グローバルCOEプログラム系統講義「ベーシックサイエンスコース」

実践バイオインフォマティクス：生物系研究者のための配列解析

Practical bioinformatics: sequence analyses for life scientists

嶋田　誠

（藤田保健衛生大学 総合医科学研究所 遺伝子発現機構学）

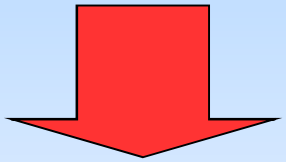
2012年 10月 9日（火） 17:00ī 18:30　名古屋大学基礎医学研究棟1階会議室2

アウトライン

- 概論
  - ねらい
  - バイオインフォマティクス：どこまで把握すべきか？
- データについて
  - データ加工変換
  - パース
- 実験研究者にとってのプログラミング
  - アルゴリズム
- ゲノム関連大規模プロジェクト
- データベース
- ツール群の分類と見つけ方

ねらい

- 実験系の研究者が、
- 今後、それぞれの研究活動において、
- 適宜、必要なデータやツールを探し出して、使えるようになること。



基本的に独学：

- 調べ方
- 他人への尋ね方
  - -->基本的用語の解説
- 最初の段階で重要な事は何か
  - データ構造の約束事＆データ入出力の仕方
  - -->将来*in silico*で解析しやすい実験データのまとめ方（キー、文字）

バイオインフォマティクス：
どこまで把握しておくべきか？

実験室での作業だけでも時間管理が大変なのに。

- バイオインフォマティクス用語
  - やたらと略語が多い？
- データベースやツール群
  - どこになにがあるのか。
- 計算手法・アルゴリズム
  - ソフト任せのブラックボックス状態でいいのか？

「どこまで」を考えるうえでのヒント1

- 開発者とユーザーとがいる
  - 開発者:
    - プログラムを組んでソフトやDBを作りだす。
    - アルゴリズムを考える。
    - バイオインフォマティクスそのものの未来を考える。
  - ユーザー:
    - 自分のデータを解析する。
    - 解析のアイデアを練る。-->開発者へ。
- ユーザーは開発者にならなくてもいい。
  - 開発者（仲介者）とコミュニケーションをとる能力は必要。
    - ヘルプ、マニュアル文書、メール、打合せ等。
    - ユーザーとして必要な用語や概念だけでも理解しておきたい。

「どこまで」を考えるうえでのヒント2

- 研究手法・解析手法における栄枯盛衰
  - DBでは激しい。
  - 手法やアルゴリズムにおける原理は様々な分野で繰り返し利用されている。
    - 例）モンテカルロ、パーシモニー、ニューラルネットワーク
- DBでは必要な時に必要なものを探し出す技術を。
- 手法やアルゴリズムは原理を（一度は）理解し、アレルギーを起こさないようにしておく。

「どこまで」を考えるうえでのヒント3

- 研究は常に新発見を求め、新しきことを試す行為なり。

ならば、

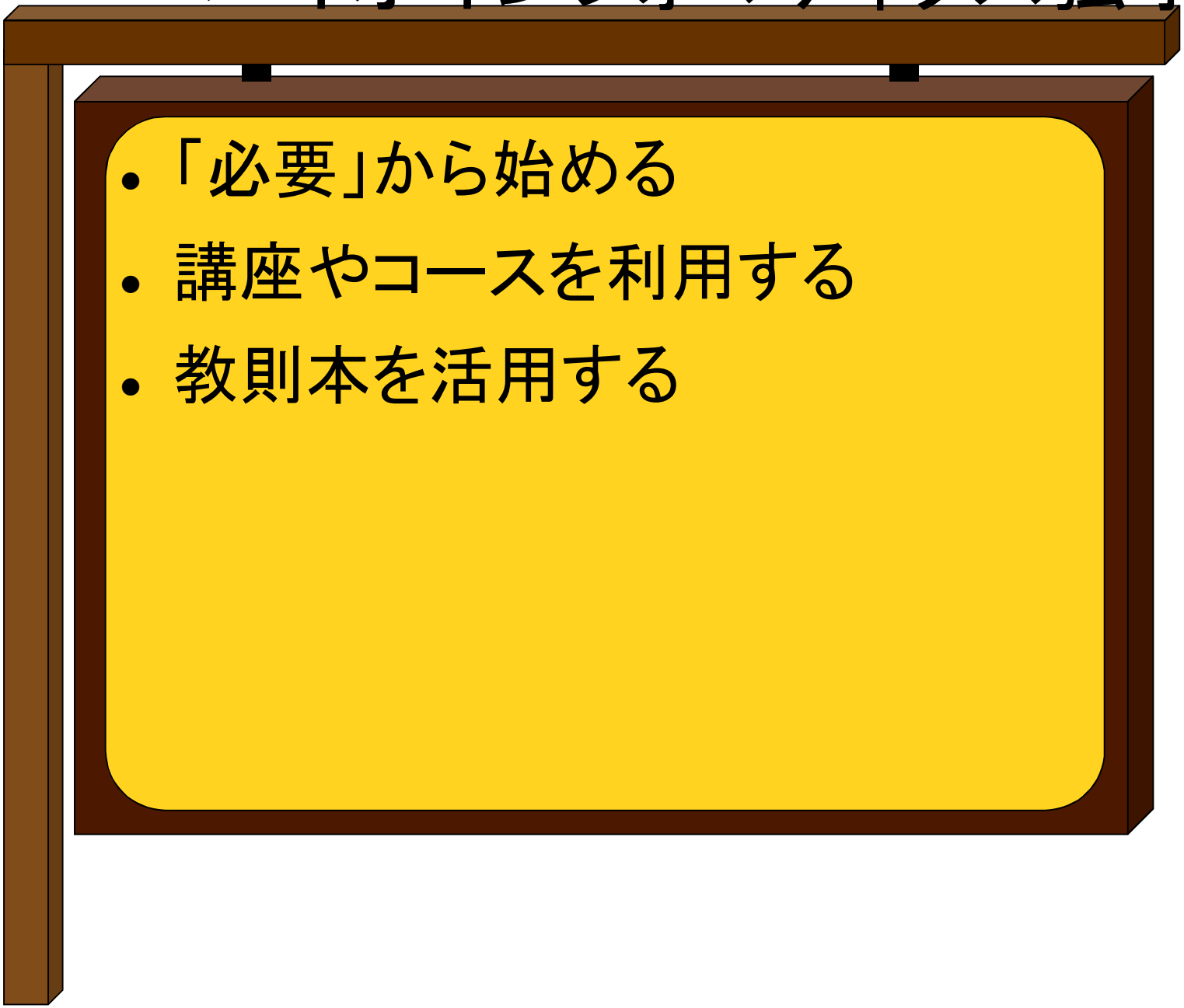
- 既存のツールでは限界がある。
  - 別の処理と組み合わせたり、修正したり、が必要。
  - 開発といっても、全く無から新規を作り出す必要はない。
    - 「車輪の再発明はするな。」

- 「数こそ力」の課題もある。

ならば、

- 繰り返し作業の得意なコンピュータを利用しよう。

バイオインフォマティクス独学のコツ



- 「必要」から始める
- 講座やコースを利用する
- 教則本を活用する

「必要」から始める

- 実験データのまとめ
  - 表計算ソフト＋目視・手作業
  - 追いつかない->改善の「必要」
- ツール、DB、開発環境、OS等は「とっつきやすさ」
  - 文字：ヘルプ、マニュアル、教則本
  - 人間関係：コミュニティー、周りの人

•講座やコースを利用

- 講座資料の利用
  - ライフサイエンス統合データベースプロジェクトの人材育成活動（お茶の水女子大学担当部分）
  - 統合データベース支援：DB構築者の養成におけるバイオDBサーバー構築演習2007年度の演習ノート
  - 理論分子生物学（京都大学理学部）講義資料
  - 本日分→ http://tinyurl.com/shimada-mk

- ビデオやストーミング配信を聴講
  - JST BIRD人材養成
  - 統合データベース:統合TV

データの形式と変換

- テキストデータとバイナリデータ
- 改行コード・文字コード
- データ形式：区切り型とタグ挿入型
- ゲノム情報学で良く使うデータ・フォーマット
- パース(parse)、パーサ(parser)
- 正規表現
- テキスト・エディターからプログラミングまで
- ID変換
- ゲノム・バージョン変換

テキストデータとバイナリデータ

- [問い]違いは？
- [回答例]
  - コンピュータ上のデータは全て２進数で表現できる。
  - そういった意味では全てバイナリデータ。
  - その中でもテキストとして読めるものがテキストデータ。
  - テキストとして読めるようにするのに、いろいろからくりがある。

テキストファイル：
意外と知らない改行コード

- 「CR」(Carriage Return : 行頭復帰)
- 「LF」(Line Feed : 改行)

タイプライタ-->テレックス：重ね打ちができるように2つを分けたのが起源

- LF:　UNIX系のシステム。Linux、Mac OS Xなど。
- CR+LF: OS/2、Microsoft Windowsなど
- CR:　Mac OS（バージョン9まで）

テキストファイル：意外と知らない文字コード（日本語）

- Shift-JIS:WindowsやMacOS（9まで）の内部コードとして使用されてきた
  - 細かくはWindowsもXp以前とVista以降では細かい部分で異なる。
- EUC-JP：Unix系の内部コードとして使用されてきた。
- UTF-7 / UTF-8：最近使われだした国際統一を目指した規格でそれぞれ7ビットと8ビット伝送路。

テキストファイル：
意外と知らない文字コード（まとめ）

- 時代の流れ
  - 初期：American National Standards Institute(ANSI)　英語のアルファベットと記号を表現するもの。
  - 一昔前：各国語でそれぞれに対応していた時代。
  - 最近：統一規格で各国語に対応できるコードを。
- 実際使用に際して
  - 汎用機<--->パソコン、OS間でファイルをやり取りする
    - 自動で変換されない場合は、要変換
      - 開くソフトでエンコードを変更。
      - または専用変換ソフトを使う。
    - ファイル名は普段から2バイト文字を使わないように注意。

バイトオーダーマーク BOM

UTF-16 では2バイト文字の1バイト目を先に書く方法とその逆の方法の2方法（エンディアン)が並立。

両者はそのままでは区別が付かないので、先頭にどちらかを区別するために記入される16ビットの値 のことをBOMという。

UTF-8などエンディアンがなくとも、Unicodeで記述されている ことを示す目的でつけられることもある。

通常Bioinfomatics dataは1バイト文字であるが、テキストファイルとして保存するときについた、BOMがファイル読み込みエラーを引き起こすこともあるので注意。

Data type　データの型(主なもの)

Statistics	Computer programming
real-valued	real data type 実数型 floating-point 浮動小数点数 fixed-point 固定小数点数
count data	integer type 整数型
binary data	Boolean
categorical data	enumerated type 枚挙型
random vector	list or array リスト型、配列型
random matrix	two-dimensional array 行列型
-	character type 文字型
-	string type 文字列型

DB検索に知っとくと得するキーワード

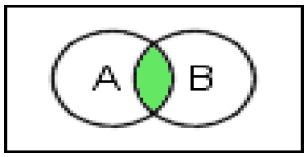
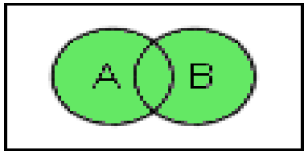
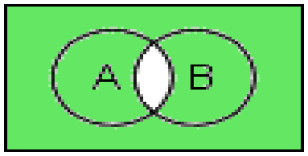
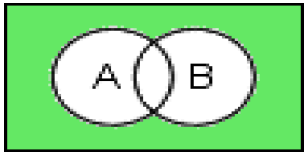
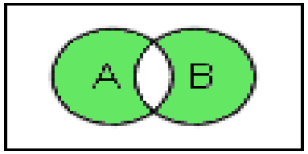
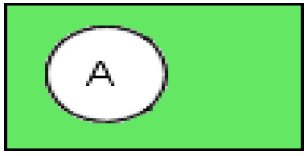
- 演算子(operator)、引数（オペランド operand）、

例）$a + 2$　の例では演算子と引数はそれぞれどれか

答え）演算子= ñ+ò、引数= ñaò, ñ2ò

- ブール演算(Boolean Operation)
  - 論理演算(Logical Operation)ともいわれる。
  - 1（真）か0（偽）かの2通りの元(要素)だけをもつ集合における演算のこと。
  - プログラミングだけでなく、データベースを複合検索(eg., AND検索、NOT検索)する際にも用いられる。
    - データベースごとに約束事が違うので注意。

ブール演算(Boolean Operation)

	式	真理値表（Truth Table）			ベン図（Venn Diagram）
		入力A	入力B	出力	
論理積（AND）	A and B A・B A×B A∩B A∧B	0	0	0	
		0	1	0	
		1	0	0	
		1	1	1	
論理和（OR）	A or B A＋B A∪B A∨B	0	0	0	
		0	1	1	
		1	0	1	
		1	1	1	
否定論理積（NAND）	A nand B $\overline{A・B}$ ¬（A∧B）	0	0	1	
		0	1	1	
		1	0	1	
		1	1	0	
否定論理和（NOR）	A nor B $\overline{A＋B}$ ¬（A∨B）	0	0	1	
		0	1	0	
		1	0	0	
		1	1	0	
排他的論理和（EOR、XOR）	A eor B A xor B $\overline{A}・B＋A・\overline{B}$ （A∨B）∧¬（A∧B）	0	0	0	
		0	1	1	
		1	0	1	
		1	1	0	
否定（NOT）	A not B $\overline{A}$ ¬A	0		1	
		1		0	

http://www.asahi-net.or.jp/~ax2s-kmtn/ref/logicope.htmlより改変

データ形式：タグ挿入型と区切り型

XML形式	HTML形式	CSV形式	フラットファイル形式
<成績表> <名前>鈴沢 一郎</名前> <国語>67</国語> <算数>70</算数> <理科>95</理科> <社会>87</社会> </成績表>	<TABLE> <TD>鈴沢 一郎</TD> <TD>67</TD> <TD>70</TD> <TD>95</TD> <TD>87</TD> </TABLE>	鈴沢 一郎, 67, 70, 95, 87	名前　鈴沢 一郎 国語　67 算数　70 理科　95 社会　87

- Markup Language(ML)では項目をタグで囲ってあらわす。
- (ML)階層的なデータ構造定義が可能
- XML（eXtensible Markup Language）はタグを自由に設定できる。
- (ML)Document Type Definition(DTD)で要素の型と、要素の親子関係を定義
- XMLはXSLスタイルシートにより、データの内容と表現を分離して管理可能
- 区切り型はComma Separated Values (CSV), Tab SV(TSV)

ゲノム情報学で良く使われる データ・フォーマット

- 配列用フォーマット
  - 単独
  - アライメント
  - 例）FASTA, PHYLIP, ALN, SAM*, BAM* (* to reference seq.)
- 多型情報表示用フォーマット
  - 例）vcf
- ゲノム領域用フォーマット
  - 例）GFF

- データ入出力はツールに依存する。まずは従うことが肝要。
- 互いに変換可能。
  - 文字列扱いに便利なプログラム言語が多様されるようになった。

代表的な配列ファイルフォーマット
FASTA (format, file)

```
>seq_1
CTCCATAATCAT
>seq_2
CTCCATAATTCAT
```

```
>seq_1
CTCCATAAT-CAT
>seq_2
CTCCATAATTCAT
```

配列ファイル関連で役立つキーワード

Phred quality score (quality value, QUAL):

$quality = -10\log_{10}p$

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

http://en.wikipedia.org/wiki/Phred_quality_score

QUAL = *quality* + 33

ASCIIコード表(10進法)を参照Ÿ 対応するASCII文字をファイルに書き込む

なぜ33を足すのか？

QUAL = *quality* + 33

ASCIIコード表(10進法)を参照Ÿ 対応するASCII文字をファイルに書き込む

文字	コード		文字	コード		文字	コード		文字	コード		文字	コード		文字	コード		文字	コード		文字	コード	
	10進	16進		10進	16進		10進	16進		10進	16進		10進	16進		10進	16進		10進	16進		10進	16進
NUL	0	00	DLE	16	10	SP	32	20	0	48	30	@	64	40	P	80	50	`	96	60	p	112	70
SOH	1	01	DC1	17	11	!	33	21	1	49	31	A	65	41	Q	81	51	a	97	61	q	113	71
STX	2	02	DC2	18	12	"	34	22	2	50	32	B	66	42	R	82	52	b	98	62	r	114	72
ETX	3	03	DC3	19	13	#	35	23	3	51	33	C	67	43	S	83	53	c	99	63	s	115	73
EOT	4	04	DC4	20	14	$	36	24	4	52	34	D	68	44	T	84	54	d	100	64	t	116	74
ENQ	5	05	NAK	21	15	%	37	25	5	53	35	E	69	45	U	85	55	e	101	65	u	117	75
ACK	6	06	SYN	22	16	&	38	26	6	54	36	F	70	46	V	86	56	f	102	66	v	118	76
BEL	7	07	ETB	23	17	'	39	27	7	55	37	G	71	47	W	87	57	g	103	67	w	119	77
BS	8	08	CAN	24	18	(	40	28	8	56	38	H	72	48	X	88	58	h	104	68	x	120	78
HT	9	09	EM	25	19	)	41	29	9	57	39	I	73	49	Y	89	59	i	105	69	y	121	79
NL*	10	0a	SUB	26	1a	*	42	2a	:	58	3a	J	74	4a	Z	90	5a	j	106	6a	z	122	7a
VT	11	0b	ESC	27	1b	+	43	2b	;	59	3b	K	75	4b	[	91	5b	k	107	6b	{	123	7b
NP	12	0c	FS	28	1c	,	44	2c	<	60	3c	L	76	4c	\	92	5c	l	108	6c	|	124	7c
CR	13	0d	GS	29	1d	-	45	2d	=	61	3d	M	77	4d	]	93	5d	m	109	6d	}	125	7d
SO	14	0e	RS	30	1e	.	46	2e	>	62	3e	N	78	4e	^	94	5e	n	110	6e	~	126	7e
SI	15	0f	US	31	1f	/	47	2f	?	63	3f	O	79	4f	_	95	5f	o	111	6f	DEL	127	7f

0Ò*quality* Ò50 としたときにQUALが取り得る範囲

一文字(1bit)で書き込める。->ファイル容量の節約

配列ファイル関連で役立つキーワード

CIGAR: referenceとのMatch, Insertion, Deletionをread配列の順にいくつつながっている状況であるか表す書式

例

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
=>
POS: 5
```

CIGAR: 3M1I3M1D5M

ゲノム領域データのフォーマット

- 単一配列やalignment配列のフォーマットは配列の文字列やそれらのalignment関係を表現。
- ゲノム領域はDNAストランド、ゲノム座標等を表現。
- MAFは両者を合わせたフォーマット。

BED format
bigBed format
PSL format
GFF format
GTF format
MAF format
BAM format
WIG format
bigWig format
Microarray format
Chain format
Net format
Axt format
.2bit format
.nib format
GenePred table format

GFF (v3) format

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene    1000  9000  .  +  .  ID=gene1;Name=EDEN

ctg123 . TF_bind 1000  1012  .  +  .  ID=tfbs1;Parent=gene1

ctg123 . mRNA    1050  9000  .  +  .  ID=mRNA1;Parent=gene1;Name=EDEN.1
ctg123 . mRNA    1050  9000  .  +  .  ID=mRNA2;Parent=gene1;Name=EDEN.2
ctg123 . mRNA    1300  9000  .  +  .  ID=mRNA3;Parent=gene1;Name=EDEN.3
ctg123 . exon    1300  1500  .  +  .  ID=exon1;Parent=mRNA3
ctg123 . exon    1050  1500  .  +  .  ID=exon2;Parent=mRNA1,mRNA2
ctg123 . exon    3000  3902  .  +  .  ID=exon3;Parent=mRNA1,mRNA3
ctg123 . exon    5000  5500  .  +  .  ID=exon4;Parent=mRNA1,mRNA2,mRNA3
ctg123 . exon    7000  9000  .  +  .  ID=exon5;Parent=mRNA1,mRNA2,mRNA3
ctg123 . CDS     1201  1500  .  +  0  ID=cds1;Parent=mRNA1;Name=edenpr.1
ctg123 . CDS     3000  3902  .  +  0  ID=cds1;Parent=mRNA1;Name=edenpr.1
```

1:seqid

2:source

3:type 4:start 5:end 6:score

7:strand

8: frame

9:attributes

parse(パース、構文解析)

- 情報学では構文解析は字句解析とプログラムの文法の正しさの判断との2つの部分を指す。
  - うらにわにはにわにわにはにわにわとりがいる
  - 裏庭には庭庭には庭鶏がいる。
  - 裏庭に葉二把、庭には丹羽鶏がいる。
  - 裏にワニ埴輪、庭には丹羽鶏がいる。
  - 裏庭に埴輪、庭に埴輪、鶏がいる。
  - 裏庭には二羽、庭には二羽、鶏がいる。

parse(パース、構文解析)

- 情報学で構文解析とは「字句解析」と「プログラムの文法の正しさの判断」との2つの部分を指す。
- XMLファイルでは、parserを介して、テキスト部分を抜き出すことにより、人の目で解釈できる。
- バイオインフォマティクスでは:
  - BLAST結果などヒトが眼でみるために書かれたファイル
  - 字句解析-->次の処理(プログラムでは)
  - いくつかのライブラリとして公開されている
    - BioPerlでBLASTをparseするライブラリ
    - http://www.bioperl.org/wiki/Parsing_BLAST_HSPs

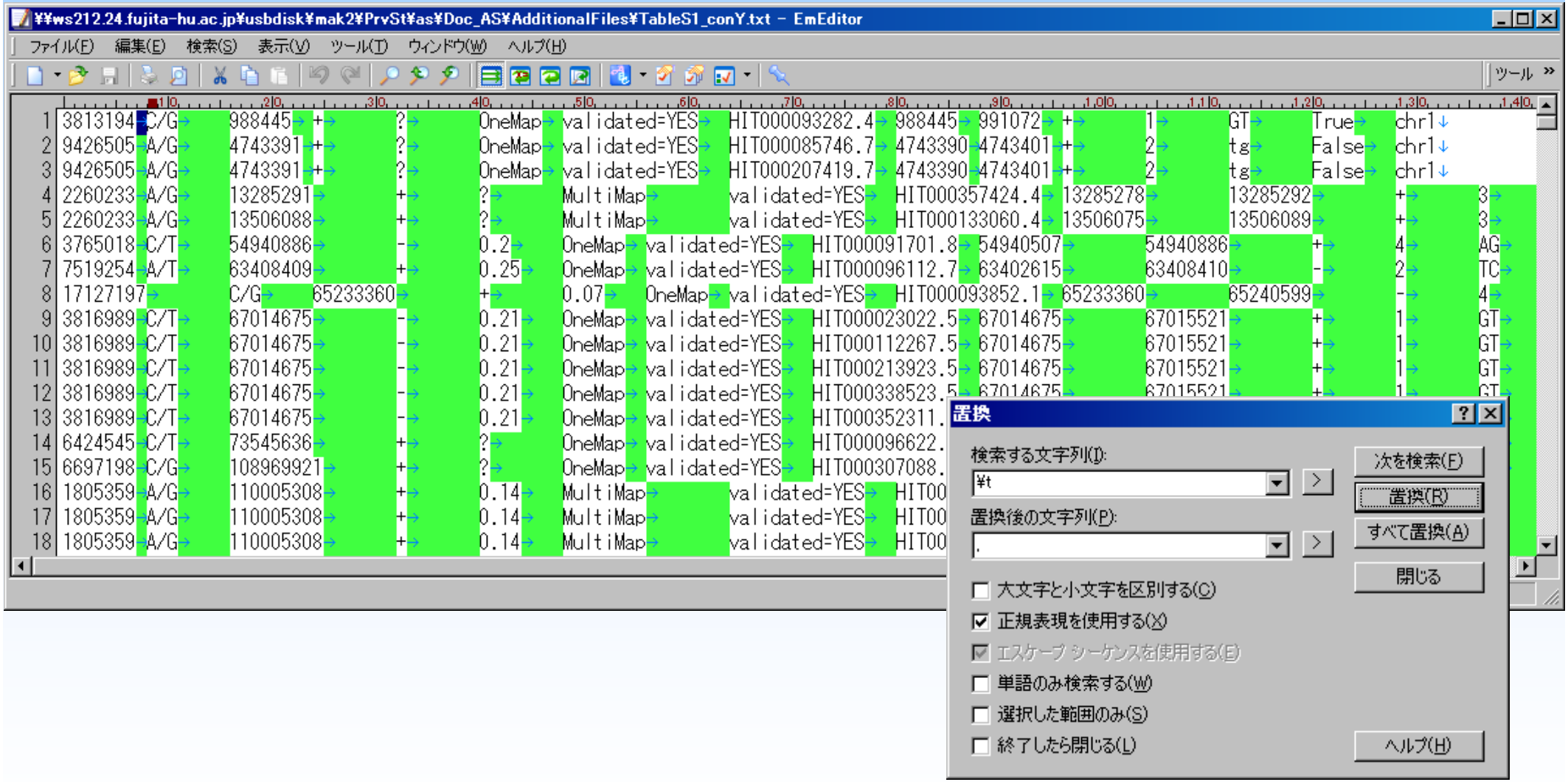
正規表現 (regular expression)

- 文字列の集合を一つの文字列で表現する方法。
- 正規表現を使うと、文字列の検索や置換をパターンで行う事ができる。
- 配列のデータは文字列のデータであるので、検索だけでなく、プログラミングにも正規表現は重要。
- 例）
  - ?：直前の文字が0か1個ある。"*colou?*r" は、
  - *：直前の文字が0個以上ある。"*go*gle*" は、
  - |：またはの意。
  - ¥dまたは[0-9]：数字にマッチ。
  - ¥n：改行コードにマッチ。

要チェックツール：テキストファイル作成

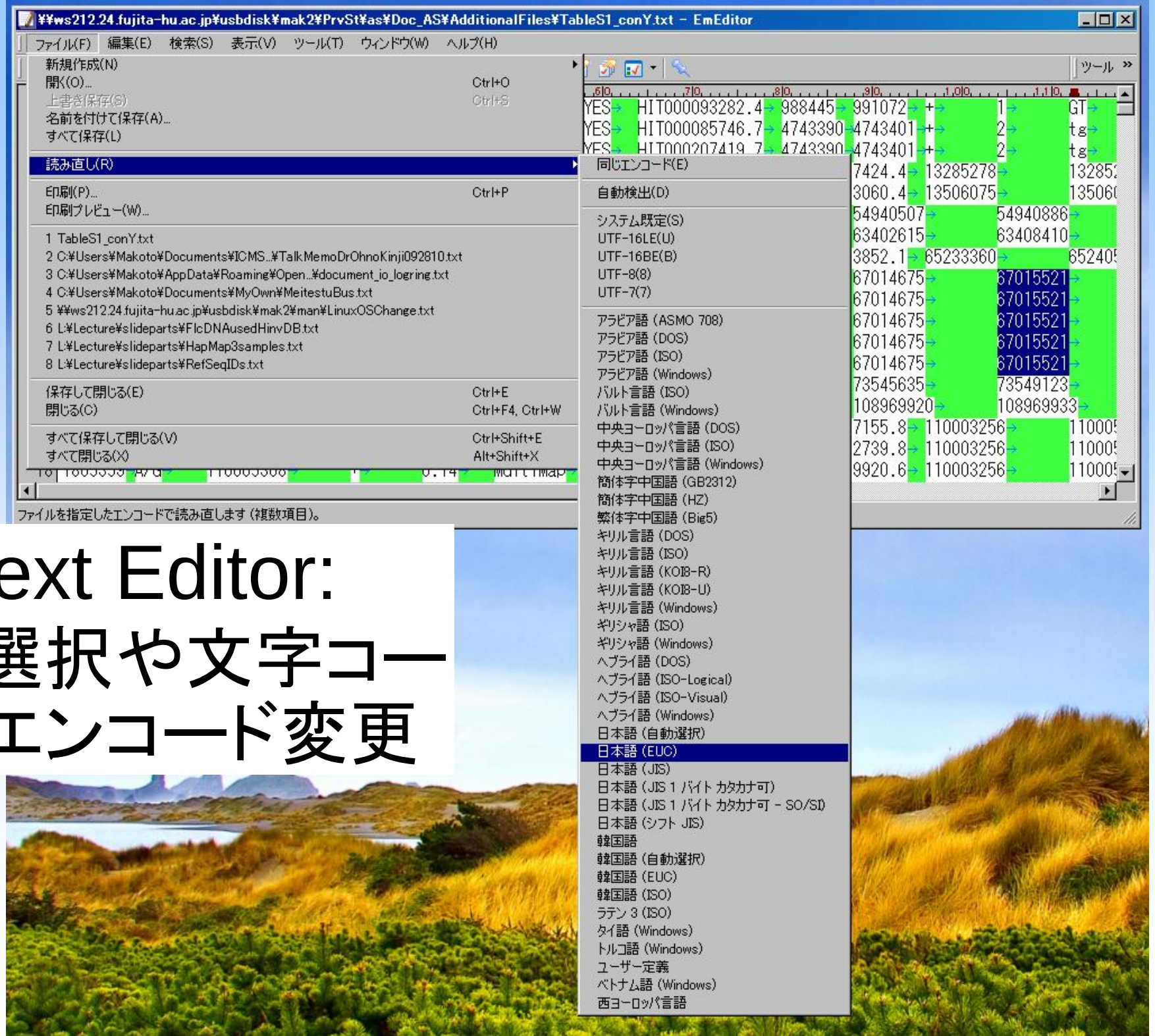
- テキスト・エディター
  - OSに付属しているものでも可能だが、正規表現使えますか？

