農業生物のゲノム情報解析研究

－国産品種ダイズのゲノム配列解析とカイコゲノムデータベースの開発－

下村　道彦

学籍番号 1356503

前橋工科大学
大学院工学研究科
博士後期課程　環境・生命工学専攻
博士論文（本審査）

2018 年 12 月

目次

概要

第 1 章　はじめに 1

1.1.　ゲノム研究の動向 1
1.1.1.　ゲノム解読の歴史 3
1.1.2.　シークエンシング 7
1.1.3.　アセンブリング 9
1.1.4.　マッピング 10
1.1.5.　遺伝子モデリング 13
1.1.6.　データベース開発 14
1.2.　農業生物ゲノム研究の課題 15
1.2.1.　ゲノム構築・解析 15
1.2.2.　データベース開発 16
1.3.　研究目標と方法 17
1.4.　本文の構成 18

第 2 章　国産ダイズゲノム構築・解析 20

2.1.　概要 20
2.2.　はじめに 20
2.3. 材料と方法 22
2.3.1.　ゲノムシークエンシング 22
2.3.2.　アセンブルとレファランスマッピング 22
2.3.3.　遺伝子モデリング 23
2.3.4.　系統解析 24
2.3.5.　アントシアニン・フラボノイド生合成系 24
2.3.6.　プロテオーム解析 24
2.4.　結果と考察 25
2.4.1.　ゲノムシークエンシングとレファランスマッピング 25
2.4.2.　一塩基多型、挿入・欠失 26
2.4.3.　遺伝子モデル 27

2.4.4. 系統解析 28
2.4.5. アントシアニン・フラボノイド生合成系 29
2.4.6. 子葉におけるタンパク質 33
2.4.7. エンレイゲノムデータベース 34
2.5. 結論 34
第 3 章　カイコゲノム統合データベース開発 36
3.1. 概要 36
3.2. はじめに 37
3.3. データセット内容 38
3.3.1. ゲノム配列情報 38
3.3.2. ゲノム配列にマップされる情報 39
3.3.3. プロテオーム情報 39
3.3.4. 発現遺伝子可視化情報 40
3.4. データベース KAIKOBASE の構成 40
3.5. 使用方法と考察 44
3.5.1. ユーザインタフェース 45
3.5.2. キーワードサーチ、ポジションサーチ 46
3.5.3. シークエンスサーチ 46
3.6. 結論 46
第 4 章 結言 48
謝辞 49
参考文献 50

概要

1865 年にメンデルが形質は遺伝することを、1913 年にモーガンらが染色体上に遺伝子は存在することを、1953 年にワトソンとクリックが DNA は相補性を持つ二重螺旋構造であることを発見した。DNA の二重螺旋構造発見以降、ゲノム DNA が細胞でどのように作用するかの研究が進んだ。ゲノム解析に着目すると、既知のガン遺伝子の変異を調べる上で、個々の遺伝子に着目した研究が行われてきたが、1986 年にダルベッコは、ヒトゲノムの塩基配列を全部決定することがブレークスルーに繋がると考え、ゲノム配列の重要性を説いた。これが契機となり、全ゲノム配列獲得の実現に向けての研究が開始された。この流れの中で、1995 年にインフルエンザ菌ゲノムを皮切りに、1998 年に線虫ゲノム、2004 年にヒトゲノムが解読された。植物分野では、2000 年にシロイヌナズナゲノム、2005 年にイネゲノム、2010 年にダイズゲノム、2015 年にダイズゲノム（エンレイ品種）、昆虫分野では、2000 年にショウジョウバエゲノム、2008 年にカイコゲノムが解読された。

ダイズ研究においては、その遺伝子構造や機能解析であれば、2010 年に解読されたダイズゲノム Williams 82 品種の使用で十分である。しかし、ダイズの育種では DNA マーカーを使用した育種が行われており、国内での育種は日本産品種同士の掛け合わせになることが多い。Williams 82 ゲノムと日本産品種ダイズは同じダイズ種ではあるが、系統が離れているため、Williams 82 から得られた DNA マーカーが使用できない場合がある。このため、国産ダイズ品種エンレイのゲノム構築・解析が必要となった。

