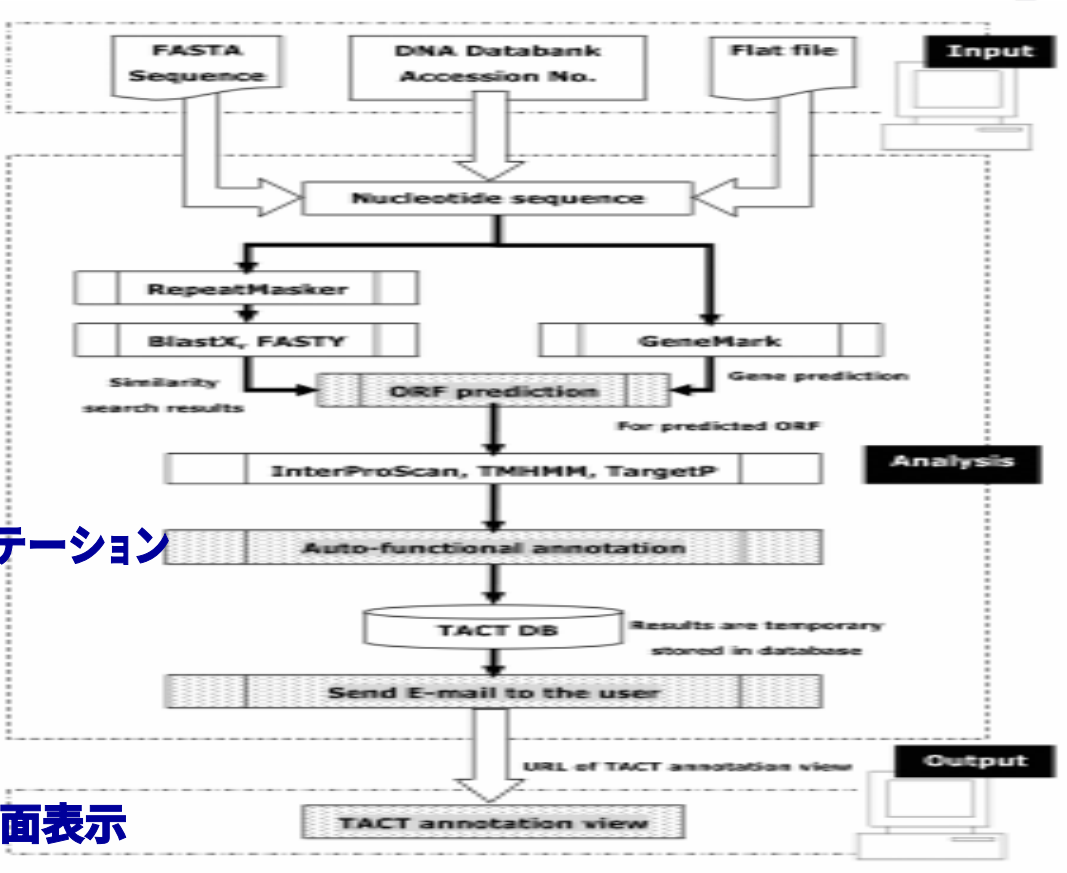
・塩基配列入力

・反復配列マスク
・相同性検索
・ORF予測
・モチーフ予測

・遺伝子機能アノテーション

・結果メール発信

・アノテーション画面表示



H-InvDB
Annotated Human Gene Database

# TACT 入力データ

**TACTには塩基配列クエリ（mRNA, cDNA, EST）を3通りの形式で入力可能.**

- **下記入力方法のいずれかに従ってクエリデータを入力**

  1)FASTA形式データのCopy&Paste
  →マルチFASTAでの入力も可能です
  2)FASTA/マルチFASTAもしくはDDBJフラットファイルのアップロード:
  →アップロードできるファイルは最大1MBまでです
  3)DDBJ/EMBL/GenBank accession numberの入力
  →カンマ・セミコロン・スペース区切りで複数入力が可能です

  [NOTES]
  * TACTに投入できるクエリ配列の最大長は30,000 bpです.
  * TACTに投入できるマルチFASTAの件数は最大で10です。

- **ID入力によるクエリ配列のリモート取得**

  塩基配列の DDBJ/EMBL/GenBank accession numberを入力すると、DDBJからデータを取得して解析を実行することができます。(DDBJ; http://www.ddbj.nig.ac.jp/).