テキスト・エディタ
正規表現を使えるか



要チェックツール：テキストファイル作成

- テキスト･エディター(便利な機能)
  - OSに付属しているものでも可能だが、正規表現使えますか？
    - タブ区切り　<-->コンマ区切りなど表の整形に便利
    - 意図しない改行を一括除去
  - 縦型の選択-->コピーやカット
  - 文字コードを変更して読みなおし



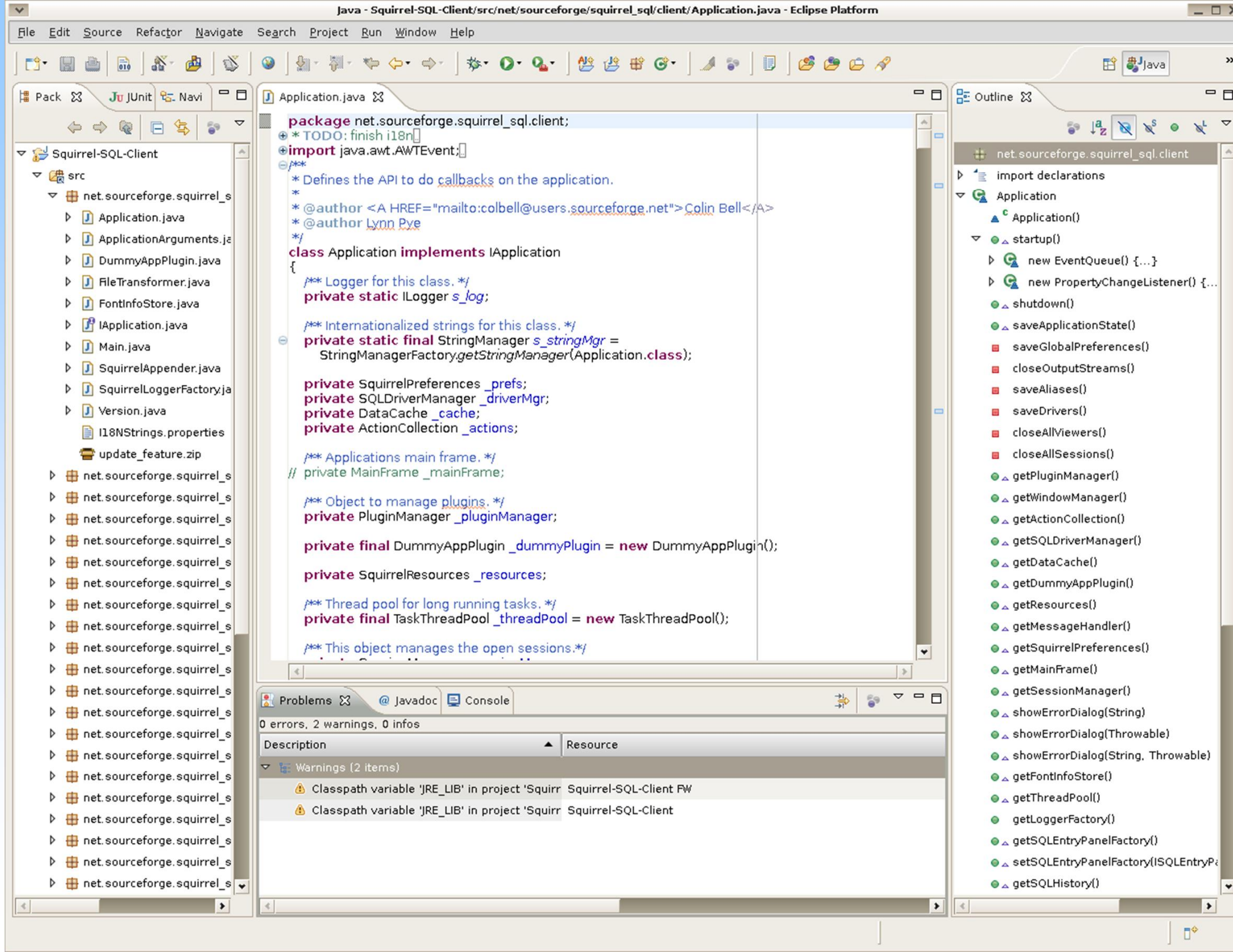
Text Editor: 縦型選択や文字コードのエンコード変更

要チェックツール：テキストファイル作成

- テキスト・エディター
  - 正規表現、縦型選択、コード変更などの機能が使える。
  - プログラミング支援機能付きはプログラム・エディターともよばれる
    - タグ挿入型：色違いで表示
  - 選ぶポイント：ショートカットキー
    - （新規習得か既知にこだわるか）
  - 例：K2Editor, 秀丸、vi, Emacs, Gedit, Unitext

Text Editorと統合開発環境(IDE)

- 統合開発環境
  - プログラム作成の際に必要な、テキスト・エディタ、コンパイラ、デバッガなどのツールを一つにしたもの。
  - 例：Eclips (IBM), KDevelop (KDE)
  - EmacsのようなText editorでもマクロを使うことでIDEとなる。
- （[参考]XML Parser: 実績あるものを選ぶ必要<---プログラマー）


Java - Squirrel-SQL-Client/src/net/sourceforge/squirrel_sql/client/Application.java - Eclipse Platform
File Edit Source Refactor Navigate Search Project Run Window Help
Pack JUnit Navi
Squirrel-SQL-Client
src
net.sourceforge.squirrel_s
Application.java
ApplicationArguments.ja
DummyAppPlugin.java
FileTransformer.java
FontInfoStore.java
IApplication.java
Main.java
SquirrelAppender.java
SquirrelLoggerFactory.ja
Version.java
I18NStrings.properties
update_feature.zip
Application.java
package net.sourceforge.squirrel_sql.client;
* TODO: finish i18n
import java.awt.AWTEvent;
/**
* Defines the API to do callbacks on the application.
*
* @author <A HREF="mailto:colbell@users.sourceforge.net">Colin Bell</A>
* @author Lynn Pye
*/
class Application implements IApplication
{
/** Logger for this class. */
private static ILogger s_log;
/** Internationalized strings for this class. */
private static final StringManager s_stringMgr =
StringManagerFactory.getStringManager(Application.class);
private SquirrelPreferences _prefs;
private SQLDriverManager _driverMgr;
private DataCache _cache;
private ActionCollection _actions;
/** Applications main frame. */
// private MainFrame _mainFrame;
/** Object to manage plugins. */
private PluginManager _pluginManager;
private final DummyAppPlugin _dummyPlugin = new DummyAppPlugin();
private SquirrelResources _resources;
/** Thread pool for long running tasks. */
private final TaskThreadPool _threadPool = new TaskThreadPool();
/** This object manages the open sessions.*/
Problems Javadoc Console
0 errors, 2 warnings, 0 infos
Description
Resource
Warnings (2 items)
Classpath variable 'JRE_LIB' in project 'Squirr
Squirrel-SQL-Client FW
Classpath variable 'JRE_LIB' in project 'Squirr
Squirrel-SQL-Client
Outline
net.sourceforge.squirrel_sql.client
import declarations
Application
Application()
startup()
new EventQueue() {...}
new PropertyChangeListener() {...
shutdown()
saveApplicationState()
saveGlobalPreferences()
closeOutputStreams()
saveAliases()
saveDrivers()
closeAllViewers()
closeAllSessions()
getPluginManager()
getWindowManager()
getActionCollection()
getSQLDriverManager()
getDataCache()
getDummyAppPlugin()
getResources()
getMessageHandler()
getSquirrelPreferences()
getMainFrame()
getSessionManager()
showErrorDialog(String)
showErrorDialog(Throwable)
showErrorDialog(String, Throwable)
getFontInfoStore()
getThreadPool()
getLoggerFactory()
getSQLEntryPanelFactory()
setSQLEntryPanelFactory(ISQLEntryP
getSQLHistory()

フリーソフト・シェアーウエアで探す

窓の杜

http://www.forest.impress.co.jp/

ベクター

http://www.vector.co.jp/

文字列の変換以外の良く使う変換

- IDの変換
- ゲノム位置の変換
  - 同種のゲノム（アッセンブリ）バージョン間

ID変換

- リンク自動管理システム（Hyperlink Management System, http://biodb.jp）
  - 生命科学の主要なデータベース間のリンクを自動で管理するツール。
  - IDの対応表を毎日自動で更新。
  - 方法１：webツールを使う。
  - 方法2:プログラムから利用する。

リンク自動管理システム：

ヒトの収録データベースとID

biodb.jp/index.cgi?lang=jp&tax=hsa

くはこちらへ。

話題のデータ(2012/8/8)
髄芽腫(ずいがしゅ)

Selected ID

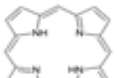
Homo sapiens IDs

[ヒトのIDs]

[マウスのIDs]

[ラットのIDs]

[ホヤのIDs]



[化合物のIDs]

H-InvDB

- Accession Number
- H-Inv transcript ID(HIT)
- H-Inv cluster ID (HIX)
- H-Inv protein ID (HIP)

NCBI

- Accession Number
- GeneID
- OMIM ID
- RefSeq ID
- New RefSeq Protein ID
- PubMed ID
- HUGO gene symbol

HUGO

- HUGO gene symbol

UniProt

- Accession Number
- UniProt Accession Number

H-GOLD

- Accession Number
- H-GOLD Marker ID

GeMDBJ

- HUGO gene symbol
- dbSNP rs#

PDBj

- PDB ID
- UniProt Accession Number

MutationView

- HUGO gene symbol
- OMIM ID

Ensembl

- HUGO gene symbol
- Ensembl Transcript ID (ENST)
- Ensembl Gene ID (ENSG)

GGDB

- HUGO gene symbol

fRNAdb

- Accession Number
- fRNAdb ID

HGPD

- Accession Number
- FLJ ID
- Clone ID

KEGG

- GeneID
- KEGG Gene ID
- KEGG Pathway ID

HPRD

- GeneID
- HPRD ID

NBRC

- Accession Number
- FLJ ID
- Clone ID

LEGENDA

- GeneID

Evola

- H-Inv transcript ID(HIT)

H-DBAS

- Accession Number
- H-Inv cluster ID (HIX)

G-Compass

- Accession Number
- H-Inv transcript ID(HIT)

CIPRO

- Accession Number
- CIPRO ID

H-ANGEL

- Accession Number
- H-Inv cluster ID (HIX)

FLJ Human cDNA

- Accession Number
- FLJ ID
- Clone ID

VarySysDB

- Accession Number
- H-Inv transcript ID(HIT)

PDBeChem

- PDB ID
- PDBeChem ID

DrugBank

Wikipedia Human proteins

JSNP DATABASE

リンク自動管理システム：プログラムから利用する

Hyperlink management sy...

biodb.jp/help/index

2. 自動管理によるリンクの設定方法

(1) 単純なリンクの方法

本サービスは、

```
http://biodb.jp/hfs.cgi?id=[ID]&type=[ID Type]&db=[Database name]
```

のように [http://biodb.jp/hfs.cgi?]にパラメータを渡すだけで利用できます。設定するパラメータは以下の通りです。

1. [ID]
   変換するIDを指定します。

2. [ID Type]
   指定したIDの形式を指定します。
   利用できるIDのリスト

3. [Database name]
   転送先のデータベース、ビューアーを指定します。
   利用できるデータベースのリスト

それぞれこの画面の下に記載可能なリストがある。

サンプル

Accession Number「BC053657」からH-InvDB (Transcript view)へリンクする場合。

```
http://biodb.jp/hfs.cgi?id=BC053657&type=ACC_ID&db=TRANSCRIPTVIEW
```

HTMLには以下のように記述します。

```
<a href='http://biodb.jp/hfs.cgi?id=BC053657&type=ACC_ID&db=TRANSCRIPTVIEW'>BC053657</a>
サンプル：BC053657
```

ヒトのID一覧

No.	形式	概要

ゲノム・バージョンの変換

http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Convert

- ゲノムの座標(coordinates)は現在もアッセンブリ（バージョン）の変更に伴い変更されている。
  - scaffold間のgapが埋まる、contigの向きや重複領域の修正などのため。
- UCSCでは2つの変換ツールがある。
  - BLAT
  - Lifting coordinates
    - Web-based
    - Command-line

genome versionとは

http://genome.ucsc.edu/FAQ/FAQreleases.html

List of UCSC genome releases

Question:
"How do UCSC's release numbers correspond to those of other organizations, such as NCBI?"

Response:

EnsemblもGRChを使用

SPECIES	UCSC VERSION	RELEASE DATE	RELEASE NAME	STATUS
VERTEBRATES				
Human	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived
	hg10	Dec. 2001	NCBI Build 28	Archived
	hg8	Aug. 2001	UCSC-assembled	Archived
	hg7	Apr. 2001	UCSC-assembled	Archived
	hg6	Dec. 2000	UCSC-assembled	Archived
	hg5	Oct. 2000	UCSC-assembled	Archived
	hg4	Sep. 2000	UCSC-assembled	Archived
	hg3	Jul. 2000	UCSC-assembled	Archived
	hg2	Jun. 2000	UCSC-assembled	Archived (data set only)
	hg1	May 2000	UCSC-assembled	Archived (data set only)
Cat	felCat4	Dec. 2008	NHGRI catChrV17e	Available
	felCat3	Mar. 2006	Broad Institute Release 3	Available
Chicken	galGal3	May 2006	WUSTL Gallus-gallus-2.1	Available
	galGal2	Feb. 2004	WUSTL Gallus-gallus-1.0	Available
Chimp	panTro2	Mar. 2006	CGSC Build 2 Version 1	Available
	panTro1	Nov. 2003	CGSC Build 1 Version 1	Available
Cow	bosTau4	Oct. 2007	Baylor College of Medicine HGSC Btau_4.0	Available

- Blat
- Table Browser
- Gene Sorter
- Genome Graphs
- In-Silico PCR
- LiftOver
- VisiGene
- Other Utilities

Lift Genome Annotations

This tool converts genome coordinates and g…en assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selecte… direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 200…hen from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome: Human

Origina… Mar. 20…

New Genome: Human

New Assembly: Feb. 2009 (GRCh37/hg19)

Minimum ratio of bases that must remap: 0.95

BED 4 to BED 6 Options

Allow multiple output regions:

Minimum hit size in query: 0

Minimum chain size in target: 0

BED 12 Options

Min ratio of alignment blocks or exons that must map: 1

If thickStart/thickEnd is not mapped, use the closest mapped base:

Paste in data (BED or chrN:start-end formats):

```
chr2    241280126       241280164
chr1    148864770       148864812
chr16   66824121        66824165
chr3    50304975        50305023
chr1    151271601       151271650
chr9    139230290       139230340
chr17   45546669        45546720
chr3    50304862        50304917
chr15   97616970        97617025
chr6    88467944        88468000
```

Submit

Clear

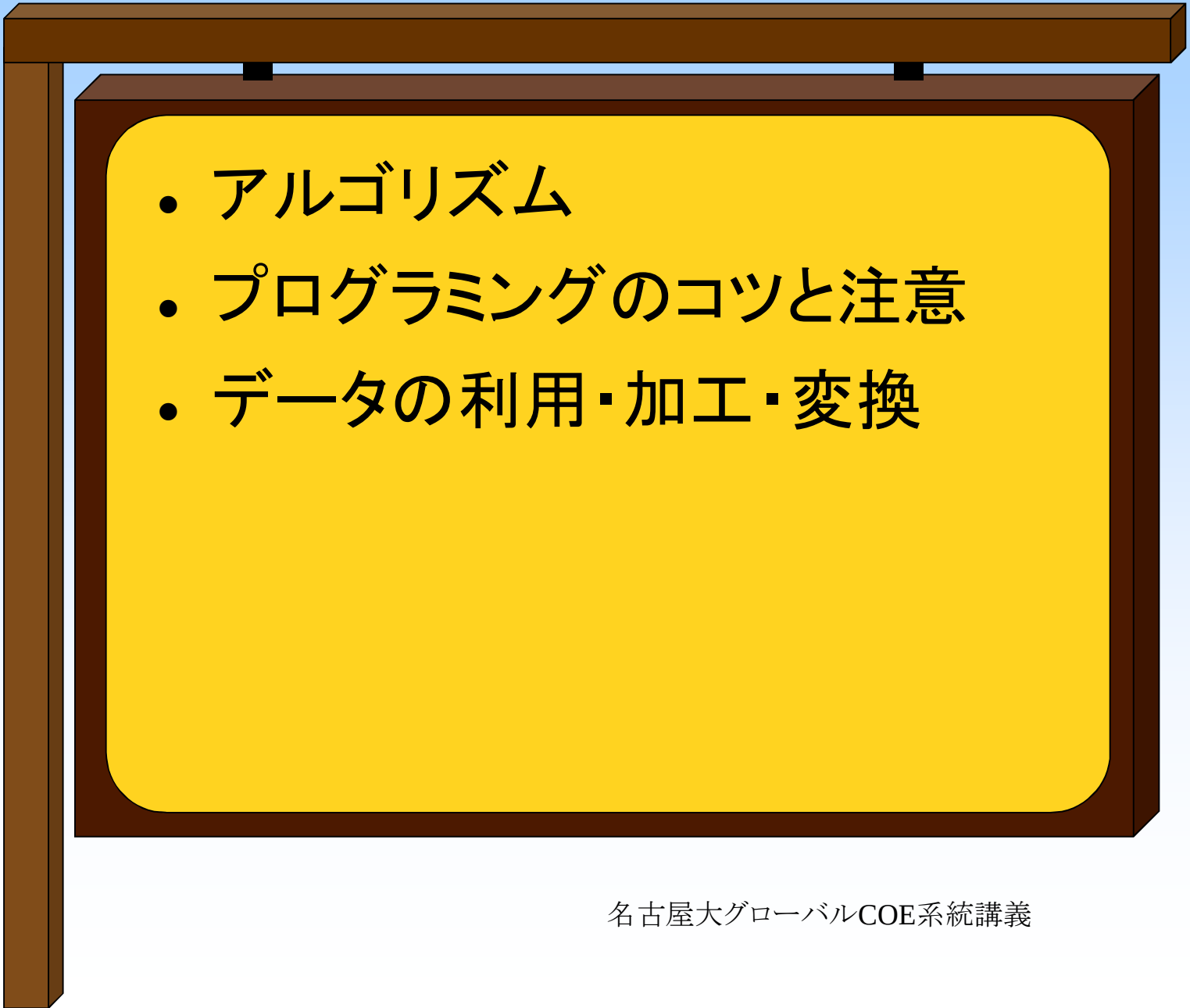
Or upload data from a file (BED or chrN:start-end in plain text format):

ファイルを選択 選択されていません Submit File

Command Line Tool

To lift genome annotations locally on Linux systems, download the *liftOver* executable and the appropriate chain file. Run *liftOver* with no arguments to see the usage message.

生命科学者にとってのプログラミング



- アルゴリズム
- プログラミングのコツと注意
- データの利用・加工・変換

algorithm アルゴリズム

- (問い)アルゴリズムとは
- (例)3，7，1，9，4を昇順に並び変えよ。
- とくに考えることもなく、やればできる。(人間)
- コンピュータにやらせるには？

algorithm:並べ替え1例

- (例)3，7，1，9，4を昇順に並び変えよ。
- [方法1：基本方針]

  1.左から順にn番目、右端をm番目とする

  2.n番目とn＋1番目を比較

  3.右隣のほうが大きければ互いに位置を交換

  4.n＝1からn＝m－1まで繰り返す。(<--最大値が右端)

  5.もう一度、n＝1からn＝m－2まで繰り返す。

  6.上記操作(5)をm－2回繰り返す。

algorithm アルゴリズム

- (問い)アルゴリズムとは
- (答え)コンピュータが実行可能な手順のこと
- 曖昧性がない
- 終了の仕方が必ず明記

生命科学者にとってのアルゴリズム

- 情報系の人たちとの打ち合わせに際し、知っていることが必要な概念
- プログラムが不得手でも、自分の考えた手順をアルゴリズムにできれば、
  - 助けがあればオリジナルなツールができる。
    - 自動化、効率化、人為ミスの解決

プログラミング言語

- プログラミング言語でアルゴリズムを記述
  - ---›コンピュータに指示をだす。
- 生命科学者にとって、プログラミング言語を学ぶべきか？
  - あなたの研究に自動化が必要ですか？
  - それはどの計算ですか、処理(判断)ですか？
  - 代表的プログラミング言語の概要だけの知識は？

生命科学者がプログラミング言語を独学するとき

- 独学の開始時(一般的に)
  - 若い方がよい
    - 遅すぎると言うことはない
  - 時代やテーマ、用途によっては、将来別の言語に乗り換えるかも
    - プログラミング言語：やること同じだからそれほど心配するな
- 情報学者やプログラマーとしての訓練を受けたものと独学者の違いは
  - あらゆるデータが渡されることを常に考える。

プログラム言語の分類

- コンパイラ方式
  - 機械語（バイナリ）に翻訳（コンパイル）してから実行
- インタプリタ方式
  - 実行時に逐次機械語に翻訳
- どちらにも分類できない言語もある

生命科学に関連深いプログラム言語

- **C/C++**
- **Java**: http://java.sun.com/
- **.NET**: http://www.microsoft.com/NET/
  - **C#**
  - **Visural Basic .NET**
- **Perl**: http://www.perl.com/
- **PHP**: http://www.php.net/
- **Python**: http://www.python.org/
- **Rubby**: http://www.ruby-lang.org/ja/
- **R言語**: http://www.r-project.org/

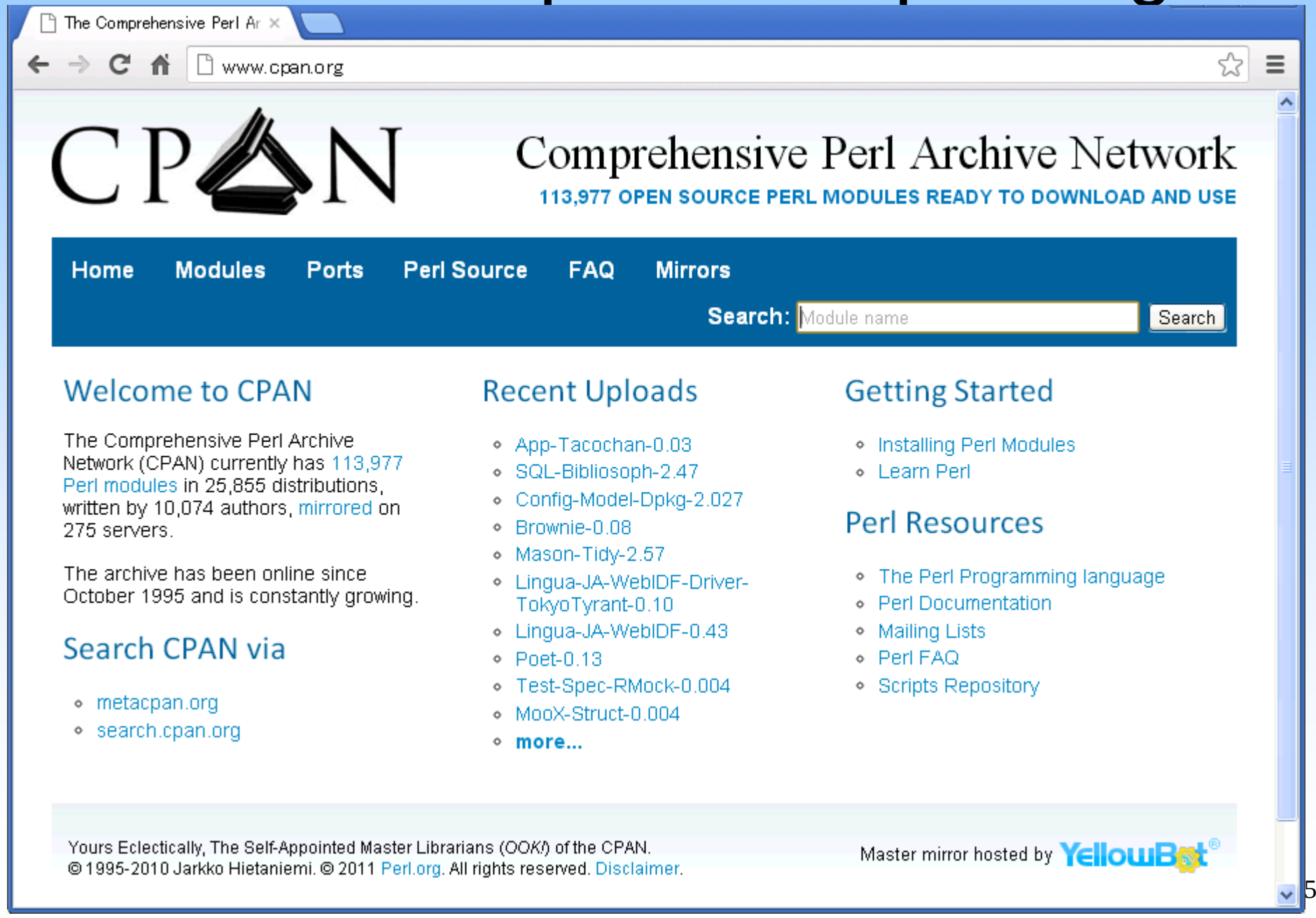
プログラム言語習得：基本は共通

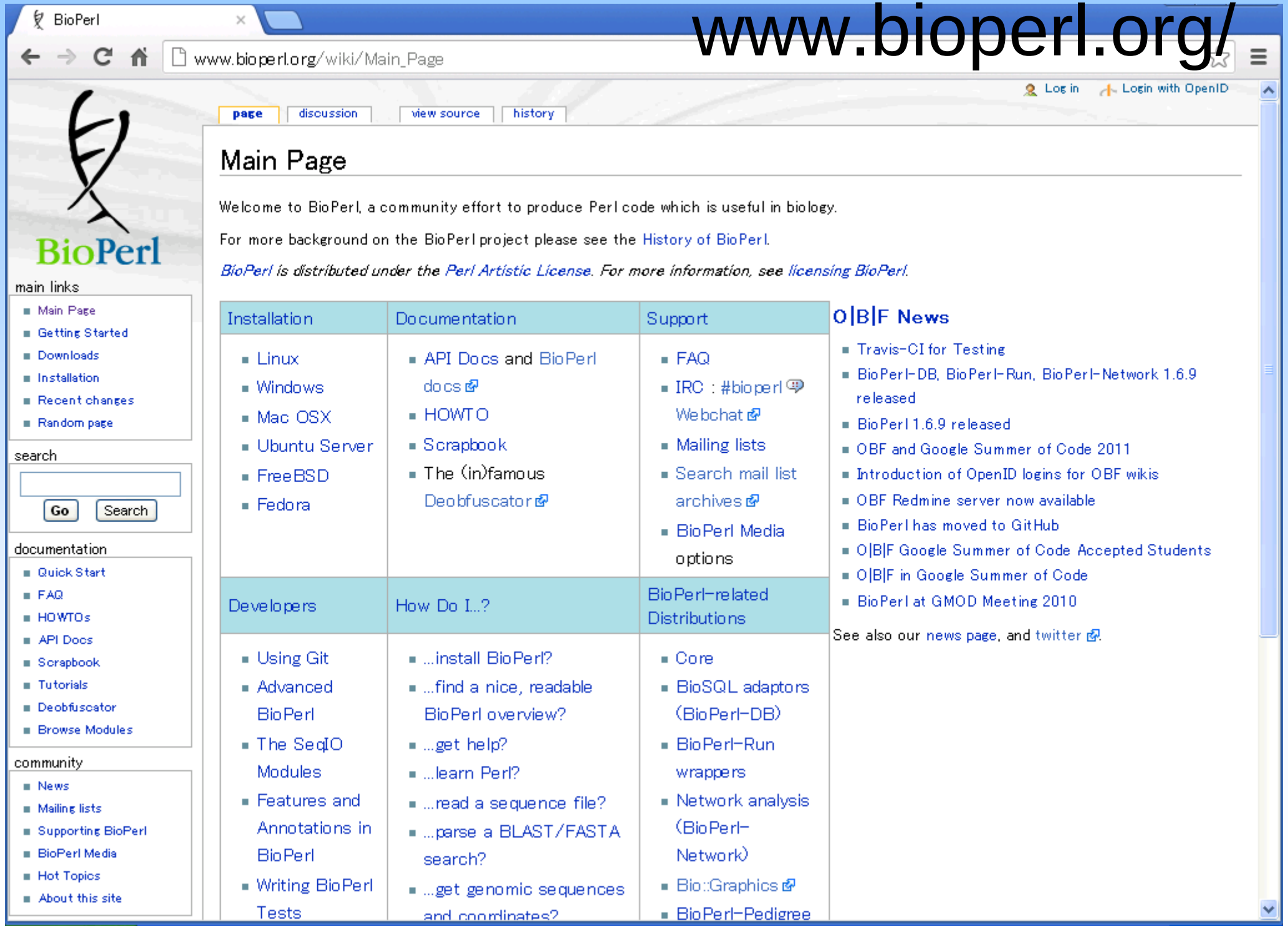
- データ型についての特性
- 変数宣言
- 変数への代入
- 四則演算
- ファイルの読み書き
- オブジェクト（属性(データ)と操作(メソッド)の集合）

libruary (ライブラリ)とは

- 汎用性の高い複数のプログラムを、他のプログラムから利用できるように、一つにまとめたもの。
  - 関数やサブルーチンの集合。
  - ライブラリーそのものは単独で機能しない。他のプログラムの部品となる。
  - 目的と出来の良さを判断する力量が必要。
- (例)Perl言語の場合:
  - Perl言語一般のライブラリ-->CPAN
    - http://www.cpan.org/
  - 生物情報のライブラリ-->BioPerl
    - www.bioperl.org/

http://www.cpan.org/




BioPerl
www.bioperl.org/wiki/Main_Page
www.bioperl.org/
Log in
Login with OpenID
page
discussion
view source
history
BioPerl
main links
Main Page
Getting Started
Downloads
Installation
Recent changes
Random page
search
Go
Search
documentation
Quick Start
FAQ
HOWTOs
API Docs
Scrapbook
Tutorials
Deobfuscator
Browse Modules
community
News
Mailing lists
Supporting BioPerl
BioPerl Media
Hot Topics
About this site
Main Page
Welcome to BioPerl, a community effort to produce Perl code which is useful in biology.
For more background on the BioPerl project please see the History of BioPerl.
BioPerl is distributed under the Perl Artistic License. For more information, see licensing BioPerl.
Installation
Linux
Windows
Mac OSX
Ubuntu Server
FreeBSD
Fedora
Documentation
API Docs and BioPerl docs
HOWTO
Scrapbook
The (in)famous Deobfuscator
Support
FAQ
IRC : #bioperl Webchat
Mailing lists
Search mail list archives
BioPerl Media options
Developers
Using Git
Advanced BioPerl
The SeqIO Modules
Features and Annotations in BioPerl
Writing BioPerl Tests
How Do I...?
...install BioPerl?
...find a nice, readable BioPerl overview?
...get help?
...learn Perl?
...read a sequence file?
...parse a BLAST/FASTA search?
...get genomic sequences and coordinates?
BioPerl-related Distributions
Core
BioSQL adaptors (BioPerl-DB)
BioPerl-Run wrappers
Network analysis (BioPerl-Network)
Bio::Graphics
BioPerl-Pedigree
O|B|F News
Travis-CI for Testing
BioPerl-DB, BioPerl-Run, BioPerl-Network 1.6.9 released
BioPerl 1.6.9 released
OBF and Google Summer of Code 2011
Introduction of OpenID logins for OBF wikis
OBF Redmine server now available
BioPerl has moved to GitHub
O|B|F Google Summer of Code Accepted Students
O|B|F in Google Summer of Code
BioPerl at GMOD Meeting 2010
See also our news page, and twitter.

