第1章　第3節

ゲノムアノテーションウェブサービスMEGANTEの果樹への応用

農業生物資源研究所　　伊藤　剛　・　沼　寿隆

1）ゲノムアノテーションとは

ゲノム解析を行う上で欠かせないのが，解読されたゲノム配列に付加される遺伝子構造や機能の情報，近縁生物ゲノムとの比較情報，関連文献等の様々な生物学的関連情報である．ゲノム配列にこのような注釈情報を付与することをゲノムアノテーション（genome annotation）と呼んでいる．なお，アノテーションという用語は注釈情報そのものを指す場合にも使われる．ゲノムアノテーションは一般的に，まずコンピュータで自動処理し，次にその結果を専門家が目で見て確認して，必要に応じて追加や修正を行いながら進めて行く．これらを区別する場合は，コンピュータによる自動処理を自動アノテーション，専門家による確認や修正をマニュアルアノテーションあるいはキュレーション（curation）と呼んでいる．人手によるチェックが必要な理由は，自動アノテーションには必ずと言って良いほど偽陽性（false positive）や偽陰性（false negative）といった間違いが含まれるからである．しかし，ゲノムの総塩基数が数百Mbになると，専門家がすべてをチェックするのは容易ではない．そのため，現在ゲノムデータベースとして公開されている遺伝子情報のほとんどはマニュアルアノテーションがされておらず，コンピュータが自動で付加した注釈情報のみが提供されている．短時間できわめて大量の塩基配列が解読されるようになった現在，多くの場合において自動アノテーションが唯一の選択肢で

Takeshi ITOH　✉taitoh@affrc.go.jp
Hisataka NUMA　✉hisataka@nias.affrc.go.jp
Application of genome annotation program MEGANTE to fruit trees

あり，マニュアルアノテーションよりも自動化手法の結果精度を高めることの方が現実的であろう．

自動アノテーションは複数の解析プログラムを組み合わせて行われることが多い．アノテーションのためのプログラムが多数存在する中，一つだけでは十分な精度を得ることが難しいとしても，複数のプログラムの解析結果を総合的に判断してより正確な遺伝子情報等を付加することが可能になる．例えば，2013年に解読されたスイートオレンジ（*Citrus sinensis*（L.) Osbeck）のドラフトゲノムでは，ゲノム中の遺伝子領域を予測するのに12個の解析プログラムが使われている（Xu *et al.* 2013）．複数の解析結果を総合的に判断する処理もまた，解析プログラムが自動で行うので，大規模なゲノム配列データであってもまとめて一度に処理することが可能である．ただし，解析プログラムの多くはLinuxやUnix OS向けに開発されている．普段からそのような環境でコンピュータ解析を行っていない研究者にとっては，そもそもLinuxなどでプログラムをインストールすること自体ハードルが高い．また，たとえ既存の複数の解析プログラムを組み合わせるだけであっても，小規模なソフトウェアを自作することが必須である場合も多いため，プログラミングの知識も必要である．

このような状況下，ウェブ上で動作するゲノムアノテーションのサービスとしてRiceGAAS（Sakata *et al.* 2002）が長らく利用されてきた．しかし基本的にはイネ（*Oryza sativa* L.）向けであるうえ，開発された時期が古いことから，新たなゲノム研究時代に対応した植物ゲノム自動アノテーションのシステム構築が待ち望まれていた．

2）MEGANTEとは

MEGANTE（URL1-3-1, Numa *et al.* 2014）は2013年に農業生物資源研究所が公開した，自動アノテーションを簡単に行うためのウェブサービスである．ゲノム配列のアップロードから解析結果の可視化まで，すべての工程がウェブブラウザ内で完結しているため，利用者が別途解析ソフトウェアやデータベースを用意する必要がない．また，難しいパラメータの設定も必要ない

ため，手持ちのゲノム配列を投入するだけで非常に簡単に利用することができる．2015年 5 月現在，29種のゲノムに対応しており，そのうち果樹は表1-3-1に示す7種である．この中には全ゲノムが解読されていない種も含まれているが，MEGANTEは既知のタンパク質配列や発現遺伝子配列断片（Expressed Sequence Tag, EST），及び近縁種の遺伝子情報を使って遺伝子予測を行うので，そのような種に対してもアノテーションを行うことが可能となっている．なお，全対応生物種はMEGANTEのウェブサイトで確認できる．