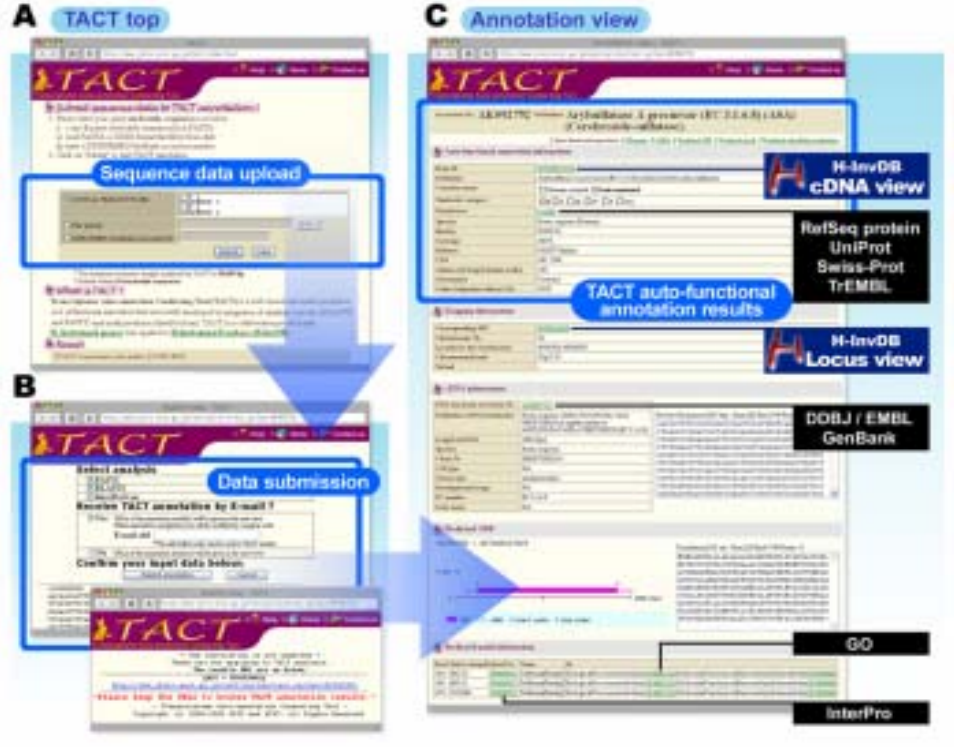
- **あらゆる生物種の解析が可能**

  TACTの解析パイプラインはヒト完全長cDNAのアノテーション の為に開発されたものですが、ヒト以外の全生物種の塩基配列クエリを用いて実行することが可能です。

H-InvDB
Annotated Human Gene Database

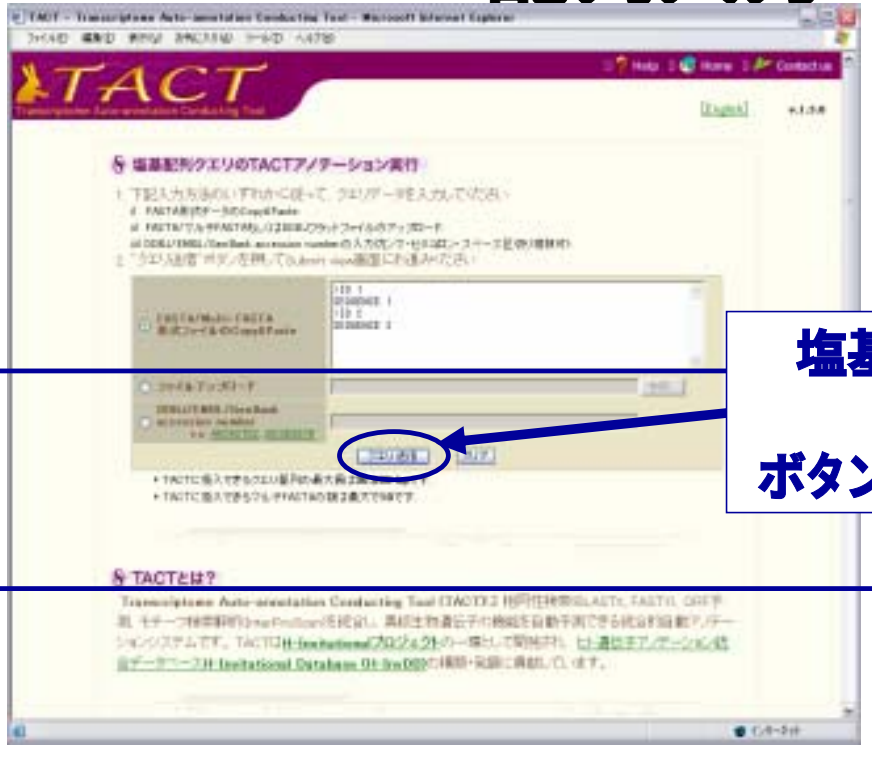
# TACT 画面構成

TACTはwebベースの解析システムで、3つのメイン画面（トップページ、Data submission、Annotation view）から構成されています。



H-InvDB
Annotated Human Gene Database

# TACT top page
# 配列入力




塩基配列クエリを入力し
「クエリ送信」
ボタンをクリックして下さい。

入力塩基配列の種類は、ラジオボタンで選択できます。

H-InvDB
Annotated Human Gene Database

# TACT data submission
## 自動アノテーション開始


解析のオプションを選択

「自動アノテーション開始」

配列クエリを確認

アノテーション結果をe-mailで受け取る事ができます。

H-InvDB
Annotated Human Gene Database

# TACTクエリ受け付け完了画面
## :解析結果URL


自動アノテーション結果詳細は
画面のURLから参照できます。

アノテーション実行中です。

H-InvDB
Annotated Human Gene Database

# TACT：annotation view


**自動アノテーション結果と関連するデータベースへのリンクが5つのセクションで提供されています。**

**>>Auto-annotation information**

**>>Mapping information**

**>>cDNA information**

**>>Predicted ORF**

**>>Predicted motif**

**自動アノテーション結果を閲覧することができます。**

H-InvDB
Annotated Human Gene Database

# TACT公開・更新

**TACTは2006/07/01論文掲載と同時に一般公開を開始し、
その後定期的に更新を行っています。**

- [2006/07/01]