生物情報のライブラリ
その他の言語

- BioPHP
  - http://www.biophp.org/
- BioRuby
  - http://bioruby.org/
- BioJava
  - http://www.biojava.org/
- BioPython
  - http://biopython.org/
- Bioconductor
  - http://www.bioconductor.org/

解析作業の自動化

- 自動化が効果的な作業とは？
  - 長大な繰り返し作業
  - 入力、ファイルの選択、等人為的なミスが重大な作業
- 代表的な自動化方法の紹介
  - プログラム言語を利用
  - OSのコマンド処理を利用: shell script
  - マクロを利用
  - web service (ウェブ・サービス)を利用

shell script
（シェル・スクリプト）

- Linux（Mac OS Xもふくむ）
- 実行するコマンドをあらかじめテキストファイルに保存して、それを実行できる形にするもの。
- 利用者が多く、ネット上に多数の教則サイトがある。

- Windowsのbatch fileと似た位置づけ。

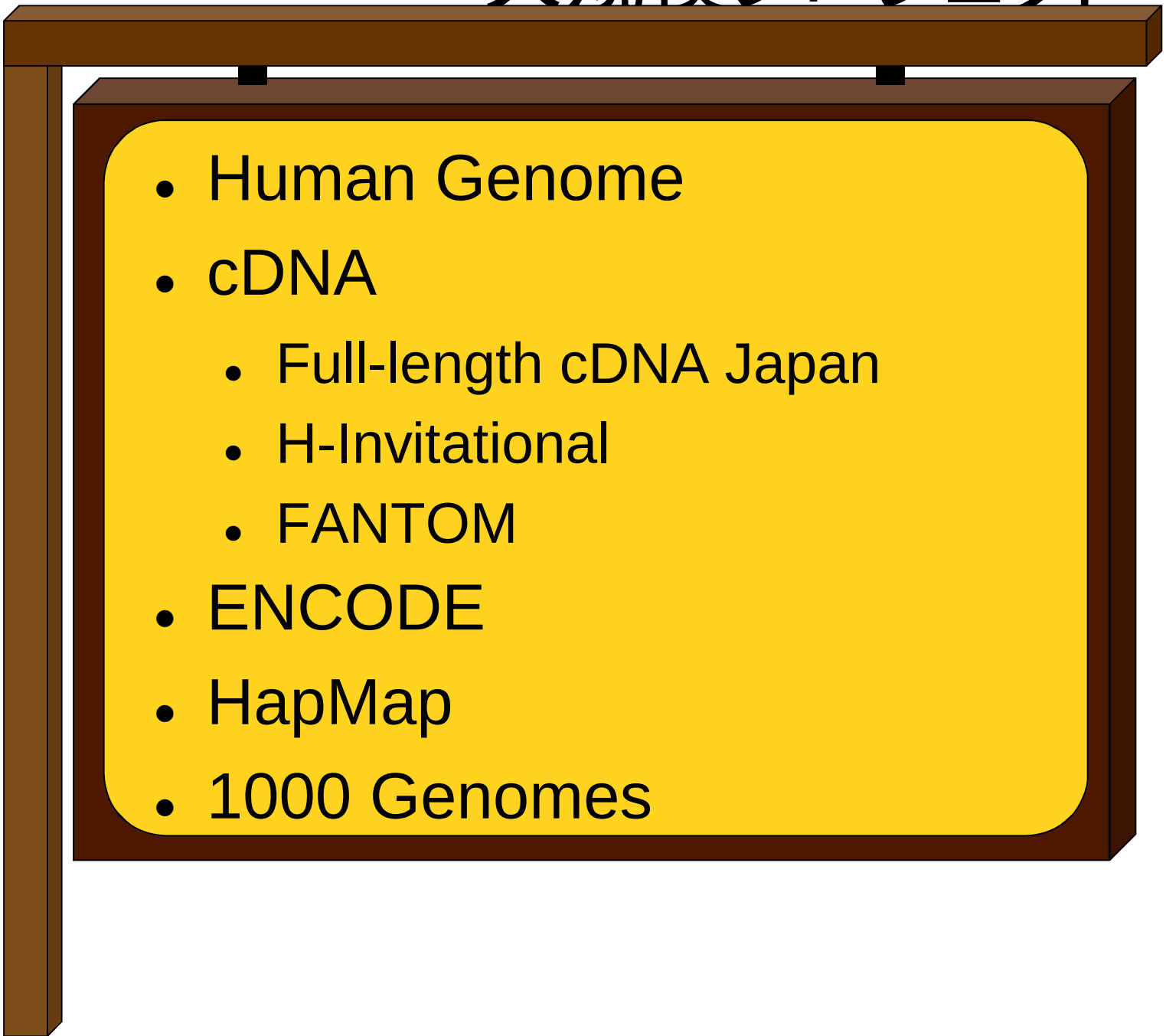
Webサービス（ウェブサービス）

- Webツール　vs. Webサービス
- Webツールは人が手で入力し、眼で理解する
- Webサービスはプログラムが入力し、理解する
- 生物情報で良く利用されるWebサービスAPI（application programming interface, アプリ開発に汎用的機能を共通利用するための手法）
  - SOAP
  - REST
  - これらは一般的にも良く利用されている
    - amazon, 楽天

生命科学系Webサービス

- EBI
  - http://www.ebi.ac.uk/Tools/webservices/
  - 英語ではあるが、丁寧なマニュアルが充実
- JST BIRD & DDBJ Web API for Biology (WABI)
  - http://wabi.ddbj.nig.ac.jp/CookBook_jp/
  - 日本語と英語のサイトがある。
- H-InvDB
  - http://hinv.jp/hinv/hws/doc/index_jp.html
- VarySysDB
  - http://hinv.jp/hinv/hws/varysysdb/doc/index_ja.html

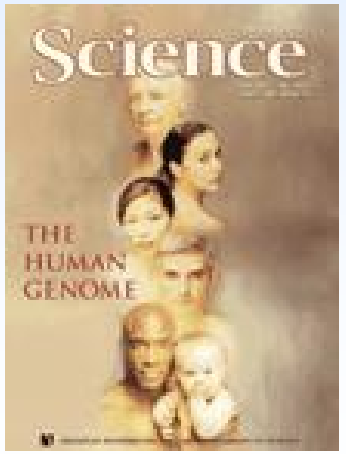
生物情報に影響を与えた大規模プロジェクト



- Human Genome
- cDNA
  - Full-length cDNA Japan
  - H-Invitational
  - FANTOM
- ENCODE
- HapMap
- 1000 Genomes

ゲノムプロジェクト

- 1987 和田昭充、機械による自動配列決定戦略提案
- 1990 アメリカ、イギリス、日本、ドイツ、フランスを中心とする国際コンソーシアム(INC)合意
  - 統一戦略なし、map作成(YAC,BAC,PAC)、新技術開発、
- 1996 バミューダ会議
  - ルール（参加、実施前公表、分担、2005年完成目標）
- 1998 Celera社３年間でヒトゲノム概要版解読宣言
  - [INC:目標変更、BAC概要版データは即座に公共DBにて公開]
- 2000 Jun. 概要版完成宣言
  - [INC] Nature (2/15/2001)
  - [Celera] Science (2/16/2001)
- 2003 Apr. [INC]全染色体解読完了宣言
  - [INC] Nature (10/21/2004)



結果の利用：階層的 vs. 全ゲノム

- ハプロタイプ由来 vs. ５人のコンセンサス
  - [INC] BACクローンのそれぞれが単一ハプロタイプ(ただし多くの不特定個人から作成されたBACライブラリー)
  - [Cerela] ５人の個人由来（ただしJ. Craig Venter氏由来が最も多い）のBAC配列の重なりを頼りに作られたコンセンサス配列

cDNA projects

- EST project
  - J. Craig Venter (USA)が先駆け
  - dbEST (release 120701, NCBI)の統計ではヒトが8,692,773エントリで最大。
- Full length project
  - H-Inv
  - FANTOM

Human Full-Length cDNA Annotation Invitational (http://hinv.jp)

- ヒト完全長cDNAプロジェクト日本(FLJ)で日本は世界をリードするコレクションを有していた(eg., Nature Genetics 36, 40 - 45, 2004)。
- H-Invitational(お台場マラソン)　2002年8月25日-9月3日

項目数	提供機関
2,031	Kazusa DNA Research Institute
397	Full Length cDNA Japan / Kazusa DNA Research Institute
6,374	Full Length cDNA Japan / 東京大学医科学研究所
22,047	Full Length cDNA Japan / Helix Research Institute, Inc.
9,212	German Human cDNA Project（DKFZ、ドイツ）
15,600	Mammalian Gene Collection（NCI/NIH、アメリカ）
758	Human cDNAs（Chinese National Human Genome Center、中国）
56,419	Total

- ヒト完全長cDNAにヒト全mRNAのデータを加えて、公開した。(2006)
  - 現在でも、完全長データに絞った検索が可能
  - (eg., AS subDB, H-DBAS, http://hinv.jp/h-dbas/)

FANTOM

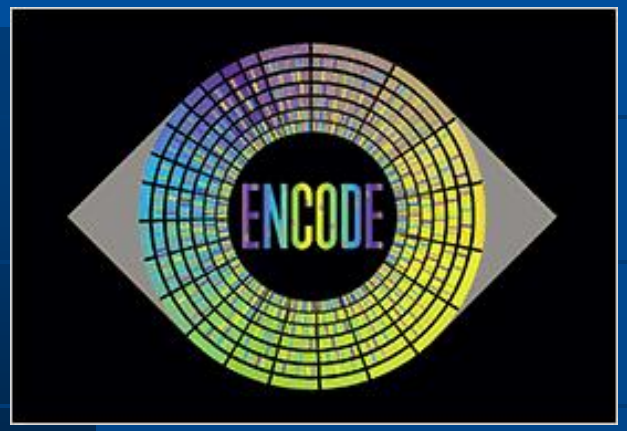
iPS細胞樹立に、cDNA DBによる、細胞分化遺伝子制御ネットワークの解明が貢献。

http://www.osc.riken.jp/contents/fantom/

- FANTOMは、理化学研究所のマウスエンサイクロペディアプロジェクトで収集された完全長cDNAのアノテーションを目的とした国際研究コンソーシアム
- 2000年に結成され、2008年12月現在、参加国数15カ国、参加機関は国内外合わせて51機関
- FANTOM データベース　http://fantom.gsc.riken.jp/
- おもな研究ツール
  - Genome Browser
  - EdgeExpressDB：制御関係と遺伝子やプロモーター活性との関連を表示
  - SwissRegulation：モチーフ活性の応答解析
- 完全長cDNA クローンバンク

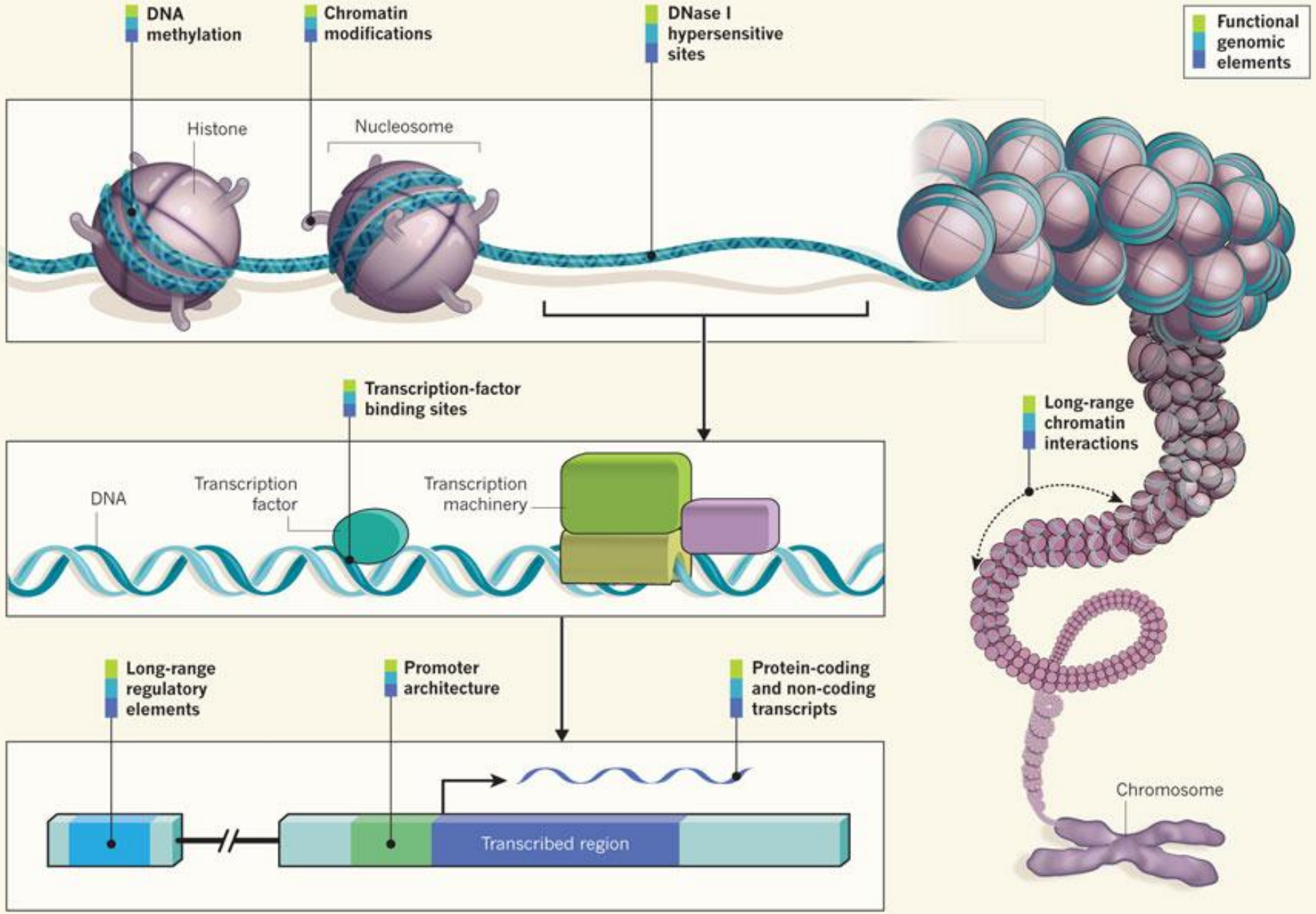
ENCODEプロジェクト

ENCODE, the Encyclopedia of DNA Elements, is a project funded by the National Human Genome Research Institute to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence. Thanks to the identification of these functional elements, 80% of the components of the human genome now have at least one biochemical function associated with them. This expansive resource of functional annotations is already providing new insights into the organization and regulation of our genes and genome.



http://www.nature.com/encode/#/threads

“Now the ENCODE consortium presents a menu of 1,640 genome-wide data sets prepared from 147 cell types, providing a six-course serving of papers



in *Nature*, along with many companion publications in other journals.”

Threads

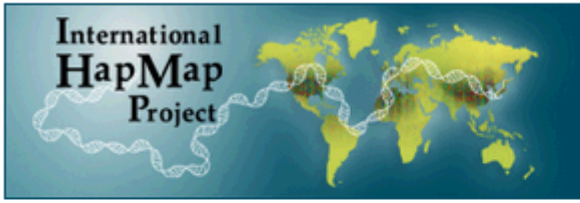
The 30 papers published across three different journals: *Nature*, *Genome Research* and *Genome Biology*

1 Transcription factor motifs
2 Chromatin patterns at transcription factor binding sites
3 Characterization of intergenic regions and gene definition
4 RNA and chromatin modification patterns around promoters
5 Epigenetic regulation of RNA processing
6 Non-coding RNA characterization
7 DNA methylation
8 Enhancer discovery and characterization
9 Three-dimensional connections across the genome
10 Characterization of network topology
11 Machine learning approaches to genomics
12 Impact of functional information on understanding variation
13 Impact of evolutionary selection on functional regions

http://www.nature.com/encode/about/nature-encode-explorer

HapMap project hapmap.org

hapmap.ncbi.nlm.nih.gov



International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants

Project Data

HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #3 (Phase 3 - genotypes & frequencies)
HapMap Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser release#24 (Phase 1 & 2 - full dataset)
GWAs Karyogram
HapMart
HapMap FTP
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

Useful Links

News

- 2011-06-13: **HapMap help desk announcement**

  There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov did not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent emails to the HapMap help desk in the past several weeks. Sorry for the inconvenience.

- 2011-04-20: **Hapmap help desk service interruption notice**

  There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.

- 2011-02-02: **Haploview issues with rel 28 data**

  Recently, there are several questions about Haploview data format errors when users tried to analyze HapMap release 28 data. The current Haploview version (4.2) does not recognize the new individuals in release 28 and the software will generate an error similar to "Hapmap data format error: NA18876" when trying to open the data.

  Haploview is developed and maintained by an organization different from HapMap. Please contact Haploview help desk (haploview@broadinstitute.org) for questions specific to this software.

- 2011-01-19: **HapMap phase II recombination rate on GRCh37**

  The liftover of the HapMap II genetic map from human genome build b35 to GRCh37 is available. Data is **available for bulk download**.

- 2010-08-18: **HapMap Public Release #28**

  Genotypes and frequency data in hapmap format are now available for data in merged HapMap phases I+II+III release #28 (NCBI build 36, dbSNP b126). Data is **available for bulk download** and also **available for browsing**. Click here to read the latest **release notes**.

- 2010-05-28: **HapMap3 Public Release #3**

  Genotypes and frequency data in hapmap format are now available for data in HapMap phase 3 release #3 (NCBI build 36, dbSNP b126). Data is **available for bulk download** and also **available for browsing**. Click here to read the latest **release notes**.

HapMap 特徴
http://hapmap.ncbi.nlm.nih.gov/

- 継続されているprojectである。
  - 最新 (2010-08-18):phase Ⅲ, release 28
- “phase”という語に注意
  - phases Ⅰ Ⅱ Ⅲ vs. phased haplotype
- [応用]ヒトのLDとその集団間差異を明らかにした。
- データのダウンロード、および、
- 多型に関する、様々なブラウザ表示およびツール群をweb通じ提供している。

HapMapのおもな成果

- Phase I / IIでは、4集団270個体について、ゲノムワイドに約3.6M個のSNPsがタイピングされた。同時に、用いたサンプル中の全多型は10M個と推定された。LD領域マップの作成とヒトの疾患に関連する数百遺伝子座を同定できた(Nature 449, 851-861, 2007)。
- Phase IIIでは、7集団を加えた計1184個体で
  - 約1.6M個のSNPsをタイピング、
  - ENCODE領域の配列決定、
  - 約1.6K箇所のCNP領域の同定、
- を行い、稀な変異の集団間差異とその解析方法を示した(Nature 467, 52ï 58, 2010)。

Samples of HapMap3

label	population sample	number of samples	QC samples
ASW	African ancestry in Southwest USA	90	83
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	180	165
CHB	Han Chinese in Beijing, China	90	84
CHD	Chinese in Metropolitan Denver, Colorado	100	85
GIH	Gujarati Indians in Houston, Texas	100	88
JPT	Japanese in Tokyo, Japan	91	86
LWK	Luhya in Webuye, Kenya	100	90
MEX	Mexican ancestry in Los Angeles, California	90	77
MKK	Maasai in Kinyawa, Kenya	180	171
TSI	Toscans in Italy	100	88
YRI	Yoruba in Ibadan, Nigeria	180	167
Total		1301	1184

1000 Genomes(千人ゲノム)

http://www.1000genomes.org/

- 多数の個人ゲノム解読により、ヒト集団の多様性を明らかにする。
- [応用]GWAS
- Pilot Project

Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centers	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

- Main Project
  - 2000 samples at 4X
    - 1st set: 1101 samples from 12 populations
    - 2nd set: 899 samples from 10 populations
    - 3rd set: 779 samples from 11 population

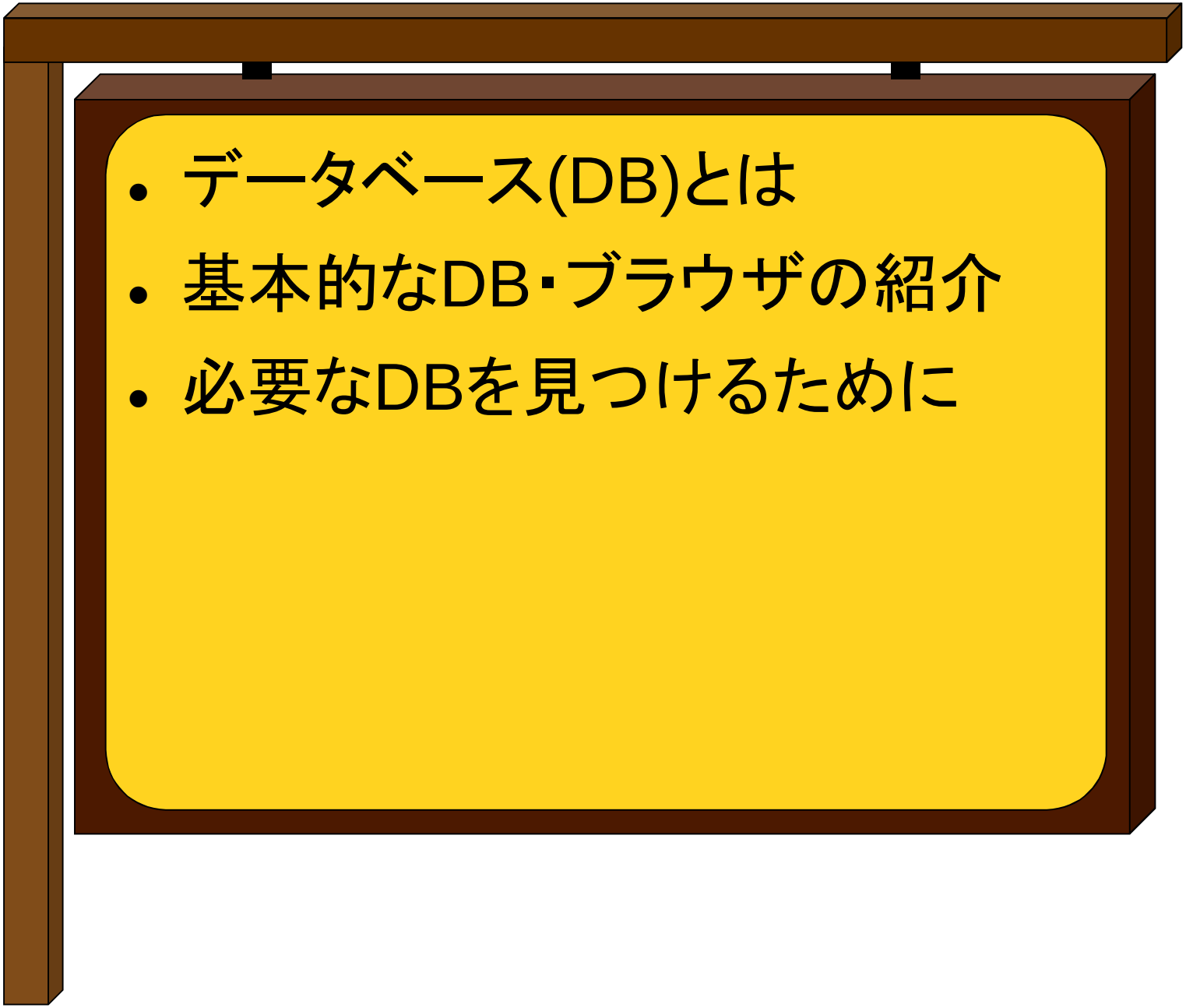
1000 genomesの結果と特徴

- viewer
http://browser.1000genomes.org/

- 新Data format (vcf)を創出
  VCFtools(コマンドベースのプログラム)を使うことにより、大規模なデータを加工することができる。
  http://vcftools.sourceforge.net/

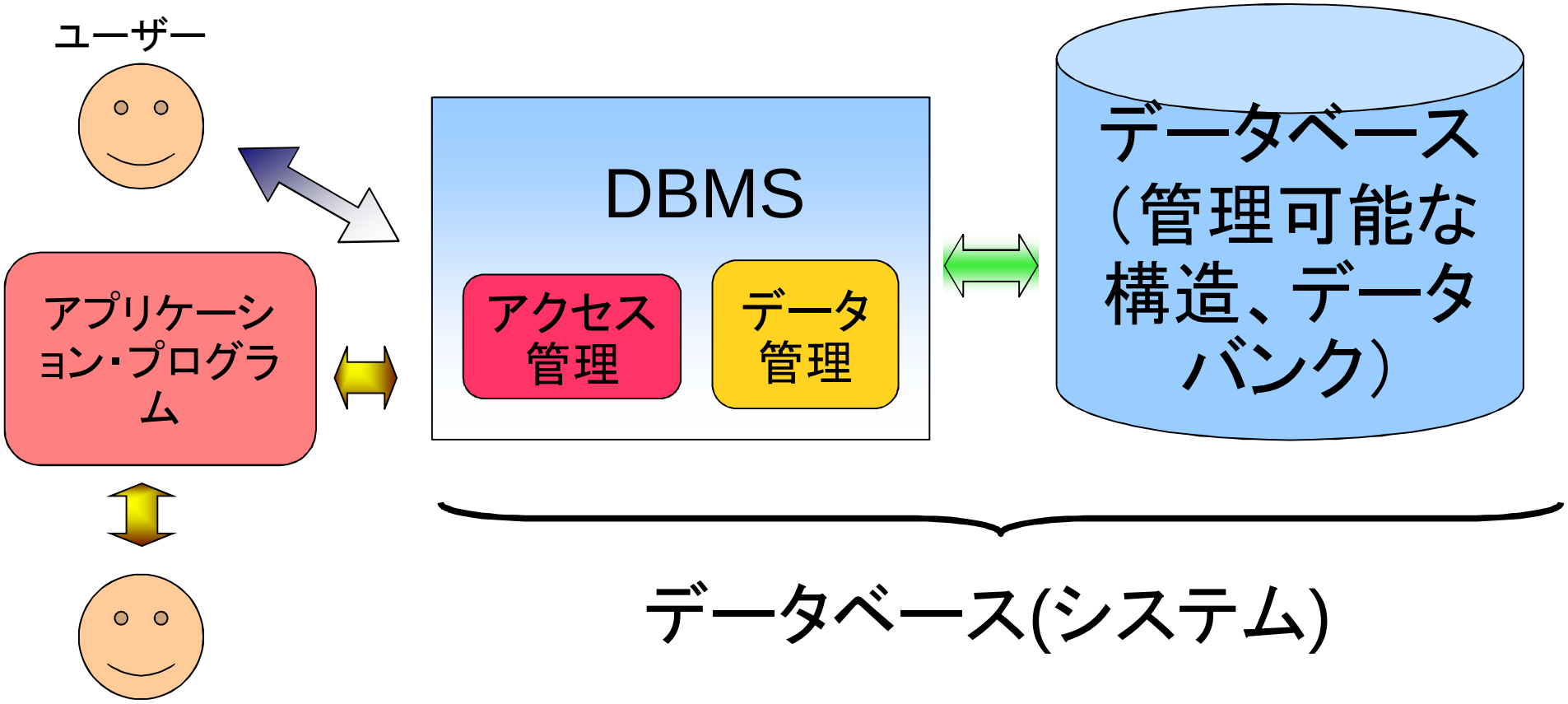
- Amazon Web Services (AWS)を通じてもalignment fileを配布
http://s3.amazonaws.com/1000genomes

データベース



データベースとは？(問い)

- 「～をデータベース化して」vs.「～をリスト化して」
- データベース管理システム (DBMS)



さらにDBMS

- DBMSが無いと、
  - プログラムから直接、ファイルに対してデータの保存や読み込みを行う
  - 個々のファイルのデータ構造を予め知っていないとアクセスできない
  - 複数のユーザーやアプリケーションに対応できない
  - 安全性・安定性に限界
- DBMS：構造の設計から管理・運用を行う
- 現場では意外と聞かないことば
  - リレーショナル・データベース（RDBMS）が一般的

リレーショナル・データベース（RDB: Relational Data Base）の概要（1）

- データを表（テーブル）形式で表し、複数の表にリレーションシップ（関係付け）を結ぶことができる。
- 通常、元データの表はタブやコンマなどで区切ったテキストファイル形式。

HIT000044994	13	-2	NHEJ1	rs10498064	A/T	T	fwd/T Y
HIT000015558	8	-1	KIAA0802	rs632423	C/G	C	fwd/B Y
HIT000041481	8	-1	C13orf18	rs2478044	C/G	G	fwd/T Y
HIT000028466	21	-1	IL16	rs4778639	G/T	T	fwd/B Y

record（行）

field, column（列）

Column Name of Table S2

#1 HIT ID: H-InvDB Transcript ID
#2 Intron No: _th intron in the transcript
#3 SitePos: see Figure 1a
#4 GeneSybl: HUGO gene symbol
#5 rs ID: rs SNP ID in dbSNP
#6 rsAlleles: alleles in dbSNP
#7 Ancestral Allele: ancestral allele estimated by NCBI and shown in dbSNP
#8 Ori/Str(rs): orientation and strand of the SNP, see ftp://ftp.ncbi.nih.gov/snp/database/Illumina_top_bot_strand.note.txt
#9 number of locations SNP mapped: 'Y'='OneMap' or 'N'='MultiMap', see List of Abbreviation

Shimada MK *et al. BMC Evol Biol*, 10:122, 2010. Additinal files.