表 1-3-1　MEGANTEでアノテーション可能な果樹ゲノム

科名	学名	和名
Rosaceae	*Malus × domestica*	リンゴ
	Prunus persica	モモ
Rutaceae	*Citrus clementina*	クレメンティン
	Citrus reticulata	マンダリンオレンジ
	Citrus sinensis	スイートオレンジ
	Poncirus trifoliata	カラタチ
Vitaceae	*Vitis vinifera*	ブドウ

MEGANTEがゲノム配列に付加するアノテーションは遺伝子構造と遺伝子機能情報である．ここで言う遺伝子構造とはゲノム中の遺伝子の位置やエクソンの並び，翻訳領域（Open Reading Frame, ORF）のことである．また，遺伝子機能はORFから翻訳されたアミノ酸配列に与える既知のタンパク質との類似性や機能ドメイン情報のことである．

（1）MEGANTEの解析処理手順

MEGANTEは表1-3-2の解析プログラムやデータベースを使って，以下の手順に従ってゲノムアノテーションを実行する（図1-3-1）．

① RepeatMaskerでゲノム配列中の反復配列を検出する．

② 完全長cDNA配列をBLATでゲノム配列にアラインメント（alignment）する．

表 1-3-2　MEGANTE内部で使用している解析プログラムとデータベース

解析プログラム・データベース	ツール名	URL	論文
反復配列検出	RepeatMasker	http://www.repeatmasker.org/	
遺伝子予測	AUGUSTUS	http://bioinf.uni-greifswald.de/augustus/	Stanke, M. *et al.* (2003)
	GeneZilla	http://www.genezilla.org/	Allen, J. E. *et al.* (2006)
	GlimmerHMM	http://cbcb.umd.edu/software/glimmerhmm/	Allen, J. E. *et al.* (2006)
	SNAP	http://korflab.ucdavis.edu/software.html	Korf, I. (2004)
	JIGSAW	http://www.cbcb.umd.edu/software/jigsaw/	Allen, J. E. *et al.* (2006)
cDNA/ESTアラインメント	BLAT	http://genome.ucsc.edu/FAQ/FAQblat.html	Kent, W. J. (2002)
	sim4db	http://sourceforge.net/projects/kmer/	Walenz, B. *et al.* (2011)
	PASA	http://pasapipeline.github.io/	Haas, B. J. *et al.* (2003)
タンパク質アラインメント	ProSplign	http://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html	Sayers, E. W. *et al.* (2012)
遺伝子機能アノテーション	NCBI BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	Camacho, C. *et al.* (2009)
	InterProScan	http://www.ebi.ac.uk/interpro/interproscan.html	Quevillon, E. *et al.* (2005)
解析結果可視化	GBrowse	http://gmod.org/wiki/GBrowse	Stein, L. D. *et al.* (2002)
データベース	NCBI GenBank (cDNA/EST)	http://www.ncbi.nlm.nih.gov/genbank/	Benson, D. A. *et al.* (2015)
	PGSB RE-cat (植物反復配列)	http://mips.helmholtz-muenchen.de/plant/recat/index.jsp	Nussbaumer, T. *et al.* (2013)
	UniProtKB (タンパク質)	http://www.uniprot.org/	Magrane, M. *et al.* (2011)
	InterPro (機能ドメイン)	http://www.ebi.ac.uk/interpro/	Hunter, S. *et al.* (2012)