昆虫分野のカイコにおいては、ゲノム研究プロジェクトが推進され、プロジェクトで得られたゲノム情報や関連する研究の情報をまとめ、効率的に研究に役立つ情報を取り出す仕組みが必要となった。

本研究では、（１）国産品種ダイズであるエンレイ品種のゲノム配列を解読した。エンレイ品種のゲノム配列は、国内の栽培事情に適したダイズの品種改良のための様々な情報を提供する。（２）カイコゲノム情報を提供する統合カイコゲノム統合データベース KAIKObase を開発した。KAIKObase は、鱗翅目の研

究だけでなく、養蚕の改善や新しい害虫駆除手法研究に向けた、データマイニングとゲノム応用を容易にする。

本論文第一章では、ゲノム研究の動向、ゲノム解読やゲノム情報解析を支える技術の背景と動向を示した後、本研究の研究目標と研究戦略を述べる。

第二章では、国産ダイズゲノム解析を実施し、栽培品種エンレイのゲノムを解読した研究について述べる。その研究では、次世代シークエンサを用いて得られた全ゲノム配列を、栽培品種 Williams 82 ゲノムにレファランスマッピングして、エンレイゲノム配列 約 928Mb の塩基配列を決定した。遺伝子予測ソフトウェアで作成した遺伝子モデル 107,423 個からリピート配列、およびトランスポゾンを除き、最終的に、60,838 個のスプライスバリアントがない遺伝子モデルを得た。系統解析では、エンレイおよび Williams 82 品種双方の系統関係、および野生ダイズを含む複合体を含む系統関係を考察した。エンレイと Williams 82 の遺伝子モデルを比較し、アントシアニン・フラボノイド生合成に関連するパスウェイ、および８番染色体上の CHS 遺伝子クラスタで両品種の違いを示した。また、登熟期の子葉のプロテオームから全体的なプロファイルを分析した。配列データは、DAIZUbase に統合化し利用可能とした。これらの研究成果は、我が国の広範なダイズ品種の比較ゲノミクスに資する包括的な情報資源と、国内外のダイズ品種の改良のための有効な情報となる。

第三章では、効果的なデータマイニングとゲノム応用のためのカイコゲノム情報を提供するカイコゲノム統合データベース KAIKObase の開発について述べる。KAIKObase に、カイコゲノム配列、ゲノム地図情報および EST データを統合した。KAIKObase は、塩基配列、遺伝子、スキャフォルド、染色体の各段階のデータを 4 種類の MapViewer（PGmap、UnifiedMap、UTGB、GBrowse）、GeneViewer、配列検索、キーワード・位置検索で表示する。さらに、プロテオームデータ用の KAIKO2DDB と遺伝子導入およびレポータデータ用の *Bombyx* trap データベースの統合により、KAIKObase の機能をさらに強化した。カイコの研究には、包括的なカイコゲノムデータベースが不可欠であり、KAIKObase は鱗翅目の研究だけでなく、養蚕の改善や新しい害虫駆除法の研究を容易にする。

第四章では、結言として本研究の成果について纏める。

図リスト

図番号	タイトル
図 1-1	ゲノム解析における概略フロー
図 2-1	分岐年代
図 2-2	アントシアニン・フラボノイド生合成のための主要なパスウェイに関与する酵素、Gmax275 とエンレイの対応する遺伝子
図 2-3	ダイズ８番染色体の CHS 遺伝子クラスタの位置を示す領域
図 3-1	KAIKObase のフローチャート
図 3-2	PGmap と UnifiedMap の通信
図 3-3	ブラウザ、ビューア、独立したデータベース間のリンク
補足図 3-1	GAL4-UAS によるカイコのエンハンサトラップを同定するための交配スキーム