リレーショナル・データベース（RDB: Relational Data Base）の概要（2）

- キー(key)の重要性：
  - 主キー(Primary key)：レコードを一意に指定するフィールド（通常非冗長ID）
  - 外部キー(Foreign key):2つの表を結び付けるキー
    - 応用例（DB間を結び付ける）, http://biodb.jp/
- SQLと呼ばれるデータベース（問い合わせ）言語が標準化されている。
- 代表的実装例
  - 商用：Oracle Database, Microsoft SQL Server
  - オープンソース： MySQL, PostgreSQL, BerkeleyDB

データベースとは(答え)

- 「様々な目的を考慮して整理整頓されたデータの集まり」である。
- 設計思想をもつ（DBMSにて体現）。
  - 生物情報DBでは普通、
    - 検索機能、
    - web利用が前提、
    - リンク付け、
    - viewerやtool群とセットになっている。
- 始まりは、第二次大戦後の米軍が点在する情報を集約し、一か所にアクセスするだけで、様々な情報が得られるようにした「情報基地」

DBに国境は？

米国系：NCBI, National Center for Biotechnology Information

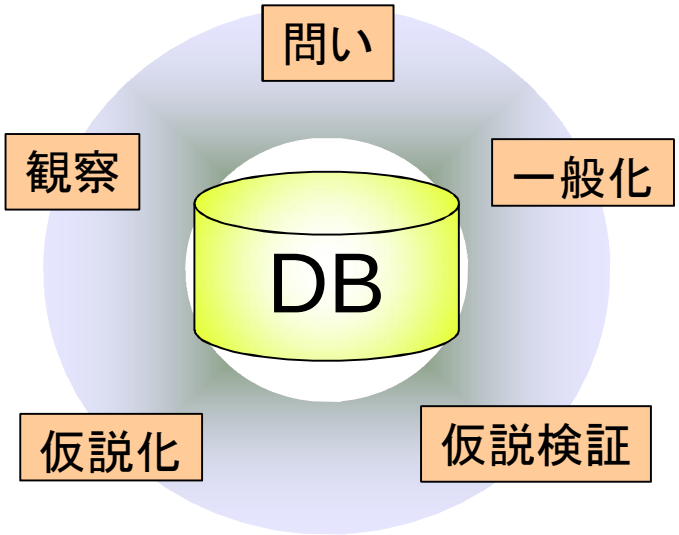
欧州系：EBI, European Bioinformatics Institute

――双璧

日本：各管轄省庁ごとにDBの統合化進行中

DB(生物情報)の生命科学における意義

- 研究活動の様々なステップにおいて活用される余地あり。
  - 膨大な量のデータ[DBに蓄積、仮説構築、実験による検証、DBに蓄積]
  - 帰納的(データ->法則)vs.演繹的(法則->検証)(仮説driven vs. Data driven)
    - 情報リテラシーが研究活動の様々な局面で要求される



OMIM：古くからあるhuman curated DB

遺伝子名

疾患名

omim.org/statistics/entry

Home | About | Statistics ▾ | Downloads/API ▾ | Help ▾ | External Links | Terms of Use ▾ | Contact Us

言語を選択

Search OMIM Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map
Search History: View, Clear

OMIM Entry Statistics:

Number of Entries in OMIM (Updated 3 October 2012) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	13,299	649	48	35	14,031
+ Gene and phenotype, combined	141	4	0	2	147
# Phenotype description, molecular basis known	3,290	264	4	28	3,586
% Phenotype description or locus, molecular basis unknown	1,628	134	5	0	1,767
Other, mainly phenotypes with suspected mendelian basis	1,773	125	2	0	1,900
Totals	20,131	1,176	59	65	21,431

核酸DBと遺伝子アノテーション

- 国際塩基配列データベース(INSDC)
- データ構造
- データのカテゴリ
- INSDC核酸DBでのID命名法
- 遺伝子アノテーション
- 遺伝子推定法とID命名法

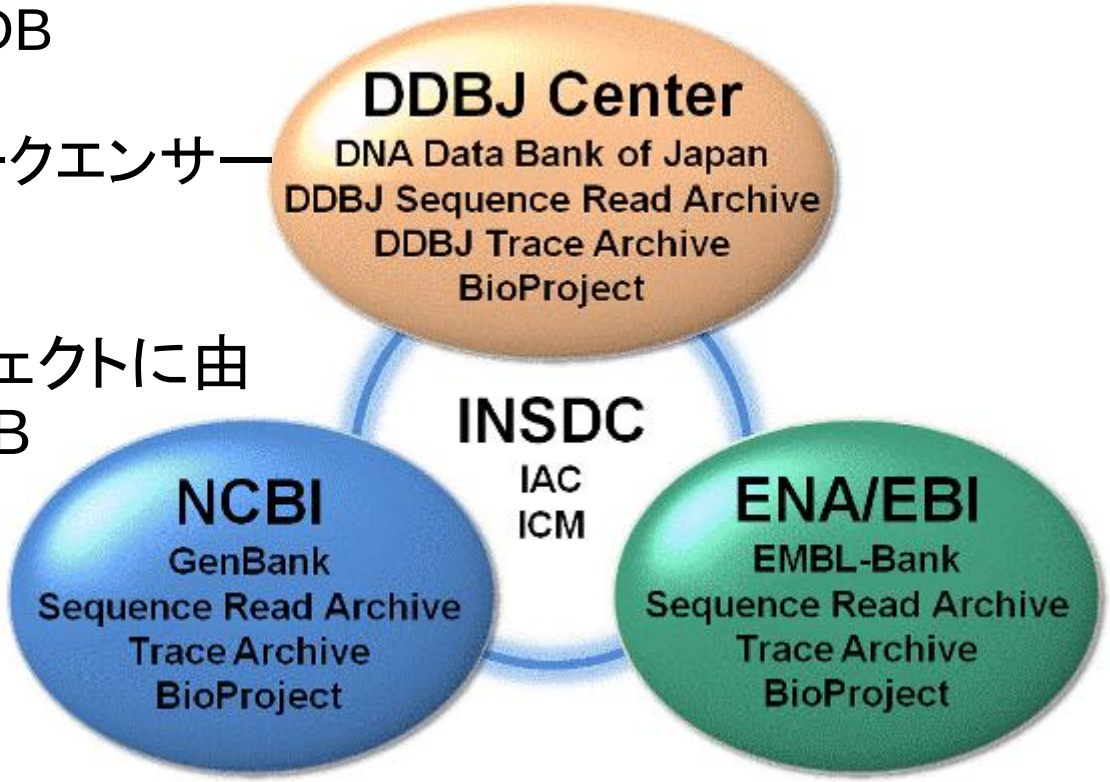
•国際塩基配列データベース(INSDC)

- ３センターの違い:
  - データアクセス(共通)とデータ登録方法(別々)
- 図:それぞれの機関で4行になっている意味

•1行目:従来からの核酸DB
•2行目:SRA (Sequence Read Archive, 次世代シークエンサーのreadの保存用DB
  - DDBJ SRA(DRA)

•3行目:Trace Archive, 従来型シークエンサーのreadの保存用DB
  - DDBJ Trace Archive (DTA)

•4行目:研究プロジェクトとプロジェクトに由来するデータをまとめるためのDB



データ構造
核酸DB
フラット・ファイル

フィールド

フラットファイルの
形式はEMBLだけ
少し異なる。

識別子
identifier

内容

識別子はDB内で統一
-->キーとして重要
核酸DBの主キーは
アクセッション

```
LOCUS       AK307560                1114 bp    mRNA    linear   HTC 12-JAN-2008
DEFINITION  Homo sapiens cDNA, FLJ97508.
ACCESSION   AK307560
VERSION     AK307560.1  GI:164692527
KEYWORDS    HTC; HTC_FLI; oligo capping.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Wakamatsu,A., Yamamoto,J., Kimura,K., Ishii,S., Watanabe,K.,
            Sugiyama,A., Murakawa,K., Kaida,T., Tsuchiya,K., Fukuzumi,Y.,
            Kumagai,A., Oishi,Y., Yamamoto,S., Ono,Y., Komori,Y., Yamazaki,M.,
            Kisu,Y., Nishikawa,T., Sugano,S., Nomura,N. and Isogai,T.
  TITLE     NEDO human cDNA sequencing project
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1114)
  AUTHORS   Isogai,T. and Yamamoto,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-JAN-2008) Contact:Takao Isogai Reverse Proteomics
            Research Institute; 1-9-11 Kaji-cho, Chiyoda-ku, Tokyo 101-0044,
            Japan E-mail :flj-cdna@nifty.com
COMMENT     Human cDNA sequencing project focused on splicing variants of mRNA
            in NEDO functional analysis of protein and research application
            project supported by Ministry of Economy, Trade and Industry,
            Japan; cDNA selection for complete cds sequencing: Reverse
            Proteomics Research Institute (REPRORI), Hitachi, Ltd., Japan
            (Hitachi) and Japan Biological Informatics Consortium, Japan
            (JBIC); cDNA complete cds sequencing: JBIC; cDNA library
            construction: Helix Research Institute supported by Japan Key
            Technology Center, Japan (HRI); cDNA 5'- & 3'-end sequencing:
            Research Association for Biotechnology, Japan, Biotechnology
            Center, National Institute of Technology and Evaluation, Japan and
            HRI; cDNA mapping to human genome: Central Research Laboratory,
            Hitachi; evaluation and annotation: REPRORI.
FEATURES             Location/Qualifiers
     source          1..1114
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /clone="NETRP2000337"
                     /cell_type="neutrophils"
                     /clone_lib="NETRP2"
                     /note="cloning vector: pME18SFL3;
                     primary culture, neutrophils"
ORIGIN
        1 agtgtcgacg gcagcggcgg cggcgggtgg gaaatggcgg agtatctggc ctccatcttc
```

primary配列DBとannotated DB

GenBank	RefSeq
Not curated	Curated
Author submits	NCBI creates from existing data
Only author can revise	NCBI revises as new data emerge
Multiple records for same loci common	Single records for each molecule of major organisms
Records can contradict each other	
No limit to species included	Limited to model organisms
Data exchanged among INSDC members	Exclusive NCBI database
Akin to **primary literature**	Akin to **review articles**
Proteins identified and linked	Proteins and transcripts identified and linked
Access via NCBI Nucleotide databases	Access via Nucleotide & Protein databases

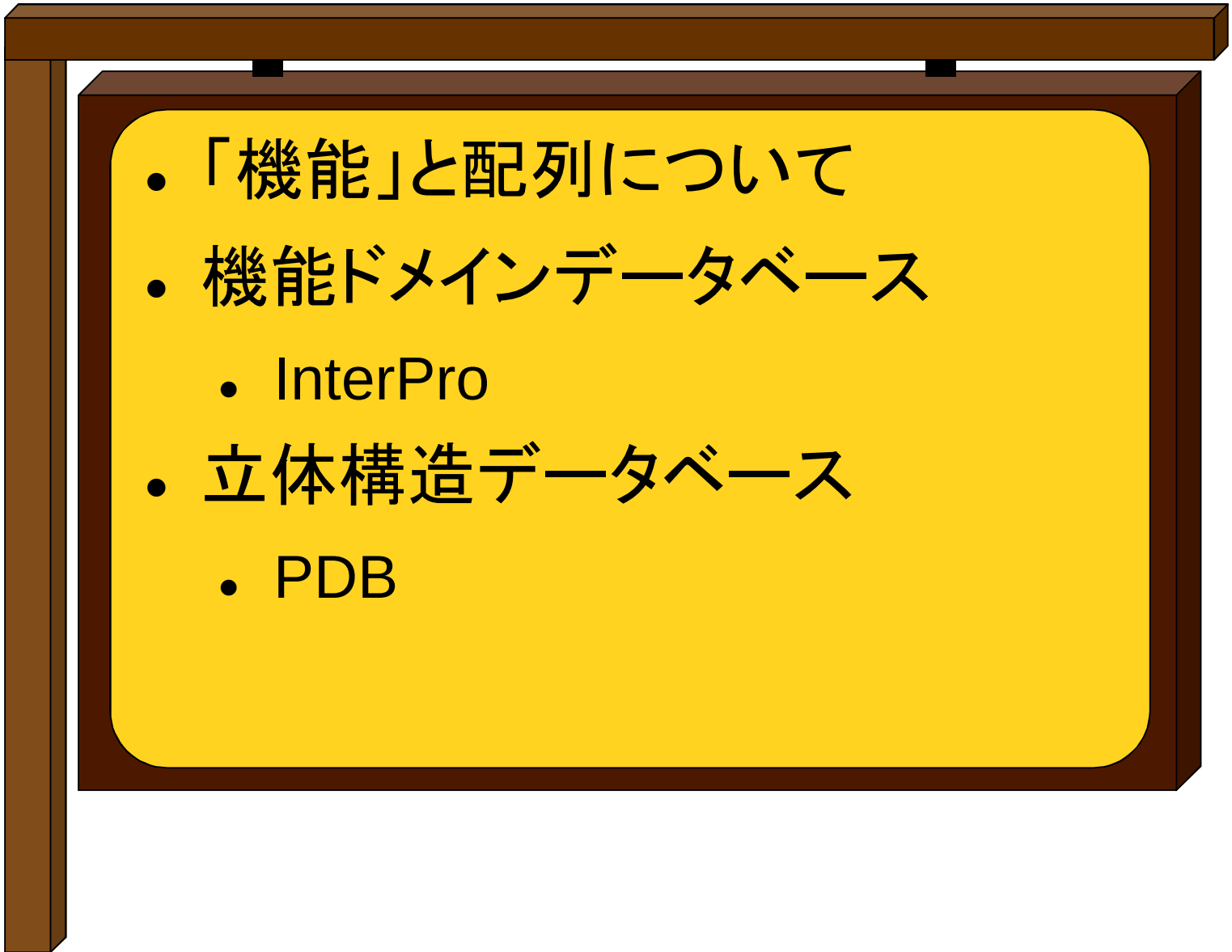
RefSeq、H-InvDBにおけるIDの付け方

- RefSeqのID
  - アルファベット2文字＋アンダーバー＋数字　例）NM_080593.2
  - アルファベット1文字目 :
    - A: Alternate assembly or annotation
    - N: kNown (A以外のゲノム配列とキュレータによりレヴューされたRNA&protein)
    - X: predictiction(prediXon)
    - Z: NZ_accession(WGS)上についたproteinのアノテーション
  - アルファベット2文字目 :
    - M: mRNA
    - R: noncoding RNA
    - P: protein
    - Z: genomic, whole genome shotgun (WGS) sequence data
    - W: genomic, Intermediate assemblies of BAC or WGS sequence data
    - T: genomic, Intermediate assemblies of BAC and/or WGS sequence data
    - S: genomic, unplaced scaffolds, etc
    - G: genomic, incomplete
    - C: genomic, complete

- H-InvDBのID
  - HIT (H-Invitational transcript):HIT + 9桁の数字 + version番号　例）HIT000000001.1
  - HIX (H-Invitational cluster): HIX + 7桁の数字 + version番号　例）HIX0000001.1
  - HIP (H-Invitational protein): HIP + 9桁の数字 + version番号　例）HIP000000001.1
  - HIF (H-Invitational gene family/group):　HIF + 7桁の数字　例）HIF0000001

http://hinv.jp/hinv/help/help_index_jp.html　http://www.ncbi.nlm.nih.gov/RefSeq/key.html#accessions

タンパク質データベースと機能推定



- 「機能」と配列について
- 機能ドメインデータベース
  - InterPro
- 立体構造データベース
  - PDB

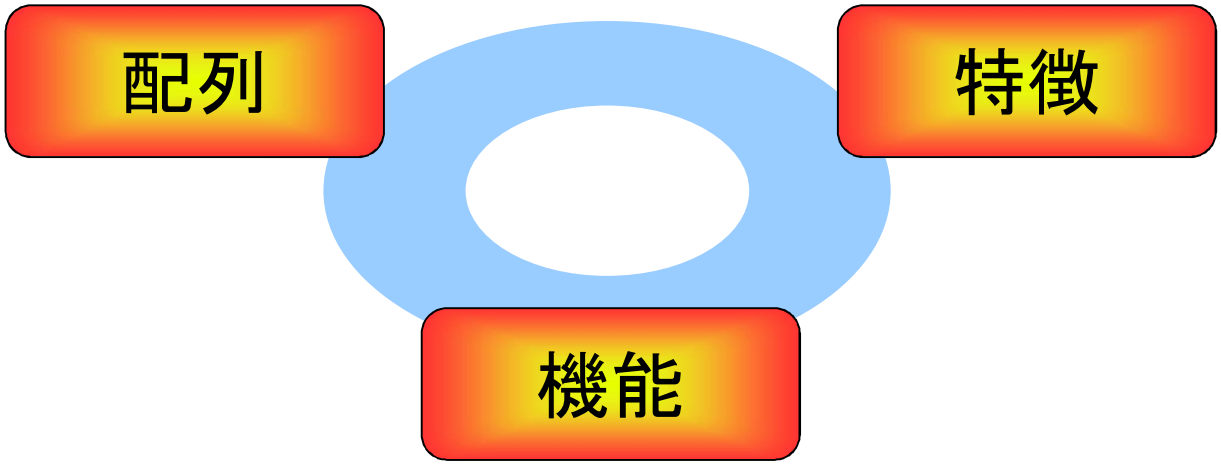
「機能」という語の多義性

Biochemical Function
生化学的機能
リガンド結合能、酵素活性

Biological Function
生物学的機能
代謝、転写、生殖、記憶

作業のscopeは何かが重要

機能と配列と特徴



- 配列に隠された機能の単位（ドメイン）。
- 機能ドメインは進化的に保存されやすい。
- 保存領域(Sequence signature)の長さ・規模は様々
  - 翻訳後修飾部位、モチーフ、ドメイン
- 配列上の特徴が機能ドメインゆえであることが多い。
- 決まった立体構造を通じて機能する

InterPro

- タンパク質(おもに機能ドメイン)DBの統合**DB**である。
  - ファミリー分類
  - 機能ドメイン
  - リピート
  - 翻訳後修飾など機能サイト
  - 機能推定に必要な情報
- 23792エントリーのデータ。

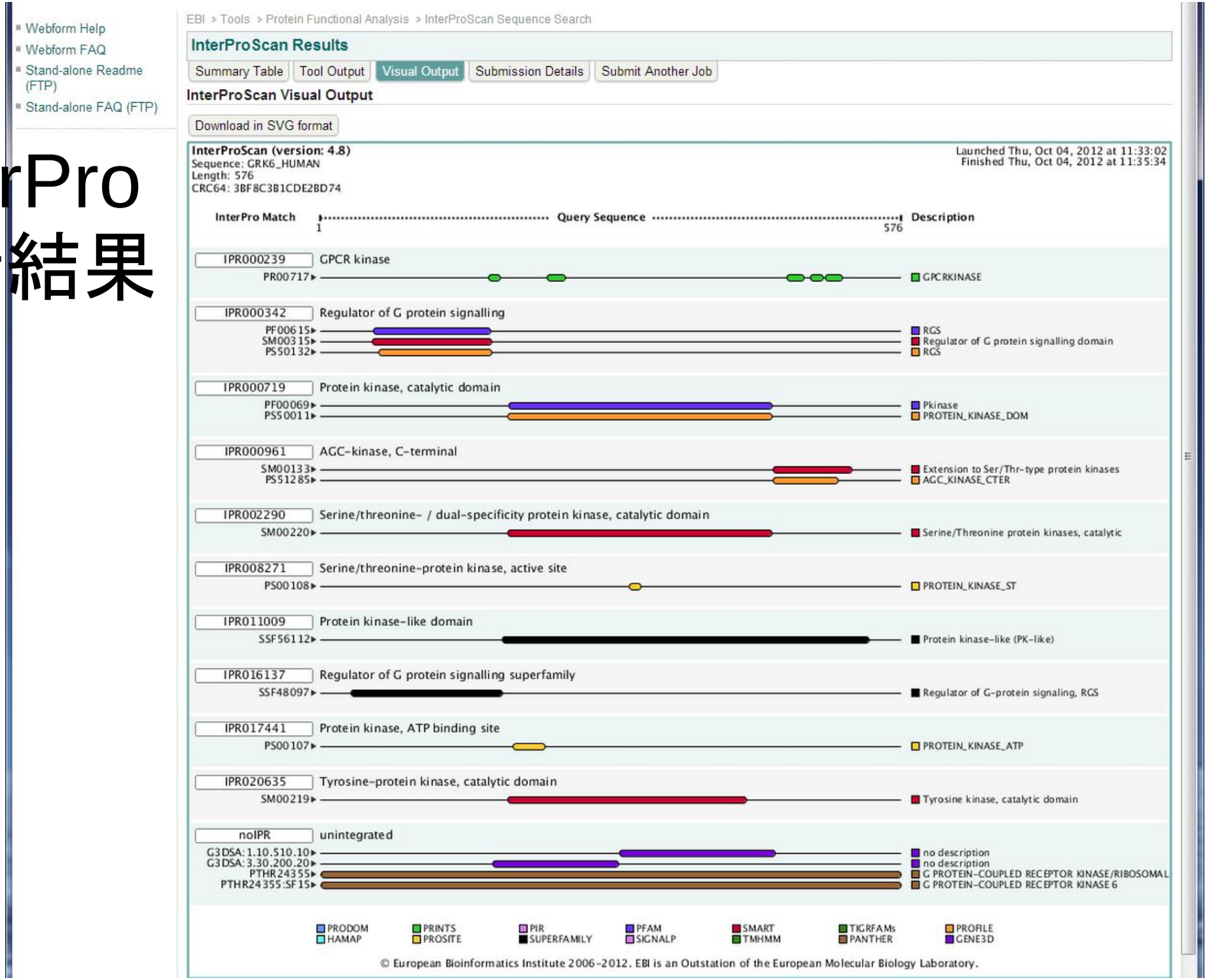
(InterPro release 39.0, Sept. 2012)

統合されたデータベース

InterProを用いた配列解析ツール InterProScan

- InterProに登録された各種機能ドメインを検索する。
- 機能未知配列中の各種機能関連sequence signatureを抽出して、特徴を発見する。
- ツールをダウンロードしてローカルで走らせることも可能。
- http://www.ebi.ac.uk/Tools/pfa/iprscan/

InterPro 実行結果



PDBとは

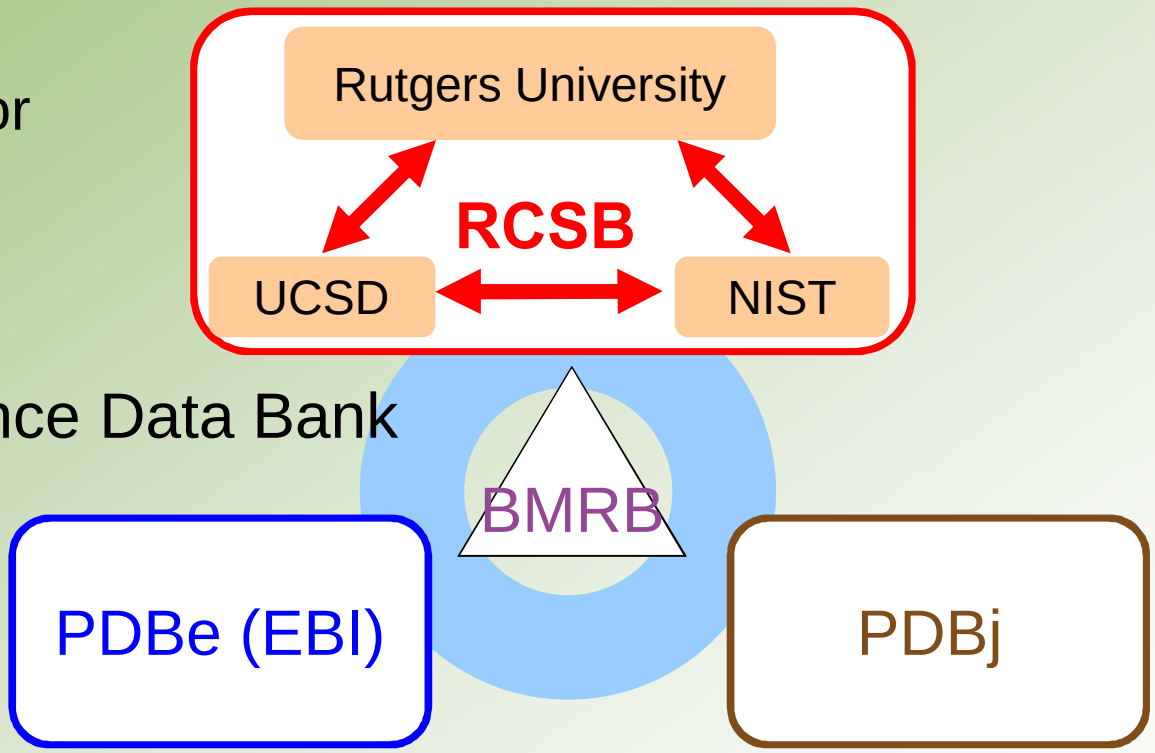
- タンパク質および核酸の三次元構造のデータベース
- X線結晶解析法やNMR(核磁気共鳴）法実験によって得られた三次元データを登録者本人とアノテータとでそれぞれ審査・検証を経て公開される。
- wwPDBにおける共同関係は核酸DBにおける国際塩基配列データベース(INSDC)に似ている。
  - データ・アーカイブは唯一で共通である。
  - 各機関はそれぞれにデータの登録受付とおよびブラウザ、ツール、webサービスを開発し公開する。

wwPDBの組織

Research Collaboratory for Structural Bioinformatics (RCSB), USA

Biological Magnetic Resonance Data Bank (BMRB), USA

Protein Databank in Europe (PDBe), Europe



日本蛋白質構造データバンク (PDBj), Japan

名称、用語、概念の共通化に関するデータベース

Gene Ontologyとは

- ñOntologyò＝”onò存在＋”logyò
  - 元々、哲学のひとつ存在論：entity（実体）の存在のあり方や他の実体との関係性（階層性、分類・体系性）について問う学問。
  - 情報学において、「概念化の明示的な仕様」と定義。
    - 同表記異義語の問題を解決、
    - 文章：単語の集まり-->意味のある実体として、コンピュータ処理が可能となった。（発展；セマンティック・ウェッブ）
- Gene Ontology Project:生物種やDBさらに分野の枠を超え、遺伝子（産物）関連用語を標準化。
  - 異祖同機能の問題解決、
  - -->DB間でのリンクや統合が容易に。
  - おもなゲノム関連研究機関が参加。
- GOは網羅的な遺伝子解析結果集計に多用。

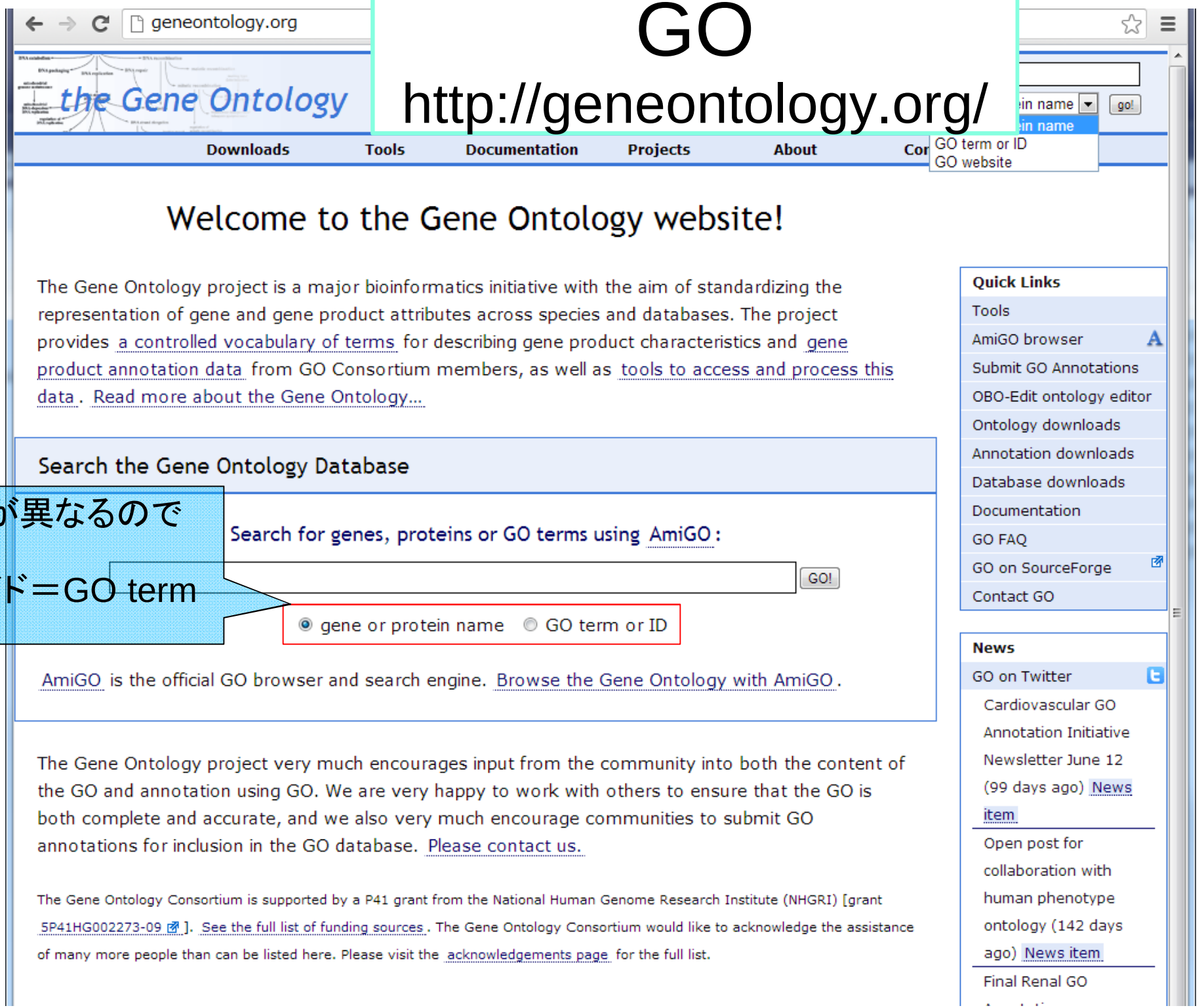
GO termの構造

- GO term全体に階層性を持つ。
- GOは3つのdomainをカバーする。
  - cellular component, 遺伝子産物細胞内外分布
  - molecular function, 遺伝子産物の機能
  - biological process, 生体内における役割

例:

Accession	GO:0015030
Ontology	cellular component

GO
http://geneontology.org/



表示結果が異なるので注意。
次のスライド＝GO term or ID

the Gene Ontology

AmiGOはGOのブラウザ

AmiGO

Search | Browse | BLAST | Homolog Annotations | Tools & Resources | Help

Search GO [Cajal body] ◉ terms ○ genes or proteins ☐ exact match [送信]

Term Search Results

4 results for **Cajal body** in terms fields **term accession, term name and synonyms**

Filter search results

Ontology

All
biological process
cellular component
molecular function

Set filters

Remove all filters

Results are sorted by **relevance**. To change the sort order, click on the column headers.