③　(2) でアラインメントされたゲノム領域を結合して仮想mRNA配列を作成し，最長ORFを探す．

④　EST配列をPASAでゲノム配列にアラインメントする．

⑤　sim4dbで他生物種の完全長cDNA配列をゲノム配列にアラインメントする．

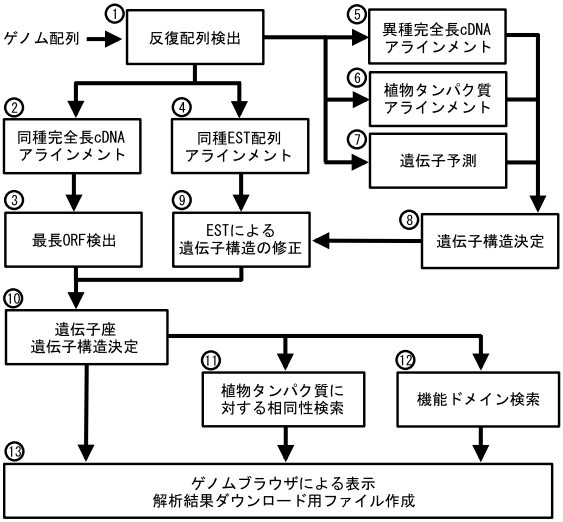

図1-3-1　MEGANTEのゲノムアノテーション解析処理手順

⑥ UniProtKBデータベースからダウンロードした植物のタンパク質配列をProSplignでゲノム配列にアラインメントする．

⑦ ４つの遺伝子予測プログラム（AUGUSTUS，GeneZilla，GlimmerHMM，SNAP）で遺伝子領域を予測する．

⑧ JIGSAWで（5）（6）（7）の結果を総合的に判断して遺伝子構造を決める．

⑨ PASAで（4）の結果を使って（8）の遺伝子構造を修正し，より正しい構造にする．

⑩ （3）と（9）の結果をまとめて最終的な遺伝子座と遺伝子構造を決定する．

⑪ BLASTPで（6）と同様のUniProtKBを検索して，（10）の結果に既知のタンパク質に対する類似性情報を付加する．

⑫ InterProScanで（10）の結果にタンパク質機能ドメイン情報を付加する．

⑬ すべての解析結果をデータベースに保存し，GBrowseで閲覧できるようにする．また解析結果を一括ダウンロードできるようにZIP形式のファイルにまとめる．

反復配列を検出した後，反復配列中の塩基をNまたは小文字に置換するリピートマスク（repeat masking）処理が行われる．この処理を行わないと，アラインメントの際に反復配列に一致する領域が大量に出てきてしまったり，あるいはトランスポゾン（transposon）を遺伝子の一部と予測してしまう．どちらもアノテーションの精度を下げることにつながるので，リピートマスクは精度向上のための重要な処理である．果樹のように，あまり多くの完全長cDNAが登録されていない生物種に対しては（2）の処理は行われない．MEGANTEが使用している完全長cDNAやEST，タンパク質配列は定期的にアップデートされており，更新日や登録件数はMEGANTEのウェブサイトで確認できる．なお，より詳しい解析手順は文献（Numa *et al.* 2014）に記載されている．

（2）果樹への適用

MEGANTEによるアノテーションの正確性を評価するため、公開されている*C. sinensis*のゲノム情報（Xu *et al.* 2013）から無作為に計1,000遺伝子を含む領域を取り出し，MEGANTEで自動アノテーションを行った．表1-3-3はMEGANTEの遺伝子予測精度を示している．比較のためMEGANTE内部で使用している4つの遺伝子予測プログラムでも同様の評価を行っている．MEGANTEは前述した通り複数の解析プログラムの結果を総合的に判断して予測を行っているので，他のプログラムよりも予測精度は高い．しかしそれでも，評価に用いた全遺伝子の内，遺伝子単位で少しのミスもなく予測できたのは5割以下である．高等真核生物の遺伝子構造は複雑であるため，MEGANTEに限らず，公開されているゲノムデータベースの予測遺伝子にも同様のミスは含まれる．したがって研究者自身も，複数の情報を加味してアノテーションの確からしさを総合的に判断する必要がある．例えばゲノムデータベースでは予測遺伝子と共に，既知のタンパク質配列との類似性情報やRNA-seqによる発現を示すデータ等が提供される場合が多いので，アノテーションされた遺伝子の確からしさを確認するためにそのような情報が活用できる．

表1-3-3　MEGANTEおよび各遺伝子予測プログラムの遺伝子予測精度

プログラム	エクソン単位の評価		遺伝子単位の評価	
	感度	特異度	感度	特異度
MEGANTE	0.83	0.81	0.45	0.40
AUGUSTUS	0.76	0.72	0.34	0.31
GeneZiila	0.67	0.61	0.25	0.19
GlimmerHMM	0.57	0.59	0.28	0.16
SNAP	0.67	0.57	0.20	0.14