表リスト

表番号	タイトル
表 1-1	ゲノム解読された主な生物
表 1-2	第一世代、第二世代、第三世代シークエンサの比較
表 1-3	DNA マーカーのいくつかの例
表 2-1	*De novo* アセンブリとレファランスマッピングで使用した配列
表 2-2	エンレイゲノムアセンブルと遺伝子アノテーション
表 2-3	エンレイゲノムにマップされた数・割合
表 2-4	エンレイゲノムの一塩基多型と挿入・欠失
表 2-5	若葉から抽出した cDNA の配列とアセンブル
表 2-6	エンレイにおける貯蔵タンパクおよび cupin 成分
補足表 2-1	連鎖距離順が一致しないマーカーと物理位置
別冊表 2-2	レファランスマッピングに使用したエンレイの DNA マーカーと物理位置
別冊表 2-3	系統解析で使用したフィルタされたシングルコピー遺伝子
別冊表 2-4	子葉タンパクデータに対応する Gmax275 とエンレイの遺伝子モデル
別冊表 3-1	ライブラリ由来のカイコ cDNA ライブラリと EST のアクセッション番号

略語リスト

略語	英語名	日本語名
ANS	anthocyanidin synthase	アントシアニジン合成酵素
BAC	bacterial artificial chromosome	BAC
BAC end	BAC end	BAC エンド
BES	BAC end sequence	BAC エンド配列
cDNA	complementary DNA	cDNA
CHI	Chalcone isomerase	カルコンイソメラーゼ
CHS	Chalcone synthase	カルコン合成酵素
CSP	chemosensory protein	感覚子タンパク質
DFR	dihydroflavonol 4-reductase	ジヒドロフラボノール 4-レダクターゼ
DNA	deoxyribonucleic acid	デオキシリボ核酸
EGFP	Enhanced GFP	EGFP
emPAI	exponentially modified protein abundance index	emPAI
EST	expressed sequence tag	EST
F3H	flavanone 3-hydroxylase	フラボノイド 3-ヒドロキシラーゼ
FL-cDNA	full-length cDNA	完全長 cDNA
FLS	flavonol synthase	フラボノール合成酵素
fosmid end	fosmid end	fosmid エンド
FPC	fingerprint of contigs	FPC
GFP	green fluorescent protein	GFP
GPCR	G protein-coupled receptor	G タンパク質共役受容体
HMM	Hidden Markov Model	隠れマルコフモデル
INDEL	Insertion Deletion	インデル（挿入・欠失）
Inverse PCR	Inverse PCR	インバース PCR
LEA	Late embryogenesis abundant protein	LEA
mol%	mol%	モル百分率

MP	mate-pair	メイトペアー
mRNA	messenger RNA	伝令 RNA
MS	Mass Spectrum	質量スペクトル
ncRNA	non-coding RNA	ノンコーディング RNA
NGS	Next Generation Sequencer	次世代シークエンサ
OBP	odorant-binding protein	匂い物質結合タンパク質
OLC	overlap-layout-consensus	オーバラップレイアウトコンセンサス
ORF	Open Reading Frame	オープンリーディングフレーム
PCR	Polymerase Chain Reaction	ポリメラーゼ連鎖反応
PE	paired-end	ペアーエンド
QV	quality value	シークエンスクオリティスコア
RNA	ribonucleic acid	リボ核酸
RNAseq	RNA sequencing	RNA シークエンシング
primer	primer	プライマ
RT-PCR	Reverse Transcription PCR	逆転写ポリメラーゼ連鎖反応
SE	single-end	シングルエンド
SNP	Single Nucleotide Polymorphism	一塩基多型
tRNA	transfer RNA	転移 RNA
UAS	upstream activation sequence	UAS 配列（遺伝子）
UniProt	The Universal Protein Resource	UniProt
WGD	Whole Genome Duplication	全ゲノム重複
95PD	95% probability density	95%の確率密度

第1章　はじめに

1.1.　ゲノム研究の動向

「形質は遺伝する」というメンデルが発見した法則(1865 年)は、遺伝子という概念の基礎となった[1]。その後、モーガンらによるショウジョウバエを使って、遺伝子が染色体上にあること(1913 年)が示され[1]、ワトソンとクリックが、DNA が相補性を持つ二重螺旋構造であることを発見(1953 年)した[1]。DNA の二重螺旋構造発見以降、ゲノム DNA が細胞でどのように作用するかの研究が進み、mRNA、コドンの発見から遺伝子発現の基礎的な仕組み[1]、ヒストンのメチル化、アセチル化、リン酸化が発現に及ぼす抑制や活性化[2]、トランスポゾンや ncRNA などによる発現抑制[3]、組織で異なったゲノム構造の空間的変化[4]などがわかってきた。