Select all | Clear all | Perform an action with this page's selected terms... | Go!

rel ↓	Accession , Term		Ontology
☑	GO:0015030 : **Cajal body** [show def]	196 gene products view in tree	cellular component
☐	GO:0072495 : host cell **Cajal body** [show def]	0 gene products view in tree	cellular component
☐	GO:0030576 : **Cajal body** organization [show def]	6 gene products view in tree	biological process
☐	GO:0061016 : snRNA import into **Cajal body** [show def]	0 gene products view in tree	biological process

Select all | Clear all | Perform an action with this page's selected terms... | Go!

AmiGO version: 1.8

Try AmiGO Labs

GO database release 2012-09-29

Cite this data • Terms of use • GO helpdesk

Cajal body

Term information ⬇ Term neighborhood ⬇ External references ⬇ 196 gene product associations ➡

Term Information

Accession GO:0015030

Ontology **Cellular Component**

Synonyms **exact**: coiled body
exact: Gems

Definition A class of nuclear body, first seen after silver staining by Ramon y Cajal in 1903, enriched in small nuclear ribonucleoproteins, and certain general RNA polymerase II transcription factors; ultrastructurally, they appear as a tangle of coiled, electron-dense threads roughly 0.5 micrometers in diameter; involved in aspects of snRNP biogenesis; the protein coilin serves as a marker for Cajal bodies. Some argue that Cajal bodies are the sites for preassembly of transcriptosomes, unitary particles involved in transcription and processing of RNA.
Source: PMID:10944589, PMID:11031238, PMID:7559785, http://genetics.cwru.edu/matera3.html

Comment None

Subset None

Community Add usage comments for this term on the GONUTS wiki.

Back to top

Term Neighborhood for Cajal body (GO:0015030)

Filter lineage gene product counts

Data source: No filter / ASAP / AspGD / CGD

Species: H. neptunium ATCC 15444 / H. sapiens / M. capsulatus str. Bath / M. grisea

Ancestors and Children | Inferred Tree View | Graph View | Other Views | Downloads | Mappings

GO:0044464 cell part [26402 gene products]

AmiGOはGOのブラウザ：例（上部）

AmiGOはGOのブラウザ：例（下部）

Back to top

Term Neighborhood for Cajal body (GO:0015030)

Filter lineage gene product counts

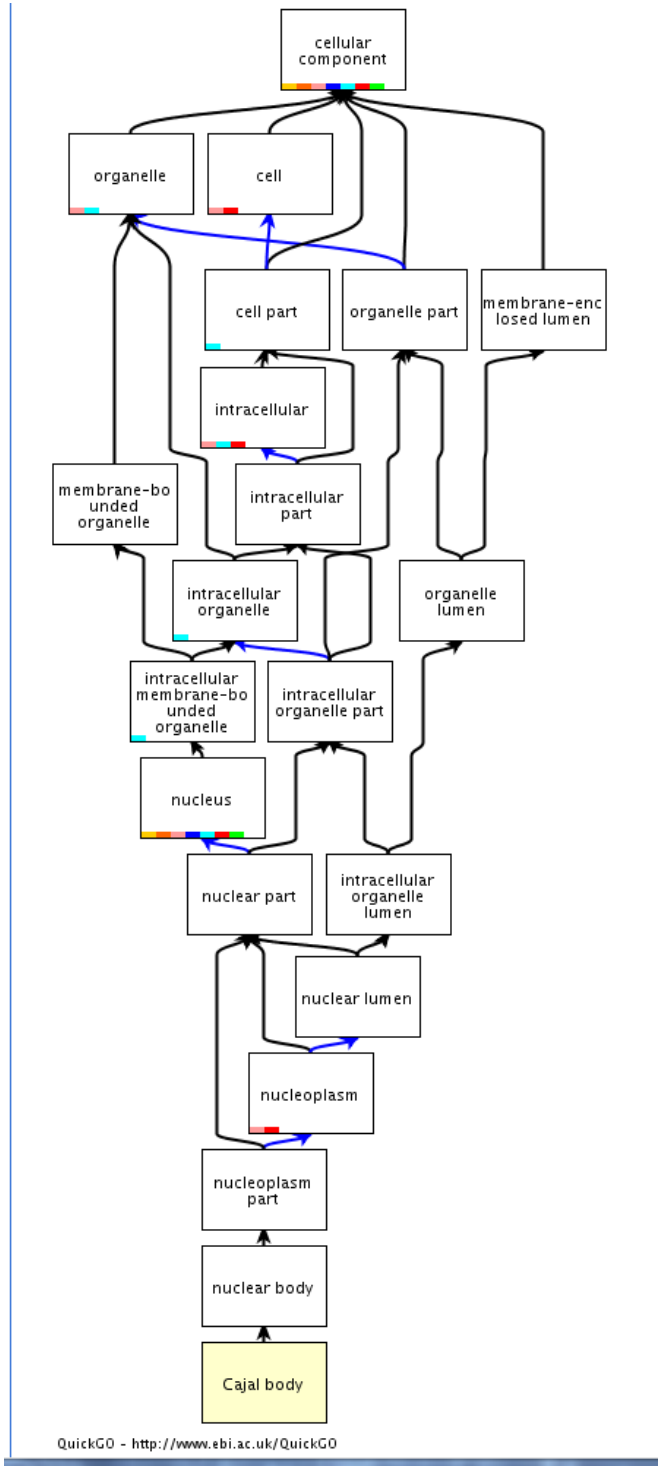
Data source: No filter / ASAP / AspGD / CGD

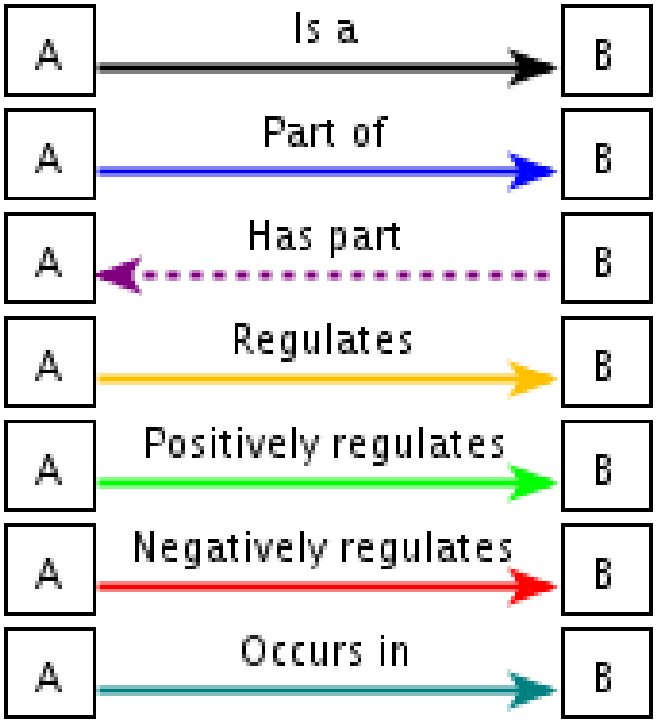
Species: H. neptunium ATCC 15444 / H. sapiens / M. capsulatus str. Bath / M. grisea

Ancestors and Children | **Inferred Tree View** | Graph View | Other Views | Downloads | Mappings

クリック

- GO:0044464 cell part [26402 gene products]
- GO:0005575 cellular_component [32185 gene products]
  - GO:0005623 cell [26403 gene products]
  - GO:0044424 intracellular part [20522 gene products]
  - GO:0031974 membrane-enclosed lumen [3465 gene products]
  - GO:0044422 organelle part [8790 gene products]
    - GO:0005622 intracellular [22940 gene products]
    - GO:0044446 intracellular organelle part [8637 gene products]
    - GO:0043227 membrane-bounded organelle [15254 gene products]
    - GO:0043226 organelle [17517 gene products]
    - GO:0043233 organelle lumen [3408 gene products]
      - GO:0043231 intracellular membrane-bounded organelle [15228 gene products]
      - GO:0043229 intracellular organelle [17485 gene products]
      - GO:0070013 intracellular organelle lumen [3360 gene products]
      - GO:0044428 nuclear part [3311 gene products]
        - GO:0031981 nuclear lumen [2746 gene products]
        - GO:0005634 nucleus [9509 gene products]
          - GO:0005654 nucleoplasm [1803 gene products]
            - GO:0044451 nucleoplasm part [1102 gene products]
              - GO:0016604 nuclear body [302 gene products]
                - **GO:0015030 Cajal body [47 gene products]**
                  - GO:0072589 box H/ACA scaRNP complex [0 gene products]


cellular component
organelle
cell
cell part
organelle part
membrane-enclosed lumen
intracellular
membrane-bounded organelle
intracellular part
intracellular organelle
organelle lumen
intracellular membrane-bounded organelle
intracellular organelle part
nucleus
nuclear part
intracellular organelle lumen
nuclear lumen
nucleoplasm
nucleoplasm part
nuclear body
Cajal body
QuickGO - http://www.ebi.ac.uk/QuickGO


A
Is a
B
A
Part of
B
A
Has part
B
A
Regulates
B
A
Positively regulates
B
A
Negatively regulates
B
A
Occurs in
B

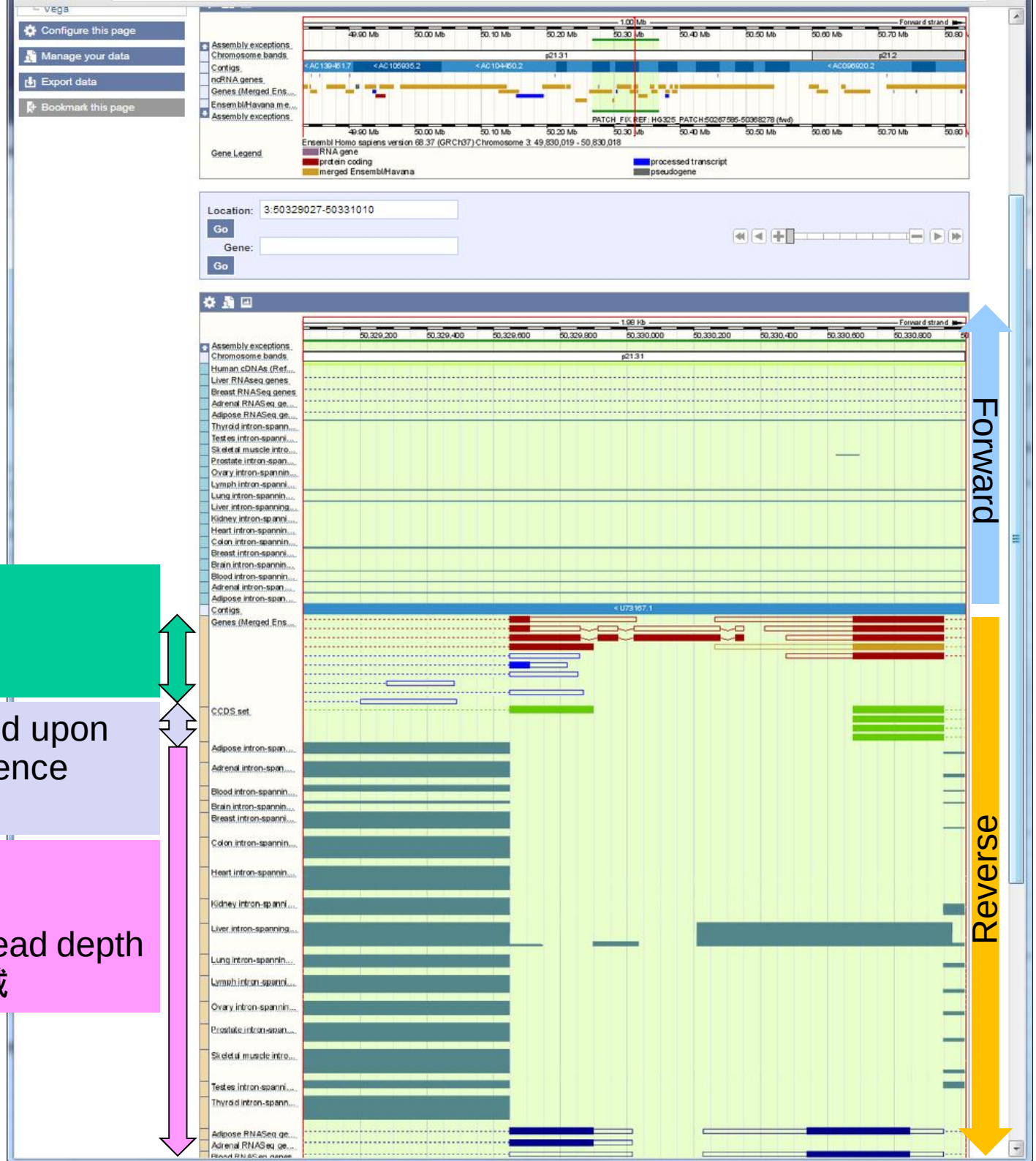
ゲノムDB・ゲノムbrowser

- NCBI
- EBI
  - Ensembl Genome browser
  - BioMart
- J.Craig Venter Institute (JCVI) HuRef
- UCSC
  - UCSC genome browser
- バージョンについて

ゲノム・ブラウザーは
遺伝子のゲノム上の位置を表示するだけでなく、
アノテーションを提供するための基盤である。

ユーザー自らのアノテーション / データを組み込んで表示できる。

Ensembl: genome browserの例

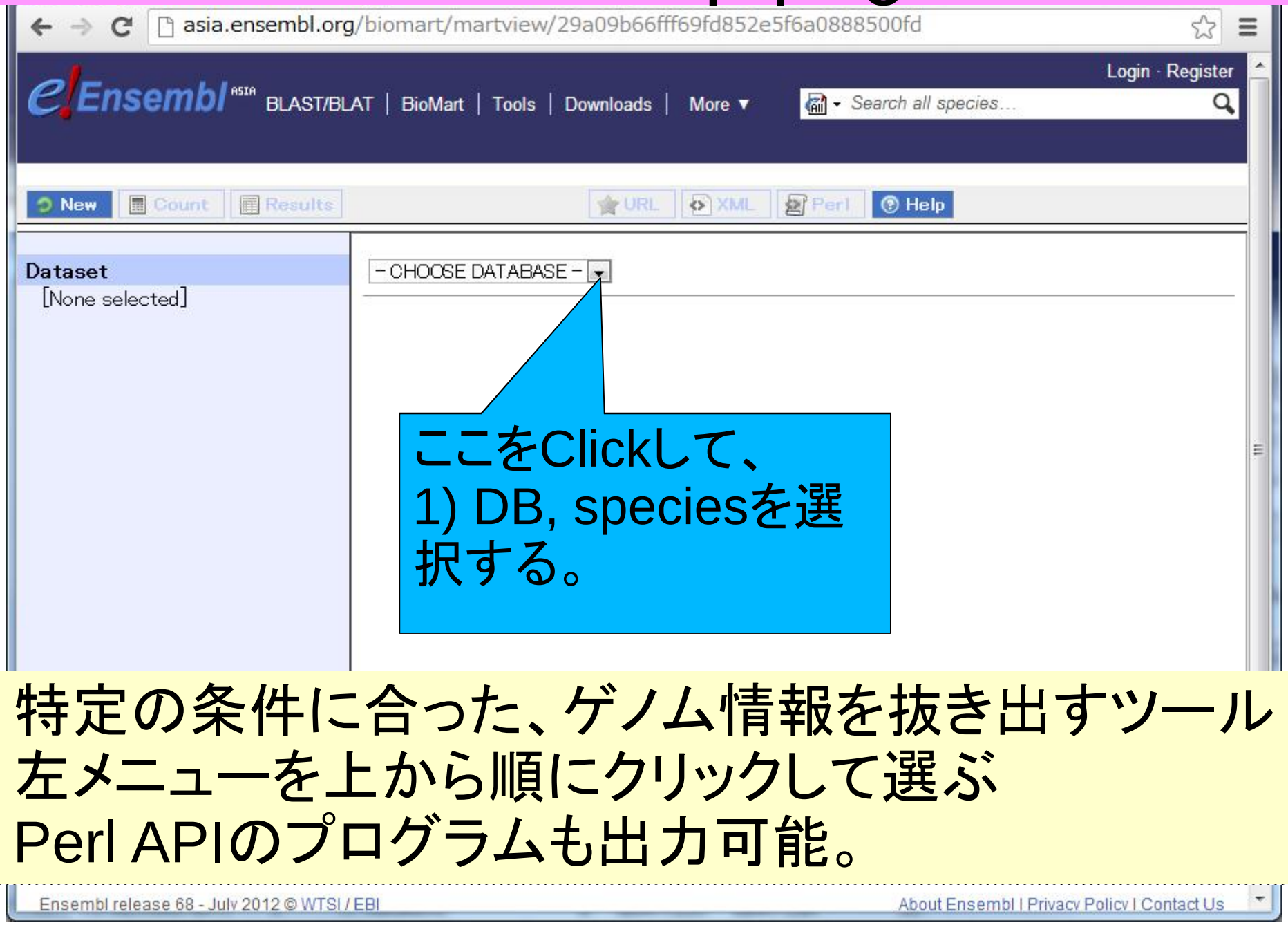


Gencode Gene Set (ENCODE)
(HAVANA: Manual)
(Ensembl: Automatic)

Protein coding sequences agreed upon by the Consensus Coding Sequence project, or CCDS.

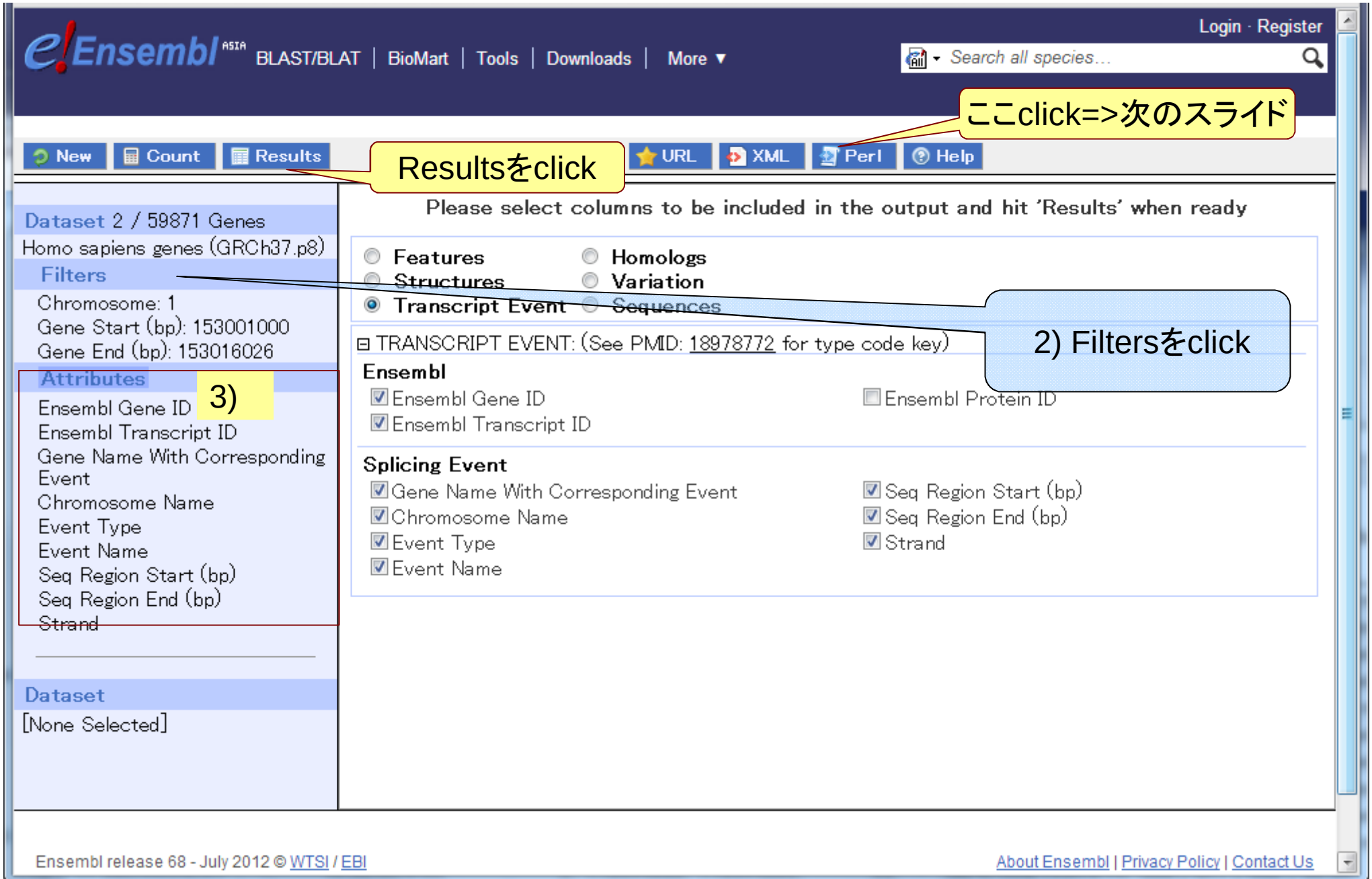
Human Bodymap 2.0 Data
32組織でのRNA-seq。
上の深緑：イントロンを支持するRead depth
下の四角：推定されたエクソン領域

BioMart top page



◆ 特定の条件に合った、ゲノム情報を抜き出すツール
◆ 左メニューを上から順にクリックして選ぶ
◆ Perl APIのプログラムも出力可能。

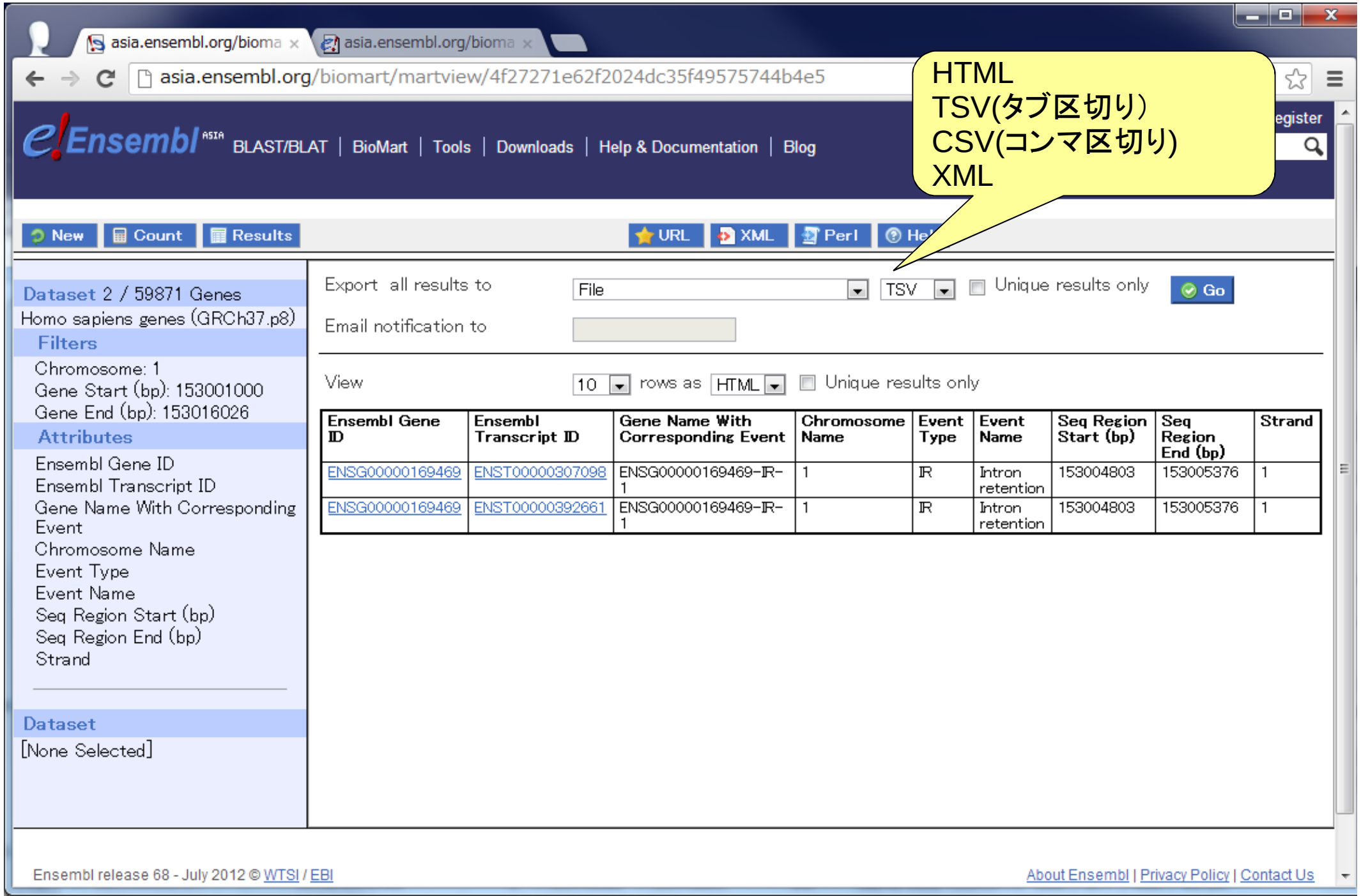
BioMart:Attributes;出力したい項目に絞り込む



asia.ensembl.org/biomart/martview/4f27271e62f2024dc35f49575744b4e5

```
# An example script demonstrating the use of BioMart API.
# This perl API representation is only available for configuration versions >=  0.5
use strict;
use BioMart::Initializer;
use BioMart::Query;
use BioMart::QueryRunner;

my $confFile = "PATH TO YOUR REGISTRY FILE UNDER biomart-perl/conf/. For Biomart Central Registry navigate to
                                                    http://www.biomart.org/biomart/martservice?type=registry";
#
# NB: change action to 'clean' if you wish to start a fresh configuration
# and to 'cached' if you want to skip configuration step on subsequent runs from the same registry
#

my $action='cached';
my $initializer = BioMart::Initializer->new('registryFile'=>$confFile, 'action'=>$action);
my $registry = $initializer->getRegistry;

my $query = BioMart::Query->new('registry'=>$registry,'virtualSchemaName'=>'default');


        $query->setDataset("hsapiens_gene_ensembl");
        $query->addFilter("chromosome_name", ["1"]);
        $query->addFilter("end", ["153016026"]);
        $query->addFilter("start", ["153001000"]);
        $query->addAttribute("ensembl_gene_id");
        $query->addAttribute("ensembl_transcript_id");
        $query->addAttribute("name_1078");
        $query->addAttribute("splicing_event__dm_name_1059");
        $query->addAttribute("splicing_event_type");
        $query->addAttribute("name_106");
        $query->addAttribute("seq_region_start_1078");
        $query->addAttribute("seq_region_end_1078");
        $query->addAttribute("seq_region_strand_1078");

$query->formatter("TSV");

my $query_runner = BioMart::QueryRunner->new();
############################## GET COUNT ####################################
# $query->count(1);
# $query_runner->execute($query);
# print $query_runner->getCount();
#############################################################################


############################## GET RESULTS ##################################
# to obtain unique rows only
# $query_runner->uniqueRowsOnly(1);

$query_runner->execute($query);
$query_runner->printHeader();
$query_runner->printResults();
$query_runner->printFooter();
#############################################################################
```

BioMart: Result表示例



www.genome.ucsc.edu/index.html?org=Human

UCSC Genome Browser genome.ucsu.edu

UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Session - FAQ - Help

- Genome Browser
- ENCODE
- Neandertal
- Blat
- Table Browser
- Gene Sorter
- In Silico PCR
- Genome Graphs
- Galaxy
- VisiGene
- Utilities
- Downloads
- Release Log
- Custom Tracks
- Microbial Genomes
- Mirrors
- Archives

表形式でBioMartのように、欲しい項目に絞り込んだ上で、表示、ダウンロード

遺伝子を任意の関係（発現、相同性、位置など）で並び変えて、表示、ダウンロード

遺伝子発現部位の顕微鏡像 マウスやアフリカツメガエルのプロジェクトより

About the UCSC Genome Bioinformatics Site

...and working draft assemblies for a ...d scrolls over chromosomes, showing ...ormation on groups of genes that can ...provides convenient access to the ...frog images to examine expression ...oup, a cross-departmental team within ...lar Science and Engineering (CBSE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

News

News Archives ▶

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

17 September 2012 - Retiring the Proteome Browser

We are announcing the imminent retirement of the Proteome Browser. We introduced the Proteome Browser in 2003 to provide access to protein-specific information independent of the genomic details presented in the Genome Browser. Since then we've incorporated more of this information in the UCSC Genes details page accessible from the Genome Browser, and in the columns of the Gene Sorter. Since very few people are still accessing the Proteome Browser, we've decided to retire it to focus our work on these other two tools. Access to the databases that supported the Proteome Browser (uniProt and proteome) will still be available through the Table Browser and our public MySQL server.