エクソン単位の評価の場合，エクソン一つの開始位置と終了位置が正しく予測されていれば良い．遺伝子単位の評価の場合は一つの遺伝子に含まれるすべてのエクソンの位置が正しく予測されていないと正解とみなさない．感度は，評価に用いたエクソンまたは遺伝子のうち正しく予測できた割合である．例えば，10個の遺伝子のうち 5個を正しく予測できていれば，感度は0.5である．特異度は，予測したエクソンまたは遺伝子のうち，正しく予測できた割合である．例えば，10個の遺伝子をすべて正しく予測できたとしても，予測した遺伝子数が20であれば特異度は10/20で0.5である．100%の精度で予測できれば，感度と特異度は，両方1になる．この表の評価に用いたのはタンパク質コード領域（coding sequence，CDS）のみで,非翻訳領域（untranslated region，UTR）は含まれていない．

3）MEGANTEを利用するには

MEGANTEを初めて使う場合を例に，使い方や解析結果の見方を概説する．事前に用意する必要があるのは，インターネット接続，ウェブブラウザ，ゲノム配列，e-mailアドレスである．ウェブブラウザさえ動作すれば良いので高性能なパソコンは必要ない．

(1) ユーザ登録とログイン

MEGANTEを初めて使う場合にはまずユーザ登録が必要である．トップページ（https://megante.dna.affrc.go.jp/）にアクセスし（図1-3-2A），「Create an account」ボタンをクリックしてユーザ登録画面に進む（図1-3-2B）．そこでe-mailアドレスとパスワードを入力し「Sign up」ボタンを押す．すると入力したe-mailアドレス宛に図1-3-3のようなユーザ登録を完了させるためのメールが届くので，その中に書かれているリンクをクリックする．クリックすればユーザ登録は完了である．ユーザ登録が完了すればMEGANTEにログイ

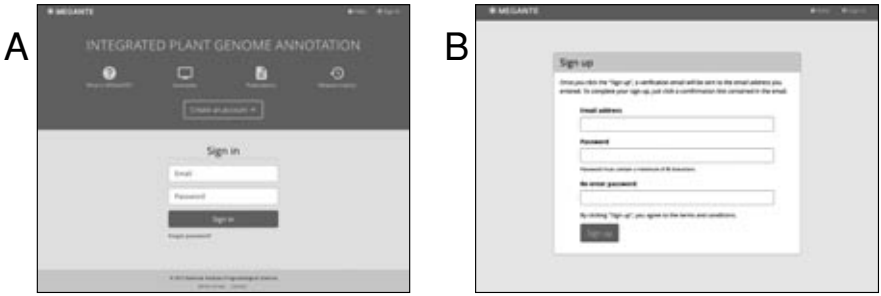

図1-3-2　MEGANTEトップページとユーザ登録画面

A:　トップページ．B:　ユーザ登録画面

Hi your@email.address,

Thanks for signing up for MEGANTE website.

Please click the following link to activate your account:
https://megante.dna.affrc.go.jp/activate/LB_R9gdo…　← このURLにアクセスするとユーザ登録が完了する

If you have any questions or comments,
please email us at megante@ml.affrc.go.jp.

--
MEGANTE support team
(https://megante.dna.affrc.go.jp/)

図1-3-3　登録確認メールの例

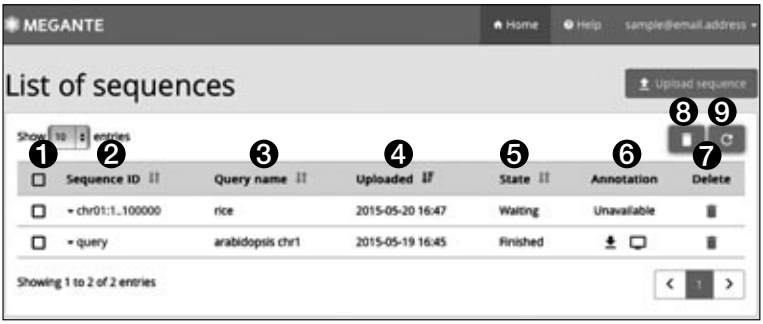

図1-3-4　ゲノム配列一覧画面

アップロードしたゲノム配列と解析結果を確認できる．①チェックボックス．②ゲノム配列のID．ゲノム配列をFASTA 形式でアップロードすると，その中に記載されているコメント行の内容がIDとして表示される．③ゲノム配列アップロード時にメモ欄に記入した内容が表示される．④ゲノム配列をアップロードした日時．⑤解析状況．⑥アノテーション結果へのリンク．⑦削除アイコン．該当行のゲノム配列および解析結果を削除する．⑧一括削除アイコン．チェックボックスがチェック済みのゲノム配列および解析結果をすべて削除する．⑨再読み込みアイコン．解析状況を更新する．