ゲノム解析に着目すると、既知のガン遺伝子の変異を調べる上で、個々の遺伝子に着目した研究が行われてきたが、ダルベッコは、ヒトゲノムの塩基配列を全部決定するのがブレークスルーに繋がると考え、ゲノム配列の重要性、これを実現するための国家的な予算支援や国際協調による作業、および解析時間短縮のための技術開発を提言(1986 年)した[1,5,6]。このことが契機となり、全ゲノム配列獲得の実現に向けての研究が開始された[6]。

ゲノム解析における概略フローは図 1-1 のとおりで、ゲノム配列を小さい断片に分け、配列を解読し、その配列を組み上げて行く方法が作られ、ヒトゲノム解読に先立ち 1995 年に、最初のゲノム解析として、インフルエンザ菌のゲノム解読[7]が報告された。

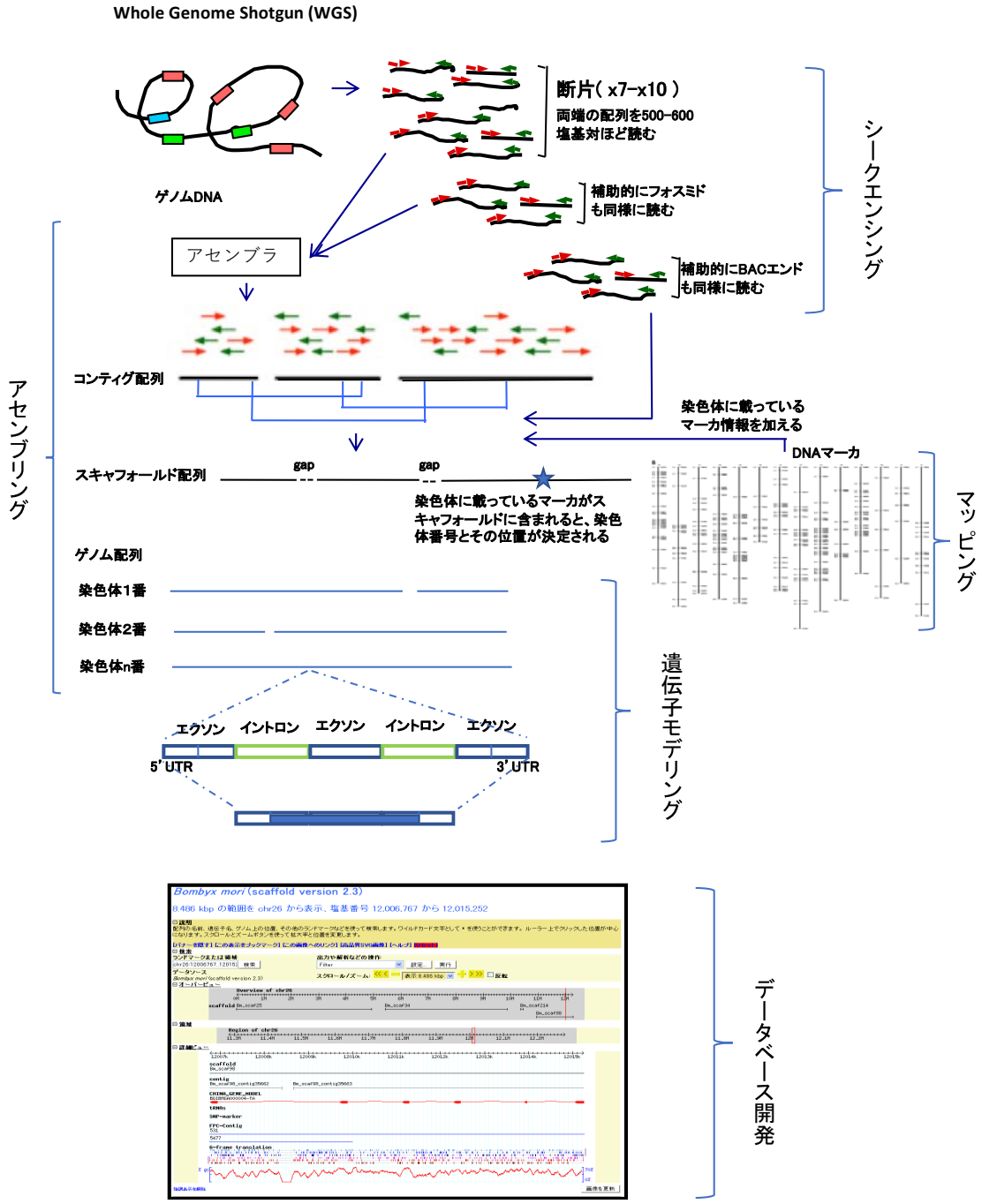


図 1-1 ゲノム解析における概略フロー

以降、ゲノム解読の歴史を示したのち、ゲノム解析における概略フローに沿って、シークエンシング、アセンブリング、マッピング、遺伝子モデリング、データベース開発について述べる。

1. 1. 1.　ゲノム解読の歴史

ゲノム解析は、1995 年にインフルエンザ菌[7]、マイコプラズマ菌[8]のゲノム解読が行われ、それ以降、様々な生物のゲノムが解読された。表 1-1 に全ゲノムが解読された主な生物を示す。これら生物の中でも、モデル生物と呼ばれるものがある。モデル生物は、生物学的現象の範囲を理解するために広範に研究されているヒト以外の生物であり、遺伝的ツールとしてのそれらの能力に密接に関連した特定の実験的特徴（短いゲノムサイズ、短い世代交代、高い出生率、容易な突然変異系統の獲得、遺伝子改変の容易性）を有する。代表例は、大腸菌、出芽酵母、ショウジョウバエ、線虫、マウス、シロイヌナズナなどである[9]。

表 1-1 ゲノム解読された主な生物

論文公開年	分類	和名	学名	参考文献
1995	Bacteria（真正細菌）	インフルエンザ菌	*Haemophilus influenzae*	[7]
1995	Bacteria（真正細菌）	マイコプラズマ・ジェニタリウム	*Mycoplasma genitalium*	[8]
1997	Bacteria（真正細菌）	大腸菌	*Escherichia coli*	[10]
1997	Fungi/Ascomycota/Saccharomycetales（サッカロミケス目）	出芽酵母	*Saccharomyces cerevisiae*	[11]
1998	Animalia/Nematoda/Rhabditidae（桿線虫目）	カエノラブディティス・エレガンス	*Caenorhabditis elegans*	[12]
2004	Animalia/Chordata/Primates（サル目）	ホモ・サピエンス	*Homo sapiens*	[13]
2002	Animalia/Chordata/Rodentia（ネズミ目）	ハツカネズミ	*Mus musculus*	[14]