---

16 August 2012 - Announcing a Genome Browser for the Medium ground finch

We have released a browser for the Medium ground finch, *Geospiza fortis*, reknowned as one of naturalist Charles Darwin's Galapagos finches. This species, which has been the subject of many evolutionary studies, is one of a group of birds that evolved over a few million years from a single ancestral species into multiple species whose beak sizes and shapes are specialized for using different food resources. The phenotypic diversity of these birds contributed to Darwin's theory of evolution. The significance of this genome assembly is described in the August 16, 2012 press release issued by the UCSC Center for Biomolecular Science and

必要なDBを見つけるために

- NAR Database Summary Paper Category List
  - http://www.oxfordjournals.org/nar/database/c/
- HGNC Useful Links
  - http://www.genenames.org/useful.html
- Integbioデータベースカタログ
  - http://integbio.jp/dbcatalog/
- 統合DBプロジェクト
  - http://biosciencedbc.jp/dbsearch/
- 生命科学系 データベース カタログ
  - http://biosciencedbc.jp/dbcatalog/dbcatalog.cgi?pg=1

生命科学解析用ツール群

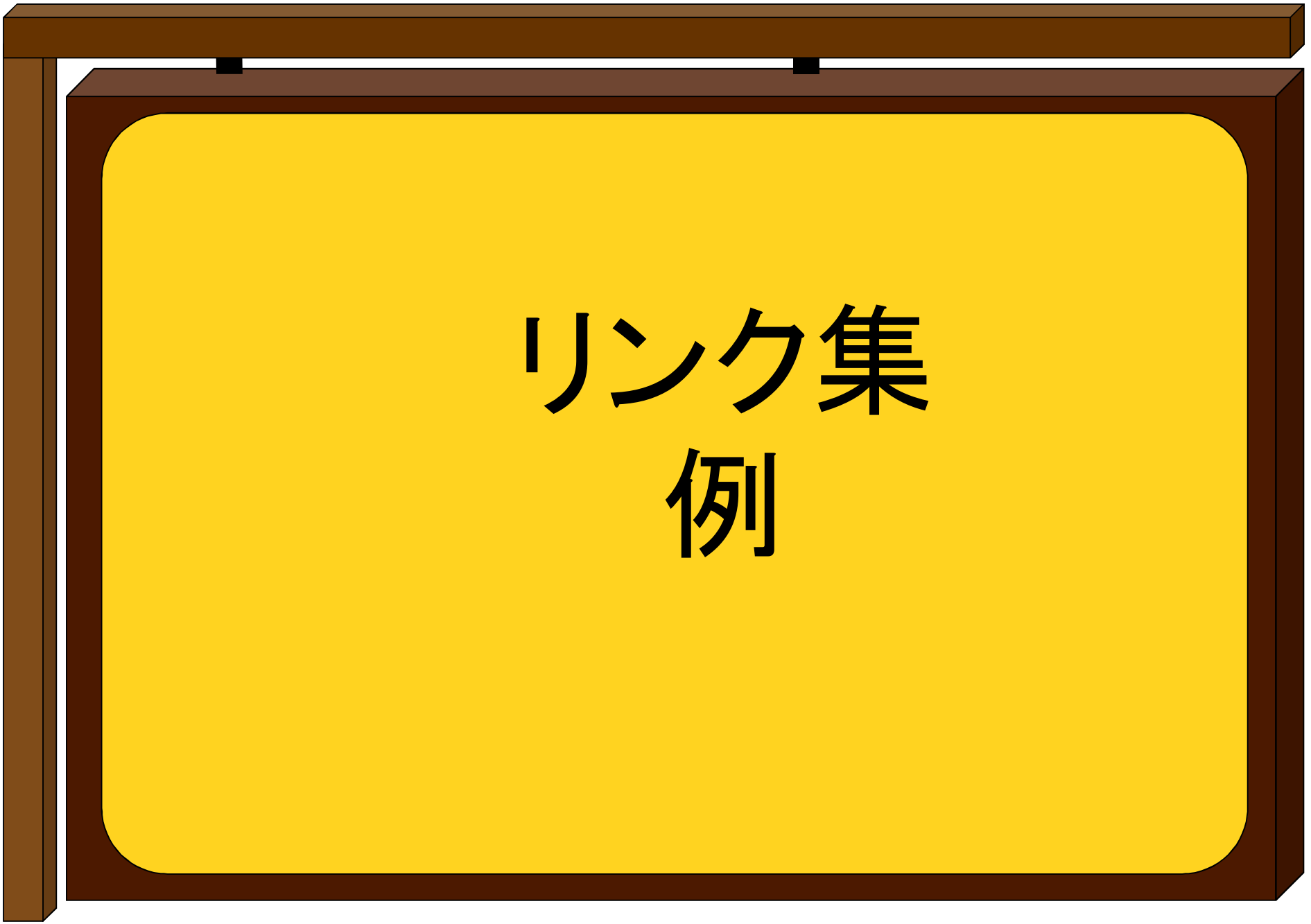
- ツール選定にあたって
  - 手順
    - 解析手順と各段階での目的の明確化
  - 基準
    - web or local, 人気, 更新年月日, 実績
- 選定ガイド
  - 論文から
    - 専門の近い論文　Method
    - NAR web server Issue (http://nar.oxfordjournals.org/content/40/W1.toc)
  - リンク集から

古いのは、使用環境や目的に合わないことも、新しいのは完成度が低いことも。

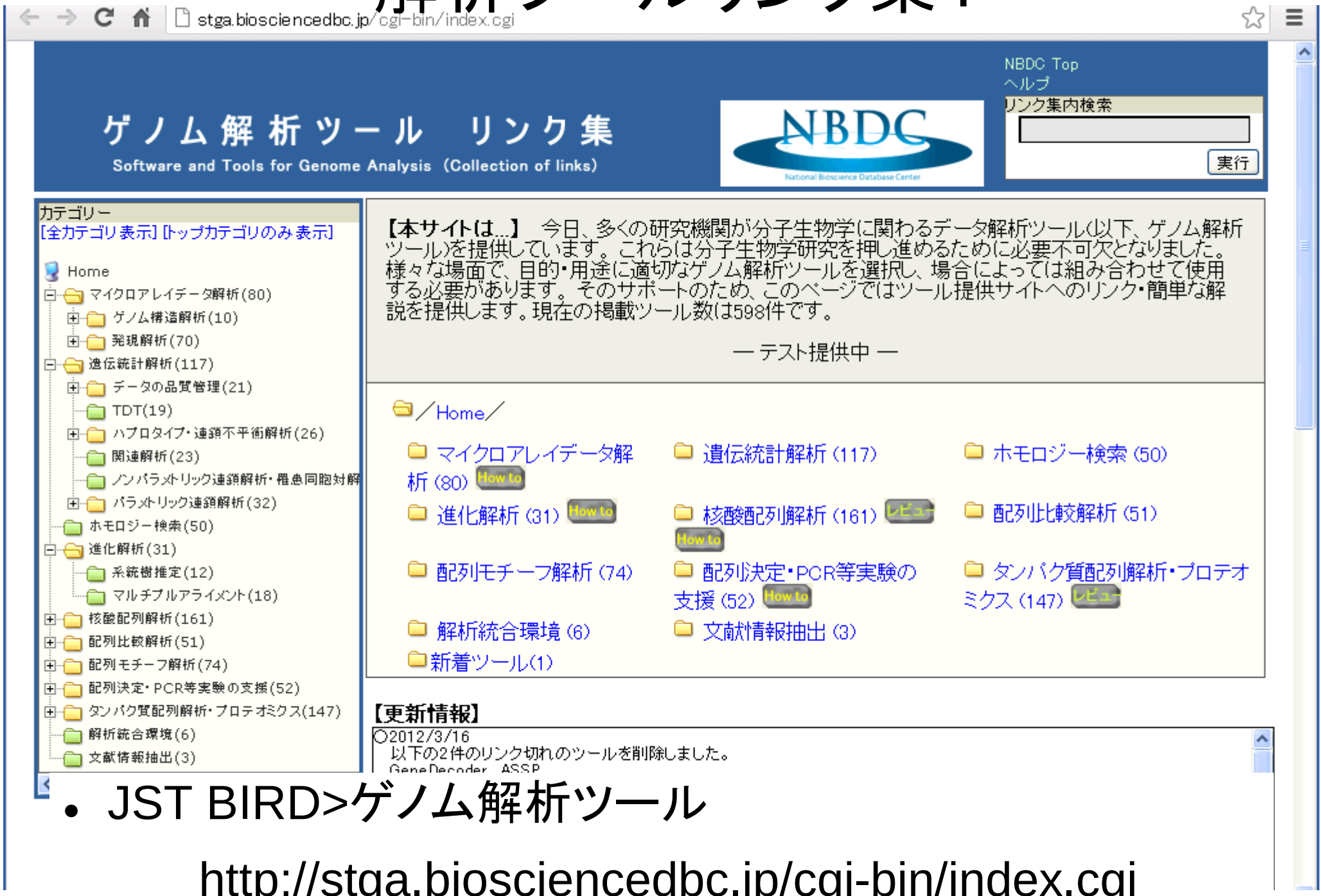
生命科学解析用ツール群

- ツール使用にあたって
  - 読んでおくべきもの
    - ReadMe等のドキュメント
    - 論文
  - 尋ねるための環境
    - Helpdesk, 作者に直接
    - Mailing List, Forum
    - インターネット検索：経験談, log
  - 尋ね方のコツ：以下を伝えたうえで
    - 目的、ドキュメント既読であること、到達点、不明点、
    - 環境（OS、マシンのスペック)

生命科学解析用ツール群



解析ツールリンク集1



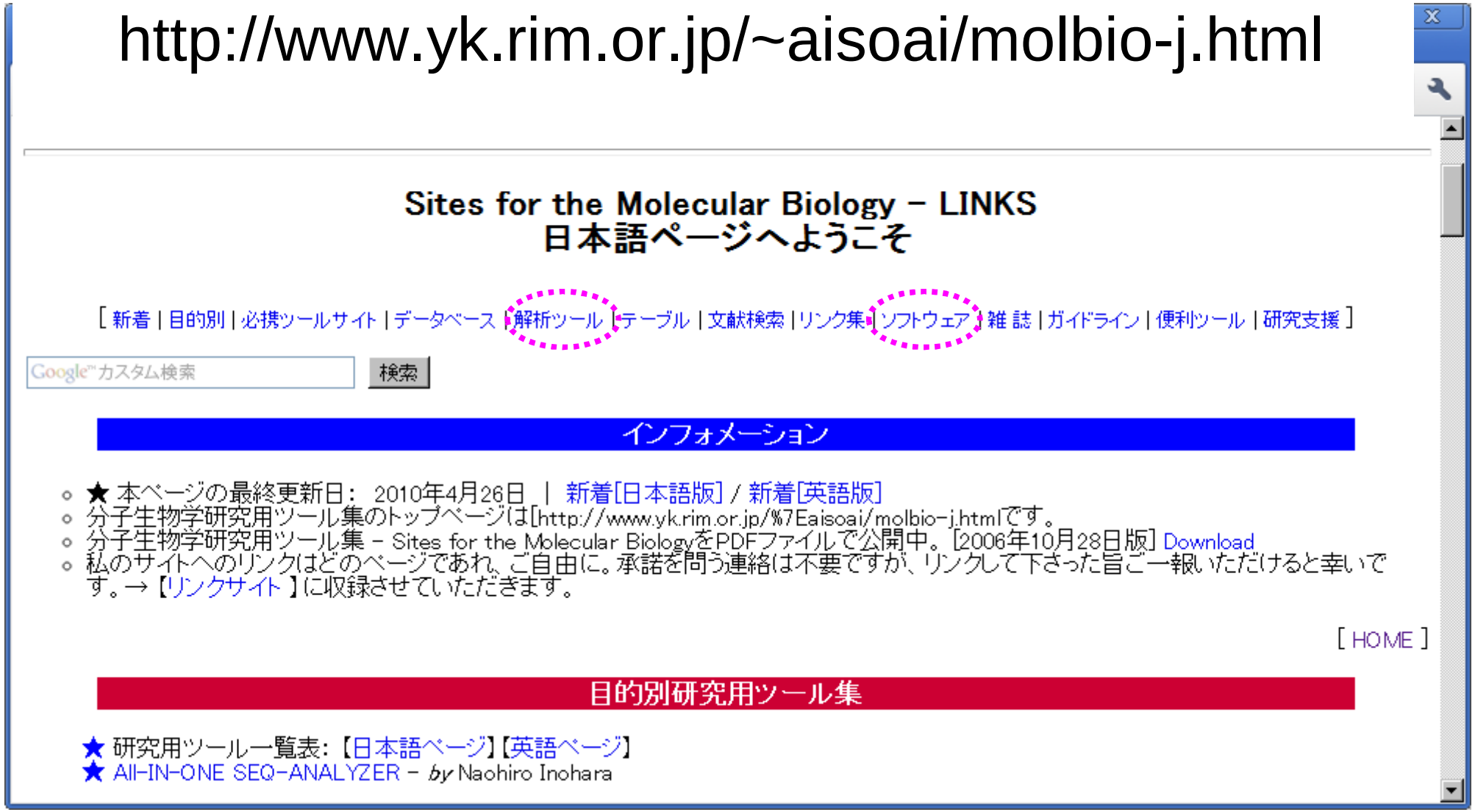
- JST BIRD>ゲノム解析ツール

http://stga.biosciencedbc.jp/cgi-bin/index.cgi

解析ツールリンク集2

- 分子生物学研究用ツール集（by Dr. Atsushi Isoai）

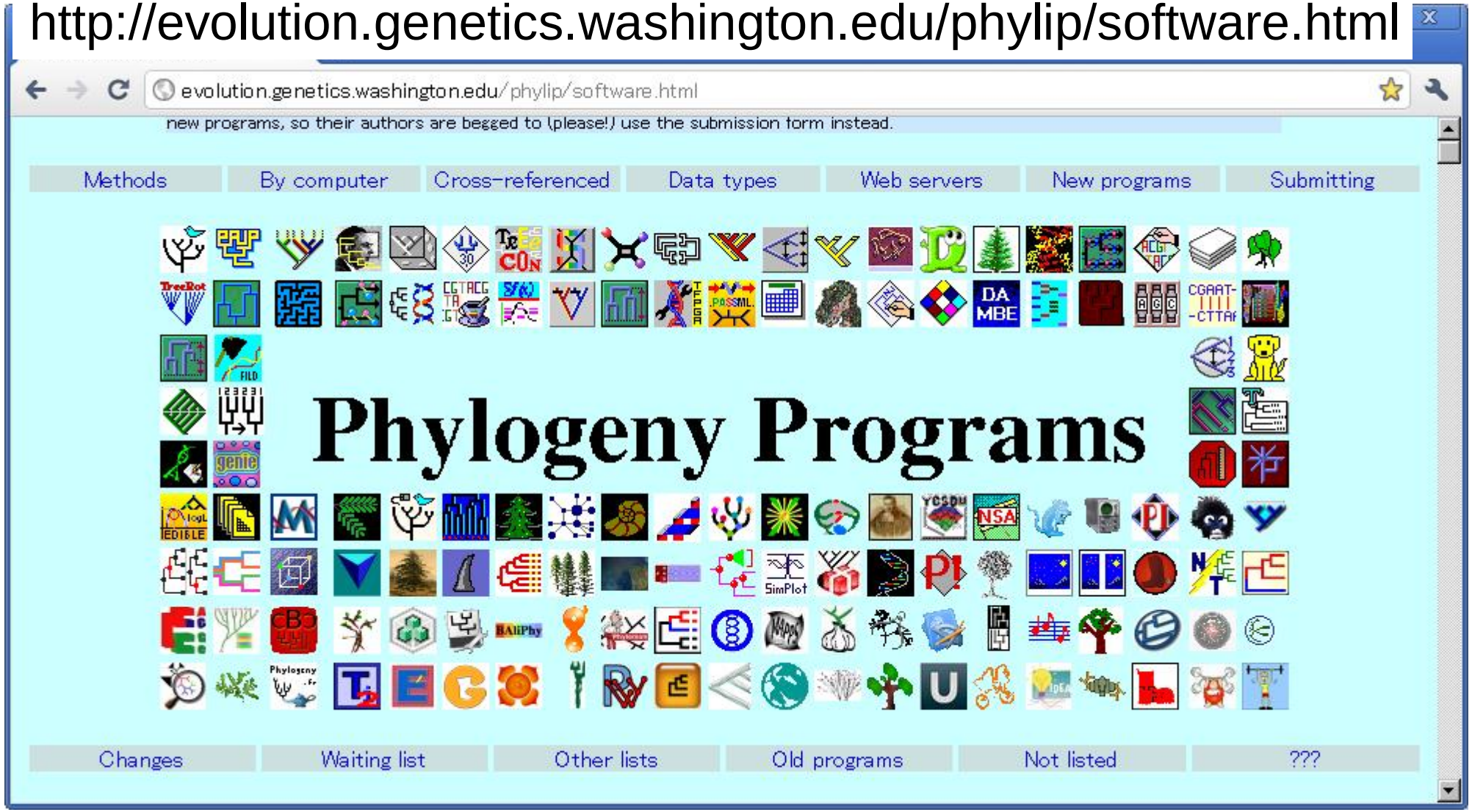
http://www.yk.rim.or.jp/~aisoai/molbio-j.html



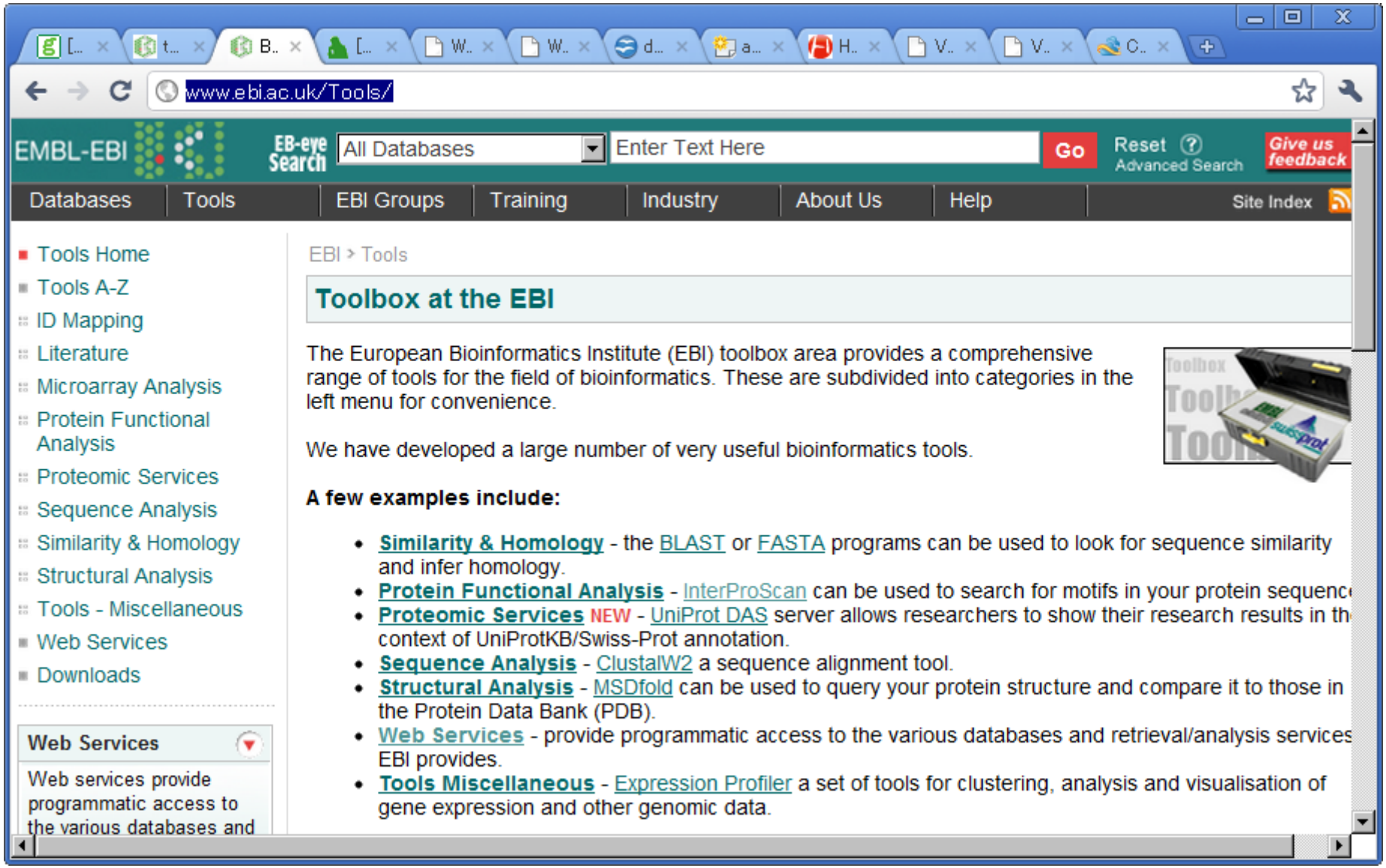
解析ツールリンク集3

- 分子進化学関係 (by Dr.Joe Felsenstein)

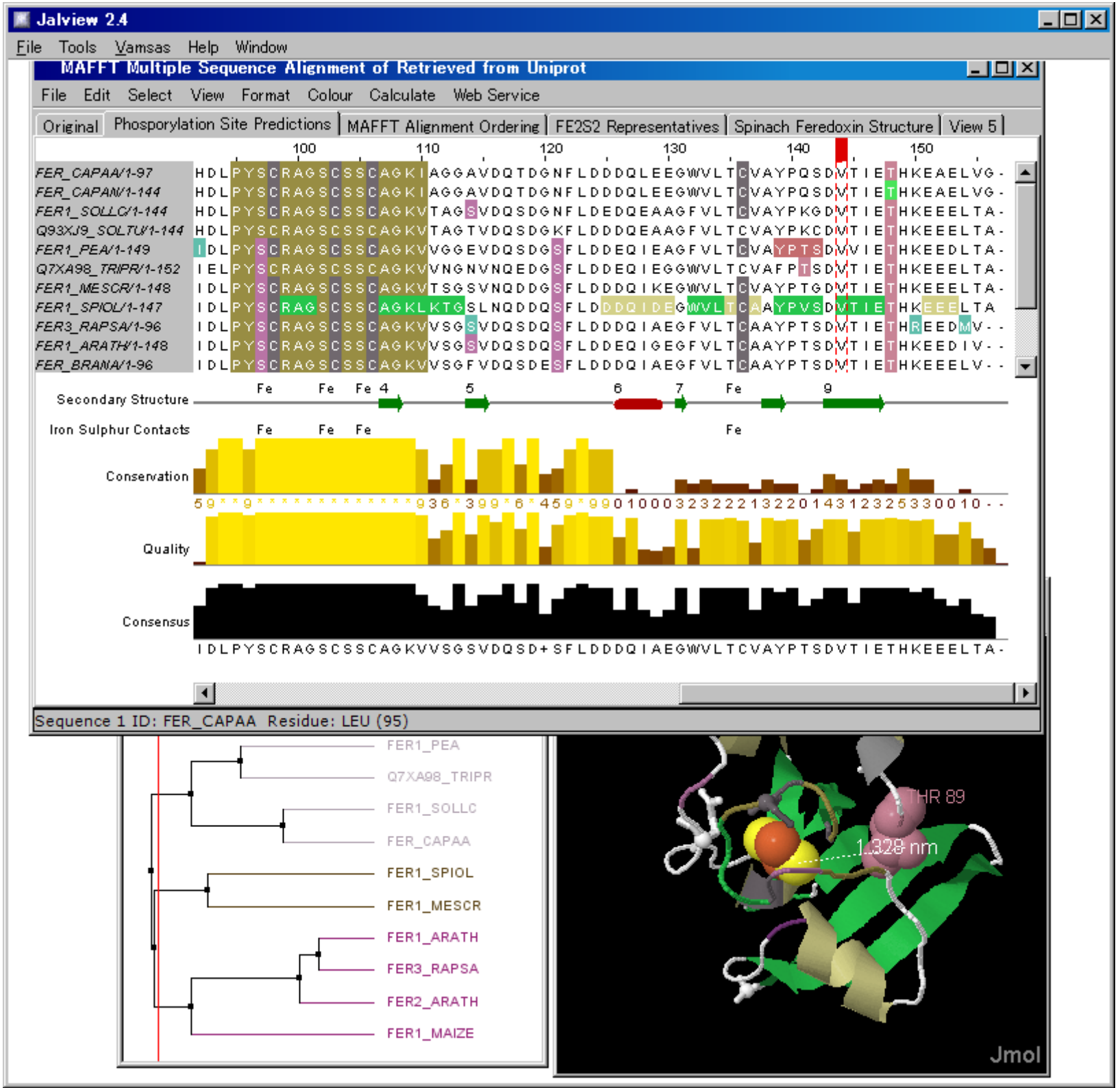
http://evolution.genetics.washington.edu/phylip/software.html



EBI toolbox
http://www.ebi.ac.uk/Tools/



配列editorの例


Jalview 2.4
File Tools Vamsas Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
MAFFT Alignment Ordering
FE2S2 Representatives
Spinach Feredoxin Structure
View 5
100 110 120 130 140 150
FER_CAPAA/1-97
FER_CAPAN/1-144
FER1_SOLLC/1-144
Q93XJ9_SOLTU/1-144
FER1_PEA/1-149
Q7XA98_TRIPR/1-152
FER1_MESCR/1-148
FER1_SPIOL/1-147
FER3_RAPSA/1-96
FER1_ARATH/1-148
FER_BRANA/1-96
Secondary Structure
Iron Sulphur Contacts
Conservation
Quality
Consensus
Sequence 1 ID: FER_CAPAA Residue: LEU (95)
FER1_PEA
Q7XA98_TRIPR
FER1_SOLLC
FER_CAPAA
FER1_SPIOL
FER1_MESCR
FER1_ARATH
FER3_RAPSA
FER2_ARATH
FER1_MAIZE
THR 89
1.328 nm
Jmol

／Home／進化解析／マルチプルアライメント／

カテゴリー
[全カテゴリ表示] [トップカテゴリのみ表示]

- Home
- マイクロアレイデータ解析
- 遺伝統計解析
- ホモロジー検索(1)
- 進化解析(18)
  - 系統樹推定
  - マルチプルアライメント(19)
- 核酸配列解析(1)
  - 繰り返し配列探索
  - ホモロジー検索(1)
  - エクソン・イントロン構造予測
  - プロモータ予測
  - UTR予測
  - 核酸高次構造推定
  - 制限酵素切断部位の検出
  - 転写因子結合サイトの抽出・検索
- 配列比較解析(19)
  - ドットプロット
  - ゲノムスケール配列アライメント(4)
  - マルチプルアライメント(19)
  - ペアワイズアライメント
- 配列モチーフ解析
  - モチーフ抽出
  - モチーフ検索
- 配列決定・PCR等実験の支援
- タンパク質配列解析・プロテオミクス(1)
- 解析統合環境
- 文献情報抽出

／Home／進化解析／マルチプルアライメント／

並べ替え：被引用件数が多い順番

ツール数:19

How to　キーワードによる絞り込み

系統樹を作成する How to

マルチプルアライメントに基づいて系統樹を作成する際の注意事項。

●ClustalW　カテゴリー

累進法によるマルチプルアライメントツール。NJ法により系統樹を作成し、その系統樹で距離が近い配列同士から累進法によってマルチプルアライメントに組み上げられて行く。入力された配列群の全ペアを対象としてペアワイズアライメントを行い距離行列を生成し、基にNJ法により系統樹を作成、系統樹の枝に沿ってペアワイズアライメント同士をアライメントすることにより最終的なマルチプルアライメントを生成する。

文献：CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.
引用数：31352(更新日:2010/6/8)　link to google scholar
提供サイト：EBI , NIG , IGBMC
ツール更新日：2009/4/16