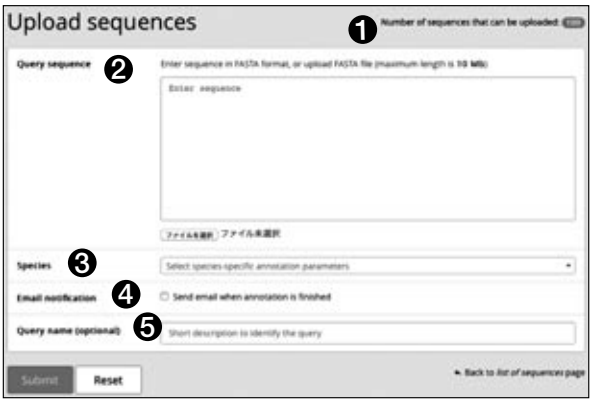

図1-3-5　ゲノム配列アップロード画面

①アップロード可能配列数．②ゲノム配列をコピーする．あるいはアップロードしたいゲノム配列ファイルを選択する．③ゲノム配列の生物種を選択する．④チェックすると解析終了後にメールが届く．⑤メモ欄（任意）．

ンできるようになるので，再びトップページにアクセスし，画面下のログインフォームにe-mailアドレスとパスワードを入力し，「Sign in」ボタンを押してログインする．

(2) ゲノム配列のアップロード

ログインすると最初に表示される画面は図1-3-4である．ここには過去に入力したゲノム配列の情報や解析結果へのリンクが表示される．初めてログインした時にはまだゲノム配列を投入していないのでそれらは表示されない．配列を入力するには画面右上の「Upload sequence」ボタンをクリックし，配列入力画面へと進む（図1-3-5）．ここで配列を直接コピーするか，もしくは「ファイルを選択」のボタンをクリックしてファイルをアップロードする．マルチFASTA形式にすれば複数のゲノム配列を一度にアップロードすることが可能である．なお 1 配列あたりの受け入れ可能な最大長は10 Mbである．また，システムが保存できる最大配列数はユーザあたり100であるので，アップロードし

た配列が100に達するとそれ以上は新規にアップロードできなくなる．しかし，最初の画面に戻って過去にアップロードした配列を削除すれば，新たにアップロードすることが可能になる．次に「Species」メニューから生物種を選択する．生物種ごとに解析パラメータや使用するデータベースが違うので，入力したゲノム配列と同じもしくは近縁の生物種を選択する．その下のチェックボックスをオンにすると解析が完了した後にメールで連絡が来るようになる．一番下のテキストボックスはメモ欄である．ここにアップロードした配列が何であるかを入力しておけば，たくさんの配列をアップロードした場合にそれぞれを識別できて便利である．入力が完了したら画面下の「Submit」ボタンを押す．アップロードが成功すれば最初の画面（図1-3-4）に戻る．ゲノム配列はサーバ上に保存されるので，アップロード後はウェブブラウザを閉じてしまっても構わない．

(3) ゲノム配列一覧

配列がアップロードされると図1-3-4のように解析状況を確認できる．アップロードされた配列は待ち行列に入れられ，順次解析処理が行われる．待ち行列に入ると「State」列が「Waiting」になり，解析処理が開始されると「Running」に変わる．解析が完了すると「State」が「Finished」になり，「Annotation」列が「Unavailable」から2つのアイコンに変わる．左のアイコンをクリックすると解析結果がZIP形式のファイルとしてダウンロードできる．右のアイコンをクリックするとゲノムブラウザが開き，解析結果を確認できる．これらの見方は次節で説明する．「Sequence ID」や「Query name」が長いとすべて表示されないが，「Sequence ID」をクリックすると行が下に開くので，そこで内容を確認できる．また，その画面からゲノム配列をダウンロードすることもできる．

データを削除する場合は，「Delete」列のゴミ箱アイコンをクリックする．確認画面が開くので「Yes」を押せばゲノム配列と解析結果は削除される．複数の配列を一度に削除する場合は，削除したい配列の一番左のチェックボックスをクリックしてチェックを入れ，右上のゴミ箱アイコンをクリックする．同様に確認ウィンドウが開くので「Yes」をクリックすれば選択した配列がすべ