2000	Animalia / Insecta / Diptera（双翅目）	ショウジョウバエ	*Drosophila melanogaster*	[15]
2002	Animalia / Insecta / Diptera（双翅目）	ハマダラカ	*Anopheles gambiae*	[16]
2006	Animalia / Insecta / Hymenoptera（膜翅目）	ミツバチ	*Apis mellifera*	[17]
2008	Animalia / Insecta / Coleoptera（鞘翅目）	コクヌストモドキ	*Tribolium castaneum*	[18]
2008	Animalia / Insecta / Lepidoptera（鱗翅目）	カイコ	*Bombyx mori*	[19]

2000	Plantae/Brassicales（アブラナ目）	シロイヌナズナ	*Arabidopsis thariana*	[20]
2005	Plantae/Poales（イネ目）	イネ	*Oryza sativa*	[21]
2006	Plantae/Malpigiales（キントラノオ目）	ポプラ	*Populus trichocarpa*	[22]
2007	Plantae/Rhamnales（クロウメモドキ目）	ヨーロッパブドウ	*Vitis vinifera*	[23]
2008	Plantae/Fabales（マメ目）	ミヤコグサ	*Lotus japonicus*	[24]
2008	Plantae/Brassicales（アブラナ目）	組換えパパイヤ	*Carica papaya*	[25]
2009	Plantae/Poales（イネ目）	モロコシ	*Sorghum bicolor*	[26]
2009	Plantae/Poales（イネ目）	トウモロコシ	*Zea mays*	[27]
2009	Plantae/Cucurbitales（ウリ目）	キュウリ	*Cucumis sativus*	[28]
2010	Plantae/Fabales（マメ目）	ダイズ	*Glycine max*	[29]
2010	Plantae/Malpigiales（キントラノオ目）	トウゴマ	*Ricinus communis*	[30]
2010	Plantae/Rosales（バラ目）	リンゴ	*Malus domestica*	[31]
2011	Plantae/Brassicales（フウチョウソウ目）	ハクサイ	*Brassica rapa*	[32]
2011	Plantae/Fabales（マメ目）	キマメ	*Cajanus cajan*	[33]
2011	Plantae/Rosales（バラ目）	イチゴ	*Fragaria vesca*	[34]
2011	Plantae/Malvales（アオイ目）	カカオ	*Theobroma cacao*	[35]
2011	Plantae/Fabales（マメ目）	タルウマゴヤシ	*Medicago truncatula*	[36]
2011	Plantae/Arecales（ヤシ目）	パームヤシ	*Phoenix dactylifera*	[37]