●T-Coffee　カテゴリー

複数のアライメントプログラムにより求められたペアワイズアライメントの結果を用いて、それらが高い整合性をもつように重み付けを行う。そのスコア体系を用いて、累進法による多重配列アライメントを構築する。塩基やアミノ酸配列のグローバルアライメントやローカルアライメントを求めるプログラムをペアワイズアライメントに用いることができる。

文献：T-Coffee: A novel method for fast and accurate multiple sequence alignment.
引用数：2648(更新日:2010/6/8)　link to google scholar
提供サイト：CNRS , SIB , EBI
ツール更新日：2010/4/24

●MultAlin　カテゴリー

累進法によるマルチプルアライメントツール。階層的クラスタリングと累進法によるマルチプルアライメントの組み上げが交互に行われ、最終的なマルチプルアライメントが生成される。

文献：Multiple sequence alignment with hierarchical clustering.
引用数：2446(更新日:2010/6/8)　link to google scholar
提供サイト：INRA
ツール更新日：2000/3/28

●MUSCLE　カテゴリー

アミノ酸配列のマルチプルアライメントを行うツール。まず、k-tupleに基づく距離の計算を行い、アライメントのペアを求める。次に、そのペアに対して距離の再計算を行い、平均距離法(UPGMA)により、系統樹を作成する。最後に最適のSPスコアを与えるように部分系統樹の再構成アライメントを繰り返す。全てのペアワイズアライメントを求めないため、処理が高速である。

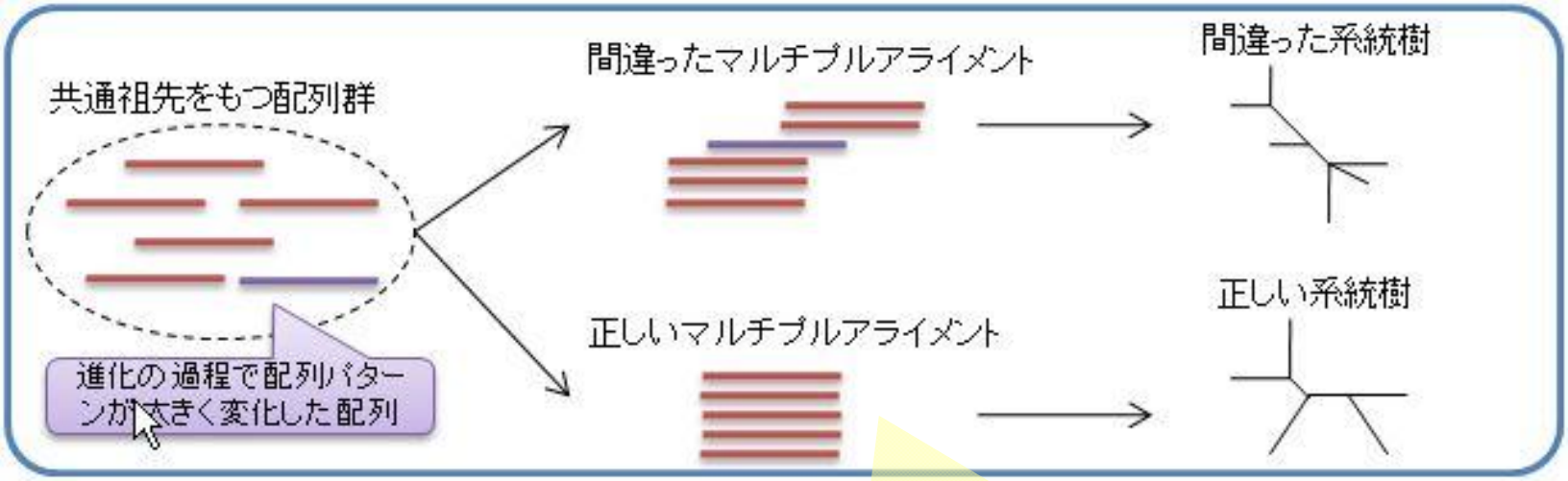
文献：MUSCLE: multiple sequence alignment with high accuracy and high throughput.
引用数：2371(更新日:2010/6/8)　link to google scholar
提供サイト：UCB , EBI
ツール更新日：2010/5/1

●MaxHom　カテゴリー

アミノ酸配列を入力データとしてデータベースを配列類似性検索して、自動的にプロファイルを生成するツール。WEBサーバではSWISS-PROTをデータベースとしている。データベース検索はBLASTPによって行われ、ヒットした配列は動的計画法によりアライメントされプロファイルに変換される。このプロファイルは別のヒットとのアライメントに使用される。

マルチプルアライメントに基づいて系統樹を作成する際の注意事項1

- 『系統樹を作成する』
  - 【マルチプルアライメントの精度と系統樹】



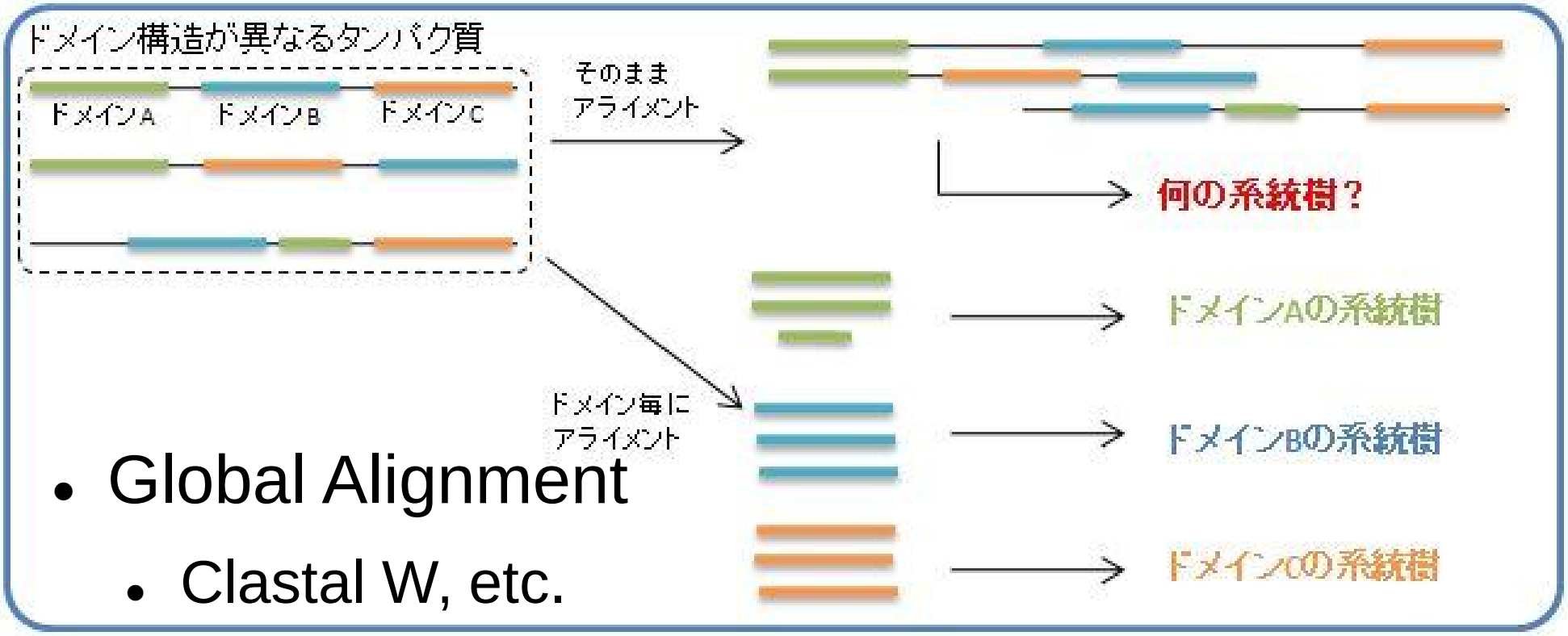
  - 【マルチドメインタンパク質のマルチプルアライメント】

配列エディターでギャップ位置を手直し、保存。
さらにマルチプルアライメントソフトにかける。

http://www-btls.jst.go.jp/Links/howto.cgi?id=4000_01

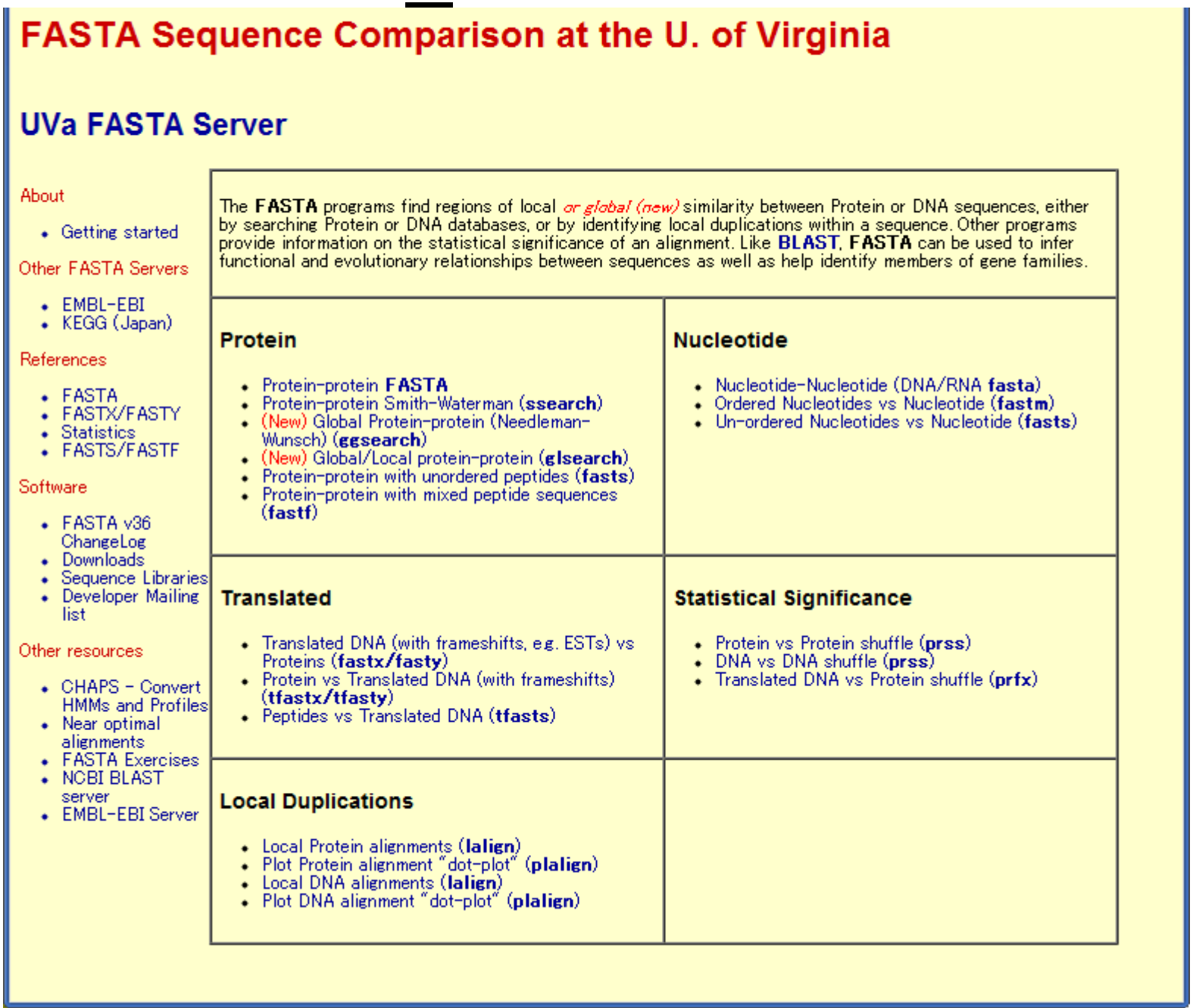
マルチプルアライメントに基づいて系統樹を作成する際の注意事項

【マルチドメインタンパク質のマルチプルアライメント】



- Global Alignment
  - Clastal W, etc.
- Local Alignment
  - FASTA, BLAST, etc.

http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml



FASTA Sequence Comparison at the U. of Virginia

UVa FASTA Server

About
- Getting started

Other FASTA Servers
- EMBL-EBI
- KEGG (Japan)

References
- FASTA
- FASTX/FASTY
- Statistics
- FASTS/FASTF

Software
- FASTA v36 ChangeLog
- Downloads
- Sequence Libraries
- Developer Mailing list

Other resources
- CHAPS - Convert HMMs and Profiles
- Near optimal alignments
- FASTA Exercises
- NCBI BLAST server
- EMBL-EBI Server

The **FASTA** programs find regions of local *or global (new)* similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like **BLAST**, **FASTA** can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Protein
- Protein-protein **FASTA**
- Protein-protein Smith-Waterman (**ssearch**)
- (New) Global Protein-protein (Needleman-Wunsch) (**ggsearch**)
- (New) Global/Local protein-protein (**glsearch**)
- Protein-protein with unordered peptides (**fasts**)
- Protein-protein with mixed peptide sequences (**fastf**)

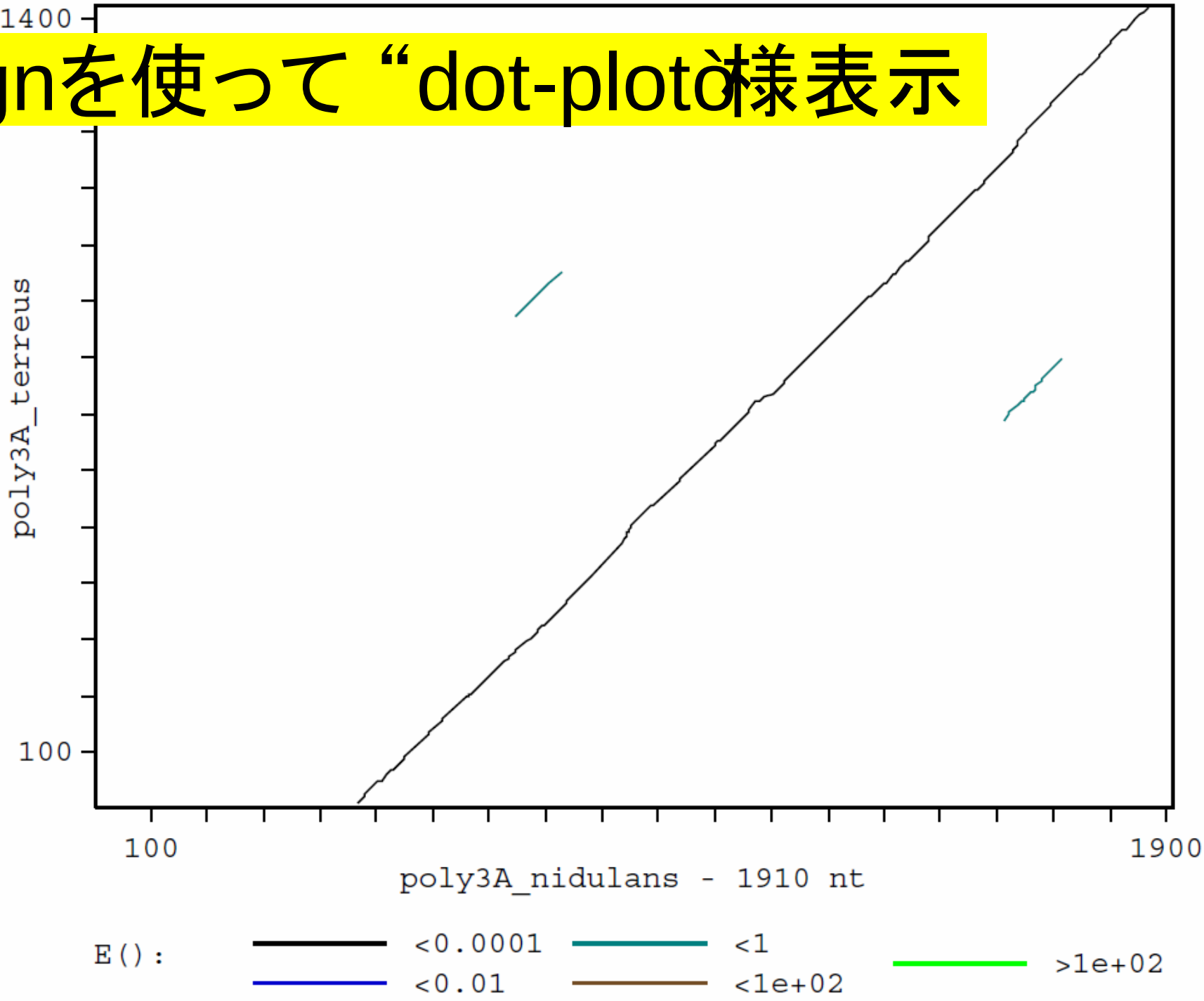
Nucleotide
- Nucleotide-Nucleotide (DNA/RNA **fasta**)
- Ordered Nucleotides vs Nucleotide (**fastm**)
- Un-ordered Nucleotides vs Nucleotide (**fasts**)

Translated
- Translated DNA (with frameshifts, e.g. ESTs) vs Proteins (**fastx/fasty**)
- Protein vs Translated DNA (with frameshifts) (**tfastx/tfasty**)
- Peptides vs Translated DNA (**tfasts**)

Statistical Significance
- Protein vs Protein shuffle (**prss**)
- DNA vs DNA shuffle (**prss**)
- Translated DNA vs Protein shuffle (**prfx**)

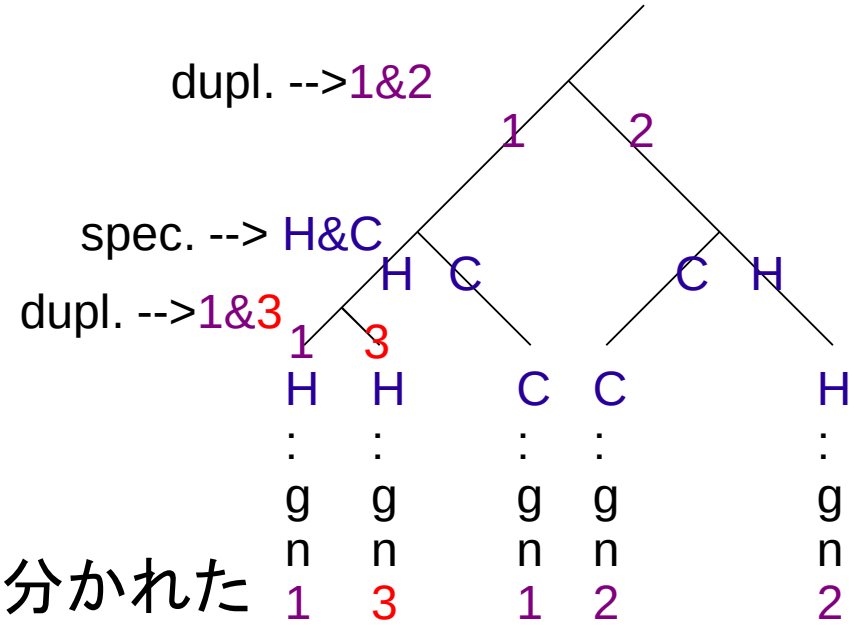
Local Duplications
- Local Protein alignments (**lalign**)
- Plot Protein alignment "dot-plot" (**plalign**)
- Local DNA alignments (**lalign**)
- Plot DNA alignment "dot-plot" (**plalign**)

palignを使って“dot-plot”様表示



／Home／進化解析／系統樹推定／

相同遺伝子



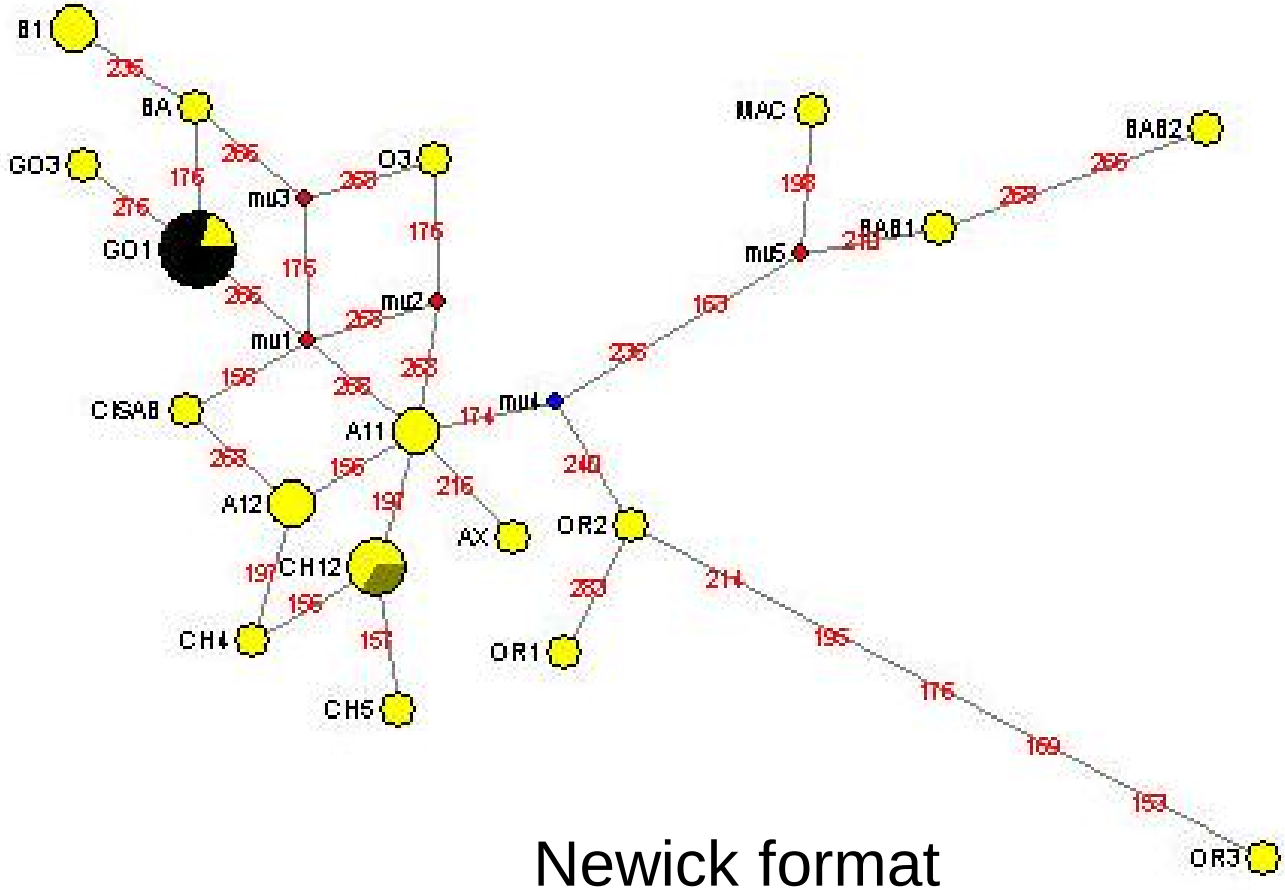
- homology
  - Orthology
    - 種分化(speciation)によって分かれた
  - Paralogy
    - 遺伝子重複 (gene duplication)によって分かれた
  - Ohnology
    - 全ゲノム重複 (whole-genome duplication)によって分かれた

＊遺伝子の機能による分類でないことに注意

系統樹・系統ネットワーク

- 系統樹 phylogenetic tree
  - 距離法
    - UPGMA法
    - NJ法
  - 最節約法
  - 最尤法
  - ベイズ法
- 系統ネットワーク phylogenetic network
  - Network 4.516
  - http://www.fluxus-engineering.com/sharenet.htm

系統樹・系統ネットワーク関連用語



node, cluster

edge, blanch

leaf node, OTU

Newick format

(MAC(RAB1, BAB2))mu5, ((OR1,OR3)OR2)mu4

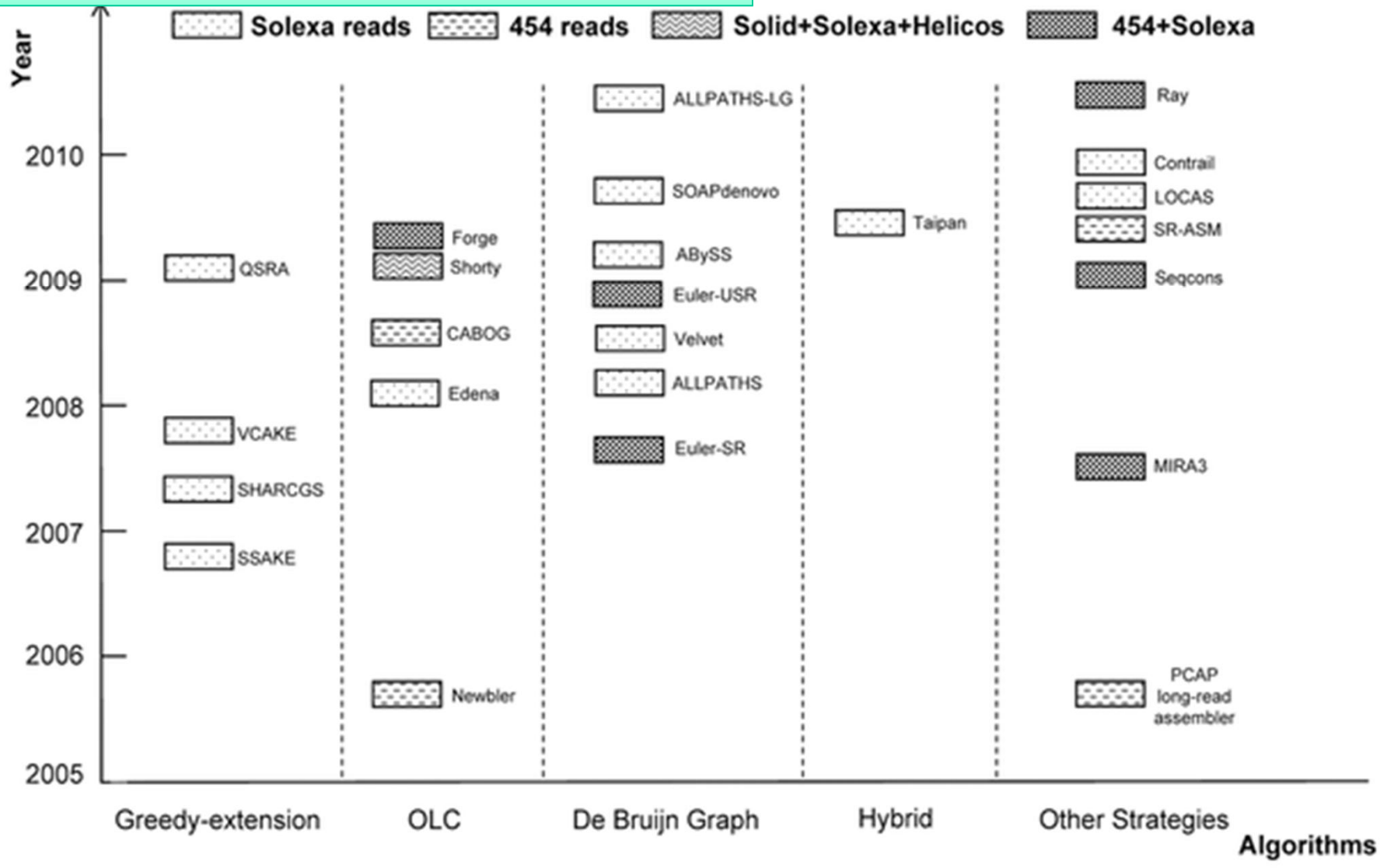
((MAC:1(RAB1:0, BAB2:2):1)mu5:2, ((OR1:1,OR3:5)OR2:1))mu4

ツールの例:
deep sequencing（NGS)関連

- 技術の特徴:
  - 膨大な数の個々のread配列を重視する傾向あり。
  - 用途の多様化:
    - 多型解析:パイプラインの条件次第でSNPになったりエラーになったりする。
      - 比較にはread配列全体が必要
    - RNA-seq:アライメントされたread配列の数が発現量を示す。
- *De novo* Genome Assemble, Mapping (Re-sequencing), SNP, RNA-seq

Figure 1. Overview of de novo short reads assemblers.

De novo ゲノムアッセンブル：要128Gくらいのメモリ。



Zhang W, Chen J, Yang Y, Tang Y, et al. (2011) A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. PLoS ONE 6(3): e17915. doi:10.1371/journal.pone.0017915
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0017915

PLoS one

Mapping (Re-sequencing)：要4Gくらいのメモリ。

Table 1: Features supported by each tool. PE.: paired-end only, mm.: mismatches, QS.: base quality score, count: total count of mismatches in the read, and missed means not supported.