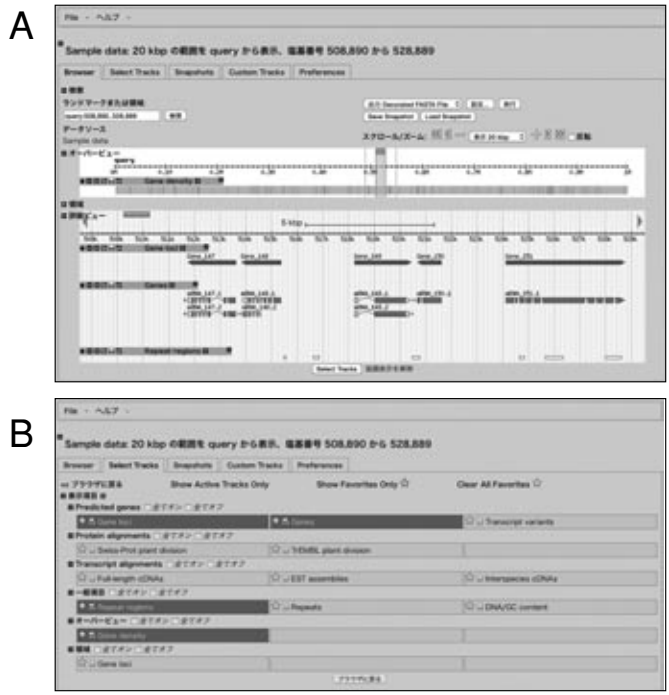

図1-3-6　ゲノムブラウザによる解析結果の表示

A: アノテーション表示画面．初期状態では遺伝子座（Gene loci），遺伝子構造（Genes），反復配列領域（Repeat regions）が表示される．Genesの赤色部分はタンパク質コード領域，ピンク色は 5′ 非翻訳領域，肌色は 3′ 非翻訳領域を表している．Genes全体が灰色になっているのはタンパク質非コード型RNAである．

B: 表示項目選択画面で，Aに表示する項目を選択できる．ここでは初期状態で表示される予測遺伝子に加えて，既知の植物タンパク質配列のアラインメント結果（Protein alignments），cDNAおよびESTのアラインメント結果（Transcript alignments），個々の反復配列（Repeats）が選択できる．「DNA/GC content」を選択すると，アノテーション画面に塩基配列（ズームイン時）もしくはGC含量（ズームアウト時）が表示される．

て削除される．なお一度削除してしまうと元に戻すことはできない．

標準ではアップロードした日時が新しい方から20件表示されるが，左上のメニューから表示件数を変更することが可能である．また，各列のヘッダ部分をクリックすれば，その項目でソートすることができる．

(4) 解析結果の表示

ゲノム配列一覧（図1-3-4）の「Annotation」列にある右側のアイコンをクリックすると，図1-3-6Aのようなゲノムブラウザが開く．このブラウザはゲノムデータベースでよく利用されているGBrowseを元に作られているので，それと同様な操作が可能である．ゲノムブラウザには標準で遺伝子座（Gene loci），遺伝子構造（Genes），反復配列領域（Repeat regions）が表示される．それ以外の項目を表示するには，画面上の「Select Tracks」タブをクリックして表示項目選択画面に切り替え，表示したい情報のチェックボックスにチェックを入れる（図1-3-6B）．

遺伝子構造（Genes）にマウスを合わせればポップアップが開いて遺伝子の概要が確認できる．クリックすると詳細を表示する画面が新たに開く（図1-3-7）．ここでは前述の，既知のタンパク質との相同性や機能ドメイン情報が確認できる．また，遺伝子のエクソン・イントロン構造，ORF配列，アミノ酸配列も閲覧できる．

(5) 解析結果のダウンロード

ゲノム配列一覧（図1-3-4）「Annotation」列の左のアイコンをクリックすると，解析結果一式がZIP形式のファイルとしてダウンロードできる．解凍すると表1-3-4に示すファイルが現れる．アノテーション情報はExcelファイルだけでなく，GFF 3 形式（Stein, 2015-05-01）でも提供している．GFF3はアノテーションデータの保存形式として幅広く採用されているので，次世代シークエンサデータの表示にもよく利用されているIGV（URL1-2-28, Thorvaldsdóttir *et al.* 2013）などの外部ツールに読み込ませれば図示することが可能である．