2011	Plantae/Solanales（ナス目）	ジャガイモ	*Solanum tuberosum*	[38]
2012	Plantae/Solanales（ナス目）	トマト	*Solanum lycopersicum*	[39]
2012	Plantae/Cucurbitales（ウリ目）	メロン	*Cucumis melo*	[40]
2012	Plantae/Zingiberales（ショウガ目）	バナナ	*Musa acuminata*	[41]
2013	Plantae/Rosales（バラ目）	モモ	*Prunus persica*	[42]
2014/2015	Plantae/Brassicales（フウチョウソウ目）	ダイコン	*Raphnus sativus*	[43, 44]
2015	Plantae/Fabales（マメ目）	ダイズ	*Glycine max* cv. enrei	[45]

1.1.2.　シークエンシング

ゲノム解析を支える配列解読技術の発展は、1977 年に発表されたサンガー法およびマクサム・ギルバート法に基づく DNA シークエンシング[46,47]や、80 年代に発展したポリメラーゼ連鎖反応（PCR）法[48]がベースにある。PCR は塩基配列情報さえあれば、プライマを設計し、使用することで、目的領域の DNA 断片を簡単に増幅することが可能となり、クローニングやシークエンスで利用できる。さらに、増幅された配列をシークエンサにかけ、シークエンシングで塩基配列を得ることができる。

サンガー法シークエンサは第一世代に分類され、ラジオアイソトープではなく蛍光試薬を用いた DNA シークエンサが登場した。1986 年の ABI370[49]を皮切りに、ガラス板型の電気泳動を用いた ABI377 その後、ガラスキャピラリを用いた ABI3700、ABI3730 が開発され、より長い配列を高精度で解読する DNA シークエンサが開発された。NGS に分類される第二世代は、Roche 社(旧 454 Life Sciences 社）の Pyrosequence 法によるシークエンサ、Illumina 社（旧 Solexa 社）による Sequence by Synthesis 法によるシークエンサ、Life Tech 社による Sequence by Ligation 法によるシークエンサが 2005 年から 2007 年にかけ開発・販売された。NGS に分類される第三世代は、Pacific Biosciences 社から平均 954bp で 2000bp 以上の配列を 5%含むリードを持つ PacBio RS[50]、Oxford Nanopore 社から平均リード長 5Kbp を持つ MinION sequencer[51]が開発された。現在まで、より長い DNA 断片を解読できるように進化している。一方、数 Mbp 以上の DNA 断片を高精度に一気にロングリードできるシークエンサは未だ現れていない。第一世代、第二世代、第三世代シークエンサの比較を表 1-2 に示す[52].

表 1-2 第一世代、第二世代、第三世代シークエンサの比較

	第 1 世代	第 2 世代	第 3 世代
基本技術	DNA ポリメラーゼによる DNA 合成でシークエンス、または分解によって生成された特異的末端標識 DNA 断片のサイズ分離	DNA 合成のリアルタイム検出、余剰塩基等の洗い流し、再反応の繰り返しによるシークエンス	DNA 分子の直接的な物理的検査（DNA ポリメラーゼによる合成のモニタリング、または 1 本鎖 DNA のナノポアの通過による塩基検出）
読取り精度	高	高	中
リード長	中（800-1,000bp）	短(33-150bp)	長（1,000bp 以上）
スループット	低	高	中
コスト	高/塩基 低/稼働	低/塩基 高/稼働	低・中/塩基 低/稼働
RNAseq 方法	cDNA	cDNA	RNA/cDNA
シークエンスを得るまでの時間	時間	日	時間
サンプル調製	中程度に複雑、PCR 増幅は不要の場合あり	複雑、PCR 増幅は要	シークエンサに応じて複雑なものから非常に単純なものまで
データ解析	検出した信号のベースコール。ルーチン	大量の画像処理によるベースコール。複雑、データ量大、ショートリードはアセンブリや	DNA 合成の映像からベースコール、またはナノポア通過時の電位変化からベースコール。複雑、データ量大、新し