	Bowtie	BWA	SOAP	MAQ	GSNAP	RMAP	FANGS
Seed mm.	Up to 3	Any	Up to 2	Any		Any	
Non-seed mm.	QS	Count	Count	QS	Count	Count	Count
Var. seed len.	> 5	Any	> 28				
Mapping qual.		Yes		Yes			
Gapped align.		Yes	PE	PE	Yes		Yes
Colorspace	Yes	Yes		Yes			
Splicing					Yes		
SNP tolerance					Yes		
Bisulfite reads					Yes	Yes	

Benchmarking Short Sequence Mapping Tools
A. Hatem, D. Bozdag, U. V. Catalyurek
IEEE International Conference on Bioinformatics and Biomedicine (BIBM11), 2011.

samtools.sourceforge.net

例）SNP検出ツール

SAMtools

SOURCEFORGE.NET®

Home

Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

SAMtools is hosted by SourceForge.net. The project page is here. The source codes are available from the download page. You can check out the latest source codes

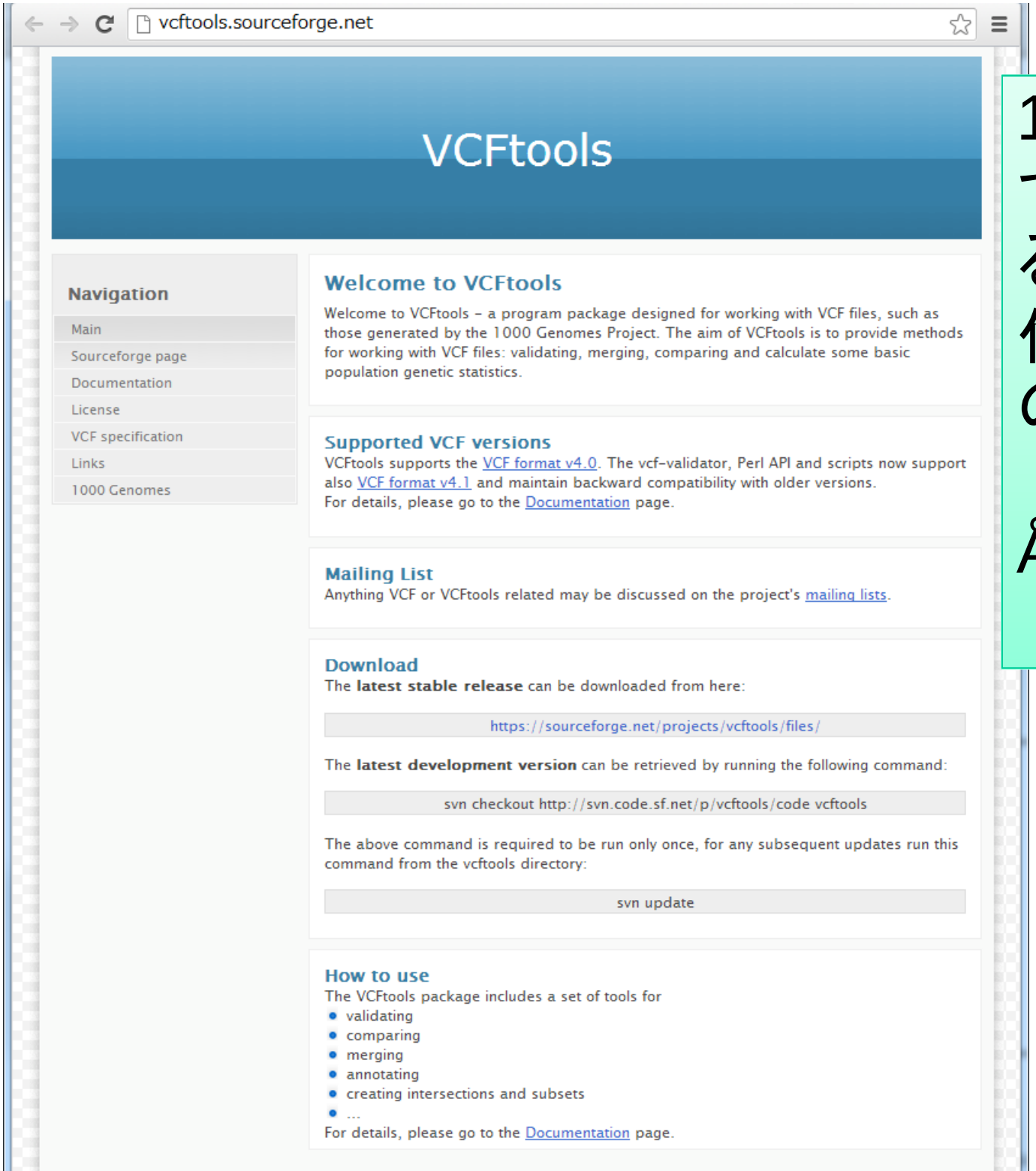
General Information

SAM Spec v1.4
SF Project Page
SF Download Page
Mailing Lists
SVN Browse
Related Software
FAQ

SAMtools in C

General Introduction
Manual Page (0.1.17)
Variant Calling (mpileup)
Text Alignment Viewer
API Documentation
Example C Program
Working on a Stream
Open Tasks
Var Calling (deprecated)
Pileup (deprecated)

Bioinformatics, 25, 2078-9. [PMID: 19505943]



1000 genomesで使用されているVCF fileを操作・加工するためのツール

- コマンド・ベース

RNA-seq & CHiP-seq関連

- RNA-seq
  - ERANGE, TopHat, MapSplice
- Chip-seq
  - ERANGE, CisGenomes, Zinba

ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

1. **Gene-based annotation**: identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected. Users can flexibly use RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, or many other gene definition systems.
2. **Region-based annotations:** identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNAse I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.
3. **Filter-based annotation:** identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF>1%) in the 1000 Genome Project, or identify subset of non-synonymous SNPs with SIFT score>0.05, or many other annotations on specific mutations.
4. **Other functionalities:** Retrieve the nucleotide sequence in any user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, identify a list of SNPs from 1000 Genomes that are in strong LD with a GWAS hit, and many other creative utilities.

SUMMARIZE_ANNOVAR is a script within the ANNOVAR package that is very popular among users. Given a list of variants from whole-exome or whole-genome sequencing, it will generate an Excel-compatible file with gene annotation, amino acid change annotation, SIFT scores, PolyPhen scores, LRT scores, MutationTaster scores, PhyloP conservation scores, GERP++ conservation scores, dbSNP identifiers, 1000 Genomes Project allele frequencies, NHLBI-ESP 5400 exome project allele frequencies and other information.

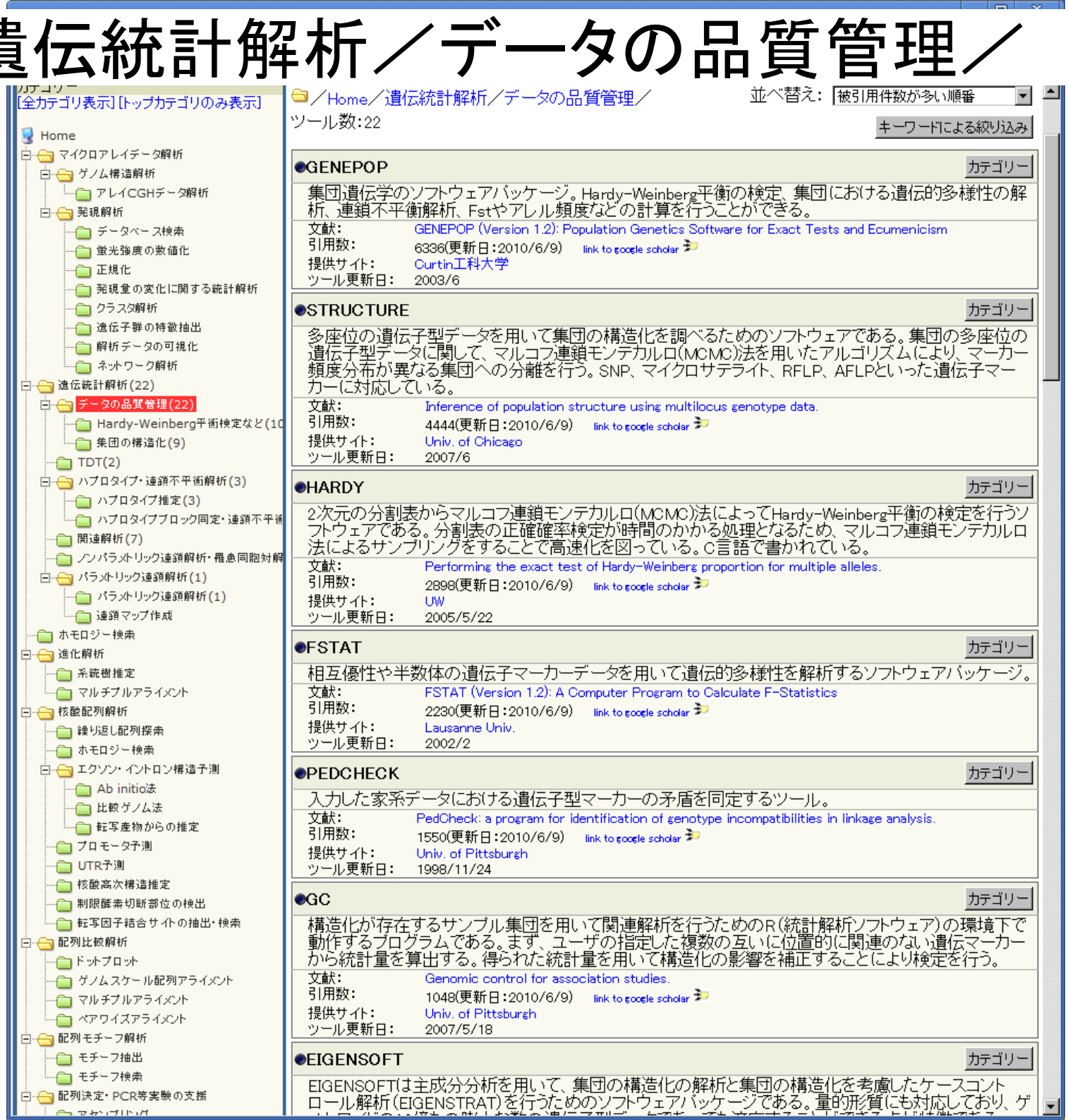
In a modern desktop computer (3GHz Intel Xeon CPU, 8Gb memory), for 4.7 million variants, ANNOVAR requires ~4 minutes to perform gene-based functional annotation, or ~15 minutes to perform stepwise "variants reduction" procedure, making it practical to handle hundreds of human genomes in a day.

What's new:

: **2012Jun24**: The **NHLBI 6500 Exome data sets** is re-uploaded as the previous version (2012Jun21) has only chr22 data. Please download again.

: **2012Jun21**: The **NHLBI 6500 Exome data sets** are available to download now. Use commands like

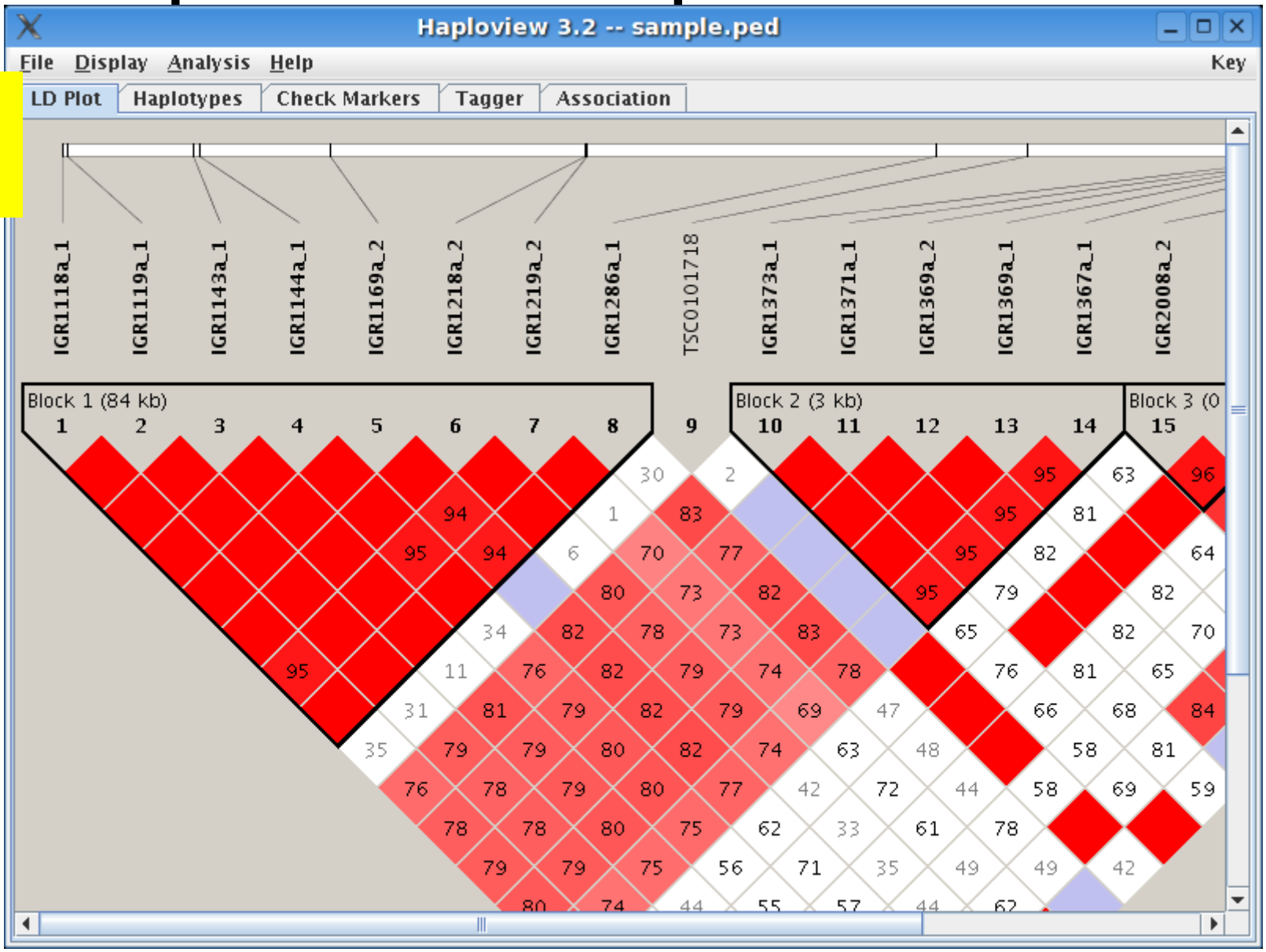
／Home／遺伝統計解析／データの品質管理／

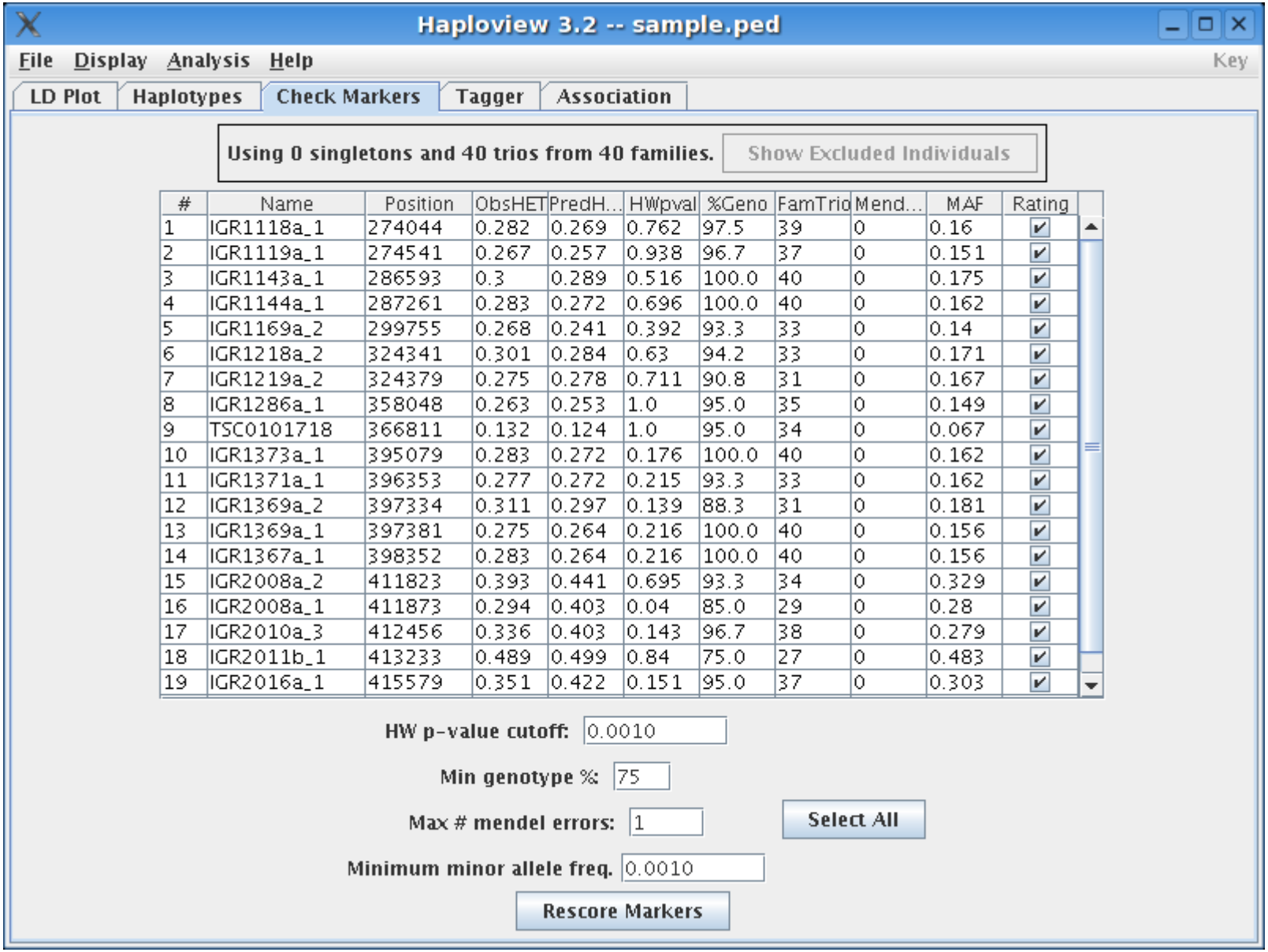



ゲノム解析ツール リンク集
User Manual | Broad Instit...
Screenshots | Broad Instit...
www-btls.jst.go.jp/Links/link.cgi?category=2300
カテゴリー
[全カテゴリ表示] [トップカテゴリのみ表示]
Home
マイクロアレイデータ解析
遺伝統計解析(27)
データの品質管理(3)
Hardy-Weinberg平衡検定など(2)
集団の構造化
TDT
ハプロタイプ・連鎖不平衡解析(27)
ハプロタイプ推定(16)
ハプロタイプブロック同定・連鎖不平衡
関連解析(4)
ノンパラメトリック連鎖解析・罹患同胞対解析
パラメトリック連鎖解析(1)
パラメトリック連鎖解析(1)
連鎖マップ作成
ホモロジー検索
進化解析
系統樹推定
マルチプルアライメント
核酸配列解析
繰り返し配列探索
ホモロジー検索
エクソン・イントロン構造予測
Ab initio法
比較ゲノム法
転写産物からの推定
プロモータ予測
UTR予測
核酸高次構造推定
制限酵素切断部位の検出
転写因子結合サイトの抽出・検索
配列比較解析
ドットプロット
／Home／遺伝統計解析／ハプロタイプ・連鎖不平衡解析／
並べ替え:
被引用件数が多い順番
ツール数:27
キーワードによる絞り込み
●GENEPOP
カテゴリー
集団遺伝学のソフトウェアパッケージ。Hardy-Weinberg平衡の検定、集団における遺伝的多様性の解析、連鎖不平衡解析、Fstやアレル頻度などの計算を行うことができる。
文献: GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism
引用数: 6336(更新日:2010/6/9) link to google scholar
提供サイト: Curtin工科大学
ツール更新日: 2003/6
●HAPLOVIEW
カテゴリー
連鎖不平衡とハプロタイプブロック解析、ブロック内のハプロタイプ推定、SNPあるいはハプロタイプによる関連解析を行い、様々な形式で出力する。HapMapプロジェクトで用いられているタグSNPを選択するアルゴリズムも組み込まれている。また、HapMapプロジェクトが提供している相分離された遺伝子型データをダウンロードする機能がある。Java言語で書かれている。
文献: Haploview: analysis and visualization of LD and haplotype maps.
引用数: 3838(更新日:2010/6/9) link to google scholar
提供サイト: MIT
ツール更新日: 2009/2/27
●PHASE
カテゴリー
家系情報のない多数個体の遺伝子型データからマルコフ連鎖モンテカルロ(MCMC)法により集団のハプロタイプ頻度を推定する。ハプロタイプの分布にコアレセンスモデルを仮定しているのが特徴である。実行形式での配布となっている。
文献: A new statistical method for haplotype reconstruction from population data.
引用数: 3118(更新日:2010/6/9) link to google scholar
提供サイト: UW
ツール更新日: Jun-04
●haplo.stats
カテゴリー
haplo.stats(旧称haplo.score)は多型の遺伝子型データからEMアルゴリズムを使用してハプロタイプ推定を行うS-PLUS/R(統計解析ソフトウェア)の環境下で動作するプログラム群である。ハプロタイプ推定の結果を用いて、関心のある表現型との関連を検定することができる。
文献: Score tests for association between traits and haplotypes when linkage phase is ambiguous.
引用数: 1141(更新日:2010/6/9) link to google scholar
提供サイト: Mayo Clinic

Haploview: LD plot

連鎖不平衡
ハプロタイプブロック
タグSNP





#	Name	Position	ObsHET	PredH...	HWpval	%Geno	FamTrio	Mend...	MAF	Rating
1	IGR1118a_1	274044	0.282	0.269	0.762	97.5	39	0	0.16	✔
2	IGR1119a_1	274541	0.267	0.257	0.938	96.7	37	0	0.151	✔
3	IGR1143a_1	286593	0.3	0.289	0.516	100.0	40	0	0.175	✔
4	IGR1144a_1	287261	0.283	0.272	0.696	100.0	40	0	0.162	✔
5	IGR1169a_2	299755	0.268	0.241	0.392	93.3	33	0	0.14	✔
6	IGR1218a_2	324341	0.301	0.284	0.63	94.2	33	0	0.171	✔
7	IGR1219a_2	324379	0.275	0.278	0.711	90.8	31	0	0.167	✔
8	IGR1286a_1	358048	0.263	0.253	1.0	95.0	35	0	0.149	✔
9	TSC0101718	366811	0.132	0.124	1.0	95.0	34	0	0.067	✔
10	IGR1373a_1	395079	0.283	0.272	0.176	100.0	40	0	0.162	✔
11	IGR1371a_1	396353	0.277	0.272	0.215	93.3	33	0	0.162	✔
12	IGR1369a_2	397334	0.311	0.297	0.139	88.3	31	0	0.181	✔
13	IGR1369a_1	397381	0.275	0.264	0.216	100.0	40	0	0.156	✔
14	IGR1367a_1	398352	0.283	0.264	0.216	100.0	40	0	0.156	✔
15	IGR2008a_2	411823	0.393	0.441	0.695	93.3	34	0	0.329	✔
16	IGR2008a_1	411873	0.294	0.403	0.04	85.0	29	0	0.28	✔
17	IGR2010a_3	412456	0.336	0.403	0.143	96.7	38	0	0.279	✔
18	IGR2011b_1	413233	0.489	0.499	0.84	75.0	27	0	0.483	✔
19	IGR2016a_1	415579	0.351	0.422	0.151	95.0	37	0	0.303	✔

Genepop

Genepop on the Web

GENEPOP is a population genetics software package originally developed by *Michel Raymond* (*Raymond@isem.univ-montp2.fr*) and *Francois Rousset* (*Rousset@isem.univ-montp2.fr*), at the Laboratiore de Genetique et Environment, Montpellier, France. The latest version of Genepop (4.0) is now available from http://kimura.univ-montp2.fr/~rousset/Genepop.htm. Genepop 4.0 runs under Windows, and can also be compiled to run under Unix or Linux. It will compile on Mac OSX machines if you have the developer tools installed. To compile under Unix or Linux, open a terminal window and cd to the Genepop source directory. Then issue the command:

'g++ -DNO_MODULES -o Genepop GenepopS.cpp -O3'

This latest version is easier to use and has some additional analyses (compared to v3.4) plus the ability to run in Batch mode.

The web version is still available for teaching purposes and for those who, for some reason, cannot run the latest version on their local PC or Mac. Below is the Genepop WWW menu with links to the data input and help pages. For further information on the Genepop program and its web implementation see the history page.

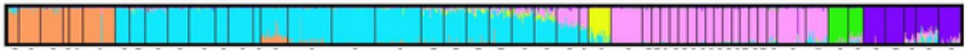
Option	Status of Web Version	Help Files
1. Hardy Weinberg Exact Tests	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 1 Help
2. Linkage Disequilibrium	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 2 Help
3. Population Differentiation	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 3 Help
4. Nm estimates	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 4 Help
5. Basic Information, Fis and gene diversities	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 5 Help
6. Fst & other correlations	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 6 Help
7. File Conversion	Equivalent to Dos versions 3.4. Includes additional file conversion to ARLEQUIN format.	Option 7 Help
8. Miscellaneous Utilities	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 8 Help

Additional Help Files

- Data input format
- Appendix 1 (null allele estimates, exact tests, markov chain probabilities, test statistics)
- Appendix 2 (Multilocus F-statistics)
- Appendix 3 (Microsatellite allele sizes, R_{ST}, and ρ_{ST}, Robertson and Hill's estimator of F_{IS}, Bootstraps)
- Bibliography

Structure

The program *structure* is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPS, microsatellites, RFLPs and AFLPs.



Download Structure 2.3.3.

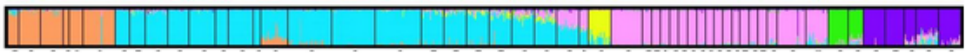
What to cite: The basic algorithm was described by Pritchard, Stephens & Donnelly (2000). Extensions to the method were published by Falush, Stephens and Pritchard (2003) and (2007) and by Hubisz, Falush, Stephens and Pritchard (2009).

Contributors: Daniel Falush, Melissa Hubisz, Matthew Stephens, Jonathan Pritchard, Peter Donnelly, William Wen, Mike Trienis, Pall Melsted.

Questions and Discussion: We have now started a Google Groups forum devoted to Structure. This replaces the Genetic Software Forum which is no longer active.

Plotting programs and other resources: ***CLUMPP*** and ***distruct*** from Noah Rosenberg's lab can automatically sort the cluster labels and produce nice **graphical displays** of *structure* results. Other plots are produced directly by the software package itself. A **free publicly available cluster** has kindly been made available for running computationally intensive *structure* jobs by CBSU at Cornell. Xavier Didelot's program xmfa2struct **converts files in eXtended Multi-Fasta (XMFA)** format into Structure input format.

Sample data sets: available here.



Taita thrush: An example of MCMC convergence based on the original paper is shown here.

Some miscellaneous applications: *structure* has been widely used for interpreting population structure of humans and other organisms. A selection of interesting references (mainly applications) is shown below.

Traces of human migrations in Helicobacter pylori populations. D. Falush, T. Wirth,

JST BIRD人材養成

http://www-bird.jst.go.jp/jinzai/


LSDB
文部科学省委託研究開発事業
統合データベースプロジェクト
Guest Account | アカウント | ログイン | English
http://lifesciencedb.jp/
ホーム
DBカタログ
横断検索
ツール
ダウンロード
About us
統合ホームページへようこそ
送信
はじめての方へ：サイトの内容をムービーやリーフレットでご紹介しています。
お知らせ 一部のサービスがJSTバイオサイエンスデータベースセンターに移動しました―
ポータル
Integbioデータベースカタログ new
(旧)生命科学系 データベース カタログ
生命科学系 学協会カタログ
生命科学系主要プロジェクト一覧
生物アイコン
ライフサイエンス 新着論文レビュー
ライフサイエンス 領域融合レビュー new
WingPro (JSTのDBポータル)
Webリソースポータルサイト (JST解析ツールポータル)
検索
生命科学データベース横断検索
蛋白質核酸酵素 全文検索
文科省「ゲノム」研究報告書 全文検索
TogoProt (蛋白質関連データベース統合検索)
OReFiL (オンラインリソースファインダー)
Allie (略語の正式名称を検索)
inMeXes (文献中の英語表現を軽快に検索)
医学・薬学予稿集全文データベース検索
データベース
DNAデータベース総覧と検索
(DDBJ/EMBL/GenBank)
遺伝子発現バンク(GEO)目次
KazusaAnnotation & Navigation (かずさDNA研究所)
KazusaMart (かずさDNA研究所)
ゲノムネット医薬品データベース (京大)
アーカイブ
生命科学系データベースアーカイブ
DDBJトレースアーカイブ (遺伝研 DDBJ)
DDBJリードアーカイブ (遺伝研 DDBJ)
ツール＆解析サービス
BodyParts3D/Anatomography
Wired-Marker
MiGAP (微生物ゲノムアノテーションパイプライン)
DBCLS Galaxy
TogoDoc (文献情報管理・推薦システム)
基盤技術開発
TogoDB(誰でもデータベースが構築できる)
TogoWS (ウェブサービスの標準化)
OpenID 認証システム
統合DB情報基盤サイト (CBRC)
辞書の構築と公開
LSDB Lab.
BioHackathon(DBCLS バイオハッカソン) 2011, 2010, 2009, 2008
教材・人材育成
統合TV (DBやツールの動画教材)
MotDB (教育・人材育成のサイト)
統合DB事業
成果報告書
パンフレット(PDF)
シンポジウム・講演会
新着情報
生命医薬情報学連合大会におけるスポンサードセッション(10月15日)他のお知らせ 2012-10-03 (Wed) 06:54:49
RefExが インターフェースを刷新。検索しやすく、結果も判りやすくなりました 2012-10-01 (Mon) 01:44:52
全サービス一時停止のお知らせ＜10月27日(土) 10:00～28日(日) 23:00＞ 2012-09-28 (Fri) 08:54:10
LSDBブログ
統合データベース講習会：AJACSりんくう～科学データは誰のものか～ 2012-10-02 (Tue) 12:14:54
バナーリンク
LEADING AUTHOR'S
ライフサイエンス 領域融合レビュー
ライフサイエンス 領域融合レビュー
FIRST AUTHOR'S
ライフサイエンス 新着論文レビュー
ライフサイエンス 新着論文レビュー
Anatomography
BodyParts3D
BodyParts3D/Anatomography
使い倒し系チャンネル
統合TV
統合TV
OReFiL
Online Resource Finder for Lifesciences
OReFiL
Allie
MAP
Allie