(6) アカウント管理

ログイン後は画面右上に表示されているe-mailアドレス（図1-3-4）をクリックするとメニューが現れるので，そこからログアウト（Sign out）や，e-mailアドレスおよびパスワードの変更ができる．ユーザのアカウントはユーザ

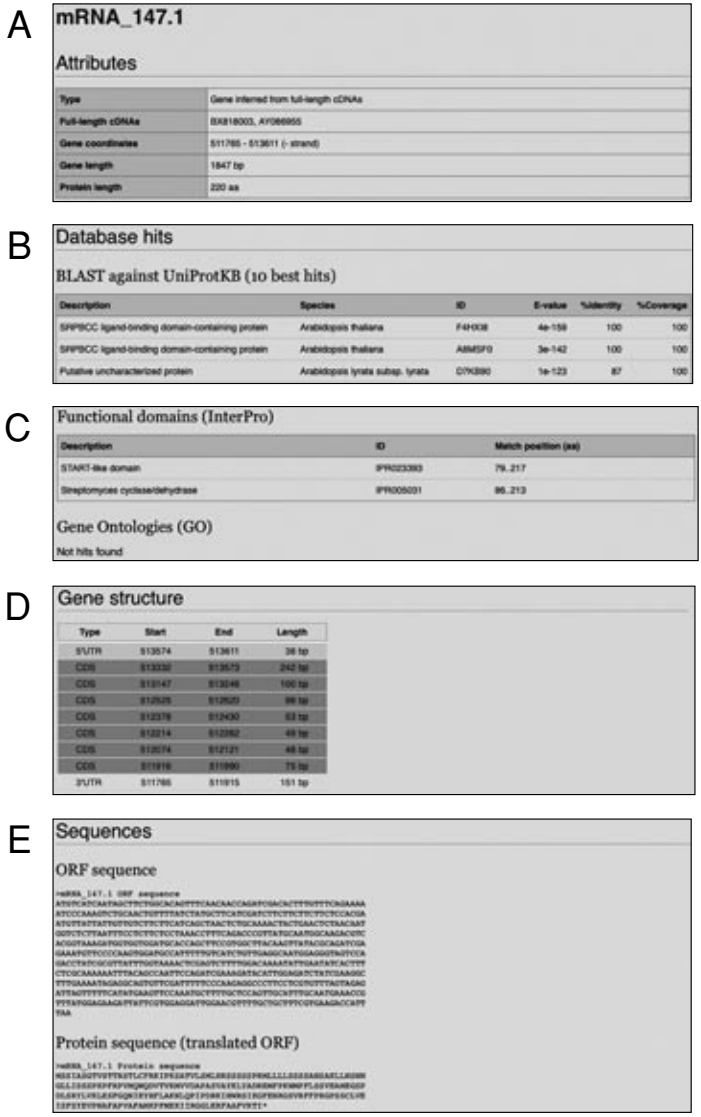

図1-3-7　遺伝子詳細画

（A）基本情報．（B）BLASTPの結果（UniProtKBタンパク質に対する相同性情報）．（C）InterProScanの結果（機能ドメイン情報）．（D）遺伝子構造情報．（E）ORFの塩基配列およびアミノ酸配列．

表1-3-4　解析結果に含まれるファイル

名前	形式	内容
annotation.gff	GFF3	下記blast.xlsx, domain.xlsx, gene.xlsxの内容をすべて含むアノテーション情報
blast.xlsx	Microsoft Excel	予測遺伝子に類似のタンパク質情報(BLASTPの結果)
domain.xlsx	Microsoft Excel	予測遺伝子の機能ドメイン情報(InterProScanの結果)
gene.xlsx	Microsoft Excel	予測遺伝子の構造情報
orf.fasta	FASTA	予測遺伝子のORF配列
protein.fasta	FASTA	予測遺伝子のタンパク質配列
query.fasta	FASTA	ゲノム配列
readme.txt	テキストファイル	上記ファイルに関する簡単な説明