  ・TACTに関する論文をNAR web-server issueに発表しました。
  "TACT: Transcriptome Auto-annotation Conducting Tool of H-InvDB"
  Yamasaki C. et al. NAR 34 Web-server Issue
- [2006/07/01] TACT_1.0.0
  ・H-InvDBの自動アノテーションツールとして無償で一般公開を開始しました。
- [2006/12/01] TACT_1.3.0
  ・日本語ページの提供を開始しました。
- [2007/03/01] TACT_1.5.0
  ・解析結果ダウンロード機能の提供を開始しました。
  形式：XMLファイル、テキストファイル、塩基・アミノ酸配列FASTAファイル
  ・解析生物種を拡張しました。（合計42種）。
- [2008/03/03] TACT_1.9.0
  ・解析参照タンパク質データベースを更新しました。
  - UniProt Knowledgebase Release 12.8
  - RefSeq release 27 (human proteins)

H-Inv DB
Annotated Human Gene Database

# ありがとうございました。

H-Inv DB
Annotated Human Gene Database
www.h invitational.jp

TACT
Transcriptome Auto-annotation Conducting Tool
www.h invitational.jp/tact/

JBiC 社団法人 バイオ産業情報化コンソーシアム
Japan Biological Informatics Consortium

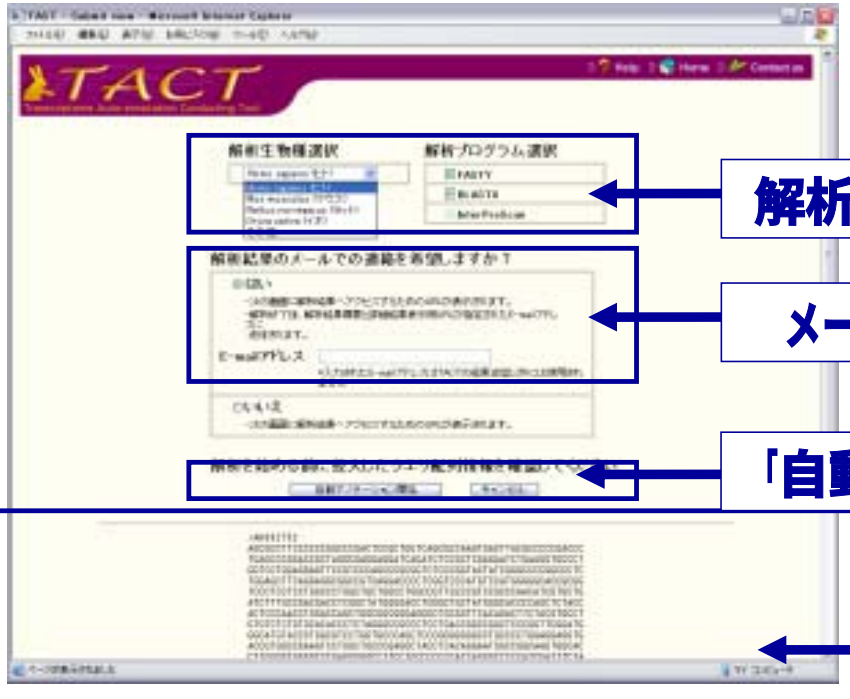
BIRC
Bioinformatics Information Research Center

DDBJ

National Institute of
Advanced Industrial Science and Technology (AIST)
独立行政法人 産業技術総合研究所

H-InvDB
Annotated Human Gene Database

# TACT data submission
## 自動アノテーション開始




解析のオプションを選択

メールアドレスを入力

「自動アノテーション開始」

配列クエリを確認

アノテーション結果をe-mailで受け取る事ができます。