		アライメントのアルゴリズムが複雑	いタイプの情報と新しい信号処理による課題
出力	QV を持った読出し	QV を持った読出し	QV、カイネティックスなどの他の塩基情報を持つ読出し

このような配列からゲノムを構築するために、シークエンスされた配列を再構築し、ゲノムに仕立て上げる方法（アセンブル）が開発された。

1.1.3. アセンブリング

シークエンサから出力される塩基配列長は、数十ベースから数百ベース、長い配列では、10 キロベースを超えるものもあるが、ゲノム全体の中の断片であることが多い。また、これらの配列には精度情報が付加される。これらの情報をもとに、全体を再構築する作業が、アセンブルである。アセンブルは、*De novo* アセンブルとレファランスマッピングの２種に大別される。*De novo* アセンブルは、新規生物の配列をシークエンスされた配列から構築する方法である。もう一方のレファランスマッピングは、参照配列が既存であり、そこに、シークエンスされた配列をマップして、新しい配列を生成する方法である。

De novo アセンブラの初期に開発されたアセンブルソフトウェア Phrap[53]、TIGR[54]は、すでに構築されたアセンブリと矛盾しない限り、最も重複する読み取りに常に結合する Greedy と呼ばれるアルゴリズム[55]を使用している。アセンブルソフトウェア Celera[56]、ARACHNE[57]は、十分によくオーバラップする全ての読み取りペアーを特定し、互いにオーバラップする読み取りペアー間でグラフを構成する。このグラフ構造は読み取り間のグローバルな関係を考慮に入れることができ、複雑なアセンブルアルゴリズム開発を可能とするオーバラップレイアウトコンセンサス（OLC）と呼ばれるアルゴリズム[55]を使用している。2004 年に Roche のシークエンサ対応で、GS *De novo* Assembler（Newbler）[58]、2007 年に Illumina シークエンサ対応で、リードから抽出された長さ k の正確な部分文字列間の関係をモデル化する De Bruijn graph と呼ばれるアルゴリズム[55]を使用した Velvet[59]、2009 年に

ABySS[60]、2010 年に SOAPdenovo[61]が開発された。これらの多くは異なる種類のシークエンサ出力（配列）がアセンブルできるハイブリッドアセンブラ[55]に発展している。

De novo アセンブルのステージは、前述したシークエンスからコンティグを作るステージ、コンティグをつなぎ合わせて、染色体配列に近づけるステージ（スキャフォールディング）がある。スキャフォールディングでは、Illumina メイトペアー(インサート長 10-20Kb)、fosmid エンド(インサート長 40Kb)や、BAC エンド(インサート長 100Kb)のメイトペアー配列が利用され、これらのメイトペアー配列でコンティグをホッチキス止めしていくようなイメージである。スキャフォールディングするためのソフトウェアは、2011 年に開発された SSPACE[62]などがある。

De novo アセンブリとは別に、レファランスアセンブリは、近縁種のゲノム配列がすでに存在している場合、シークエンサから出力されるショートリード配列を BWA[63]などのソフトウェアを使用して、レファランスにするゲノムにマップして、ゲノムに沿って１塩基ずつ比較を行い、ターゲットの配列を決定するという方法である。

1.1.4. マッピング

マーカーを染色体上に並べていったものは、連鎖地図、もしくは遺伝的連鎖地図（リンケージマップ）と呼ばれる。ゲノム構築上、マーカーは、例えば、コンティグやスキャフォールドに DNA マーカーが１つ乗っていれば、そのコンティグまたはスキャフォールドがどの染色体に含まれるかわかる。また、コンティグやスキャフォールドに２つ以上の DNA マーカーが乗っていれば、染色体上でのコンティグやスキャフォールドの方向が決定できる。この方法はゲノム配列のアセンブルの際に利用される。