自身が明示的に削除しない限りサーバ上に保存される．アカウントを削除したい場合は「Delete account」をクリックするとアカウント削除フォームに進めるので，そこでアカウントを削除する．一度削除を行うとアカウントに加えてサーバ上に保存されている配列ファイルや解析結果はその場で削除されるので，元に戻すことはできない．

パスワードを忘れてしまった場合にはホームページ（図1-3-2A）の「Forgot password?」をクリックしてパスワード再発行フォームに進む．そこで登録したe-mailアドレスを入力するとパスワード再発行のための手順がメールで届くので，それに従ってパスワードを再発行する．

4）応用事例と今後の展望

MEGANTEは遺伝子単離のための補助ツールとして開発されたウェブサービスであり，せいぜい数百kbから1 Mb程度のゲノム領域に含まれる候補遺伝子を絞り込む目的で利用されることを想定していた．しかし，サービス公開後は全ゲノムアノテーションに利用される場合も多く，例えばゲノムサイズが約540 Mbであるアズキゲノムのアノテーションにも利用されている（Sakai *et al.* 2015）．利用者一人当たり最大1 Gbのゲノム配列を処理できるので，ゲノムサイズが1 Gb以下なら全ゲノムアノテーションを一度に実行できる．一方，

遺伝子単離が目的の利用者にとっては大量のゲノム配列を処理できることよりも，使いやすさの方が重要であると考えられる．現状MEGANTEは予測した遺伝子に対して遺伝子機能の予測は行っていない．その理由は一般的にコンピュータ解析だけで正確に遺伝子機能を予測することは困難だからである．しかし，候補領域にどのような機能の遺伝子があるか一覧できれば有用であると思われるので，MEGANTEに遺伝子機能予測機能を実装し，その結果をゲノムブラウザ上で表示できるよう準備を進めている．

引用文献

Allen, J. E. *et al.*（2006）JIGSAW, GeneZilla, and GlimmerHMM:puzzling out the features of human genes in the ENCODE regions. Genome Biology. 7（Suppl. 1）:S9.1-S9. 13.

Benson, D. A. *et al.*（2015）GenBank. Nucleic Acids Research. 43:D30-D35.

Camacho, C. *et al.*（2009）BLAST+:architecture and applications. BMC Bioinformatics. 10:421.

Haas, B. J. *et al.*（2003）Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 31:5654-5666.

Hunter, S. *et al.*（2012）InterPro in 2011:new developments in the family and domain prediction database. Nucleic Acids Research. 40:D306-D312.

Kent, W. J.（2002）BLAT-the BLAST-like alignment tool. Genome Research. 12:656-664.

Korf, I.（2004）Gene finding in novel genomes. BMC Bioinformatics. 5:59.

Magrane, M. *et al.*（2011）UniProt Knowledgebase:a hub of integrated protein data. Database（Oxford）. 2011:bar009.

Numa, H. *et al.*（2014）MEGANTE:a web - based system for integrated plant genome annotation. Plant and Cell Physiology. 55:e2.

Nussbaumer, T. *et al.*（2013）MIPS PlantsDB:a database framework for comparative plant genome research. Nucleic Acids Research. 41:D1144-

D1151.

Quevillon, E. *et al.* (2005) InterProScan:protein domains identifier. Nucleic Acids Research. 33:W116-W120.

Sakai, H. *et al.* (2015) The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. Scientific Reports. 5:16780.

Sakata, K. *et al.* (2002) RiceGAAS:an automated annotation system and database for rice genome sequence. Nucleic Acids Research. 30:98-102.

Sayers, E. W. *et al.* (2012) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 40:D13-D25.

Stanke, M. *et al.* (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 19 (Suppl. 2):ii215-ii225.

Stein, L. D. *et al.* (2002) The generic genome browser:a building block for a model organism system database. Genome Research. 12:1599-1610.

Stein, L. (2013) Genetic Feature Format Version 3 (GFF). The Sequence Ontlogy Project. http://www.sequenceontology.org/gff3.shtml 2015-05-01.

Thorvaldsdóttir, H. *et al.* (2013) Integrative Genomics Viewer (IGV):high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 14:178-192.

Walenz, B. *et al.* (2011) Sim4db and Leaff:utilities for fast batch spliced alignment and sequence indexing. Bioinformatics. 27:1869-1870.

Xu, Q. *et al.* (2013) The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics. 45:59-69.