遺伝マーカーとして使用される DNA マーカーは、個体または種を同定するために使用することができ、染色体上の既知の位置を有する遺伝子または DNA 配列をもとに作られたものである。表 1-3 に DNA マーカーのいくつかの例、制

限断片長多型(RFLP)［64］、無作為増幅多型（RAPD）［65］、一塩基多型（SNP）［66］、単純配列反復（SSR）［67,68］、増幅断片長多型（AFLP）［69］、単純配列反復と EST を組み合わせた（EST-SSR）［70］、SNP-BAC エンド（SNP-BACend）［71,72］を示す。

表 1-3 DNA マーカーのいくつかの例

マーカー名称	和名	英名	説明	参考文献
RFLP マーカー	制限酵素断片長多型	Restriction Fragment Length Polymorphism	Y.W.Kan らにより 1978 年に発表された。ある特定の DNA 領域について、制限酵素識別部位の塩基置換や認識部位に挟まれた部位での欠失や挿入があると、制限酵素によって切断された断片のサイズに違い(多型) が現れるので、電気泳動で区別することが可能となり、これをマーカーとする。	［64］
RAPD マーカー	無作為増幅多型	Random Amplified Polymorphic DNA	J.G.Williams らにより 1990 年に発表された。ゲノム DNA を鋳型として、無作為に合成したプライマを用いた PCR によって増幅したとき、DNA の塩基配列に違いがあるとプライマの結合に差異がでるため、増幅された DNA 短編のサイズや数に違いが現れ、これをマーカーとする。	［65］
SNP マーカー	一塩基多型	Single Nucleotide Polymorphism	SNP は 1991 年の Ligtenberg の論文で使用された。ある特定 DNA 領域の塩基配列を比較することにより、一塩基の違いを見つけ、PCR などを利用して検出し、マーカーとする。検出方法には、いくつかの方法が知られ、より効率・低コスト化を目指して開発が進んでいる[73]。また、遺伝学的手法を用いた解析では DNA マーカーは遺伝子座との連関を示すことが可能となり、形	［66］

			質を調べる代わりに DNA マーカーを用いることで、早期の育種選抜が可能となる[74]。	
SSR マーカー	単純配列反復	Simple Sequence Repeat	縦列型反復配列（short tandem repeat: STR）、マイクロサテライト（microsatellite）とも呼ばれる。これらのマーカーは 1994 年、Zietkiewicz ら、A.Utquhart らにより発表された。2 から 4 塩基を単位とした縦列反復は、ゲノム上多数見られ、この反復単位の繰り返し回数に違いが見られることがある。この違いを PCR により増幅、電気泳動等でサイズの差を検出し、これをマーカーとする。	[67, 68]
AFLP マーカー	増幅断片長多型	Amplified Fragment Length Polymorphism	1995 年に Pieter Vos らにより発表された。制限酵素によって切断した DNA 断片を PCR で増幅することにより、RFLP の場合のように違いが現れた場合、それをマーカーとする。	[69]
EST-SSR マーカー	EST-単純配列反復	EST-Simple Sequence Repeat	EST と SSR を組み合わせたマーカーで、2002 年 Eujayl らにより発表された。EST-SSR マーカー作成では、自殖系などが使用され、EST から得られた SSR 配列の解析とユニークで非冗長な EST から連鎖解析で、連鎖群上に EST-SSR マーカーが構築される。	[70]
EST-BACend マーカー	一塩基多型—BAC エンド	SNP-BACend	SNP と BAC エンドを組み合わせたマーカーで、2004 年に Weil らにより発表された。SNP マーカーでは、ゲノム中に豊富にあるため、構築が簡単である。BAC エンド配列を用いた SNP マーカー作成では、バッククロスなどにより系統が作られ、さらに BAC エンド配列に含まれる SNP を見つけるた	[71, 72]

	めに PCR アンプリコンが使用され、連鎖解析で連鎖群上に SNP-BAC エンドマーカーが構築される。SSR を用いた SNP マーカーでは、バッククロスなどが使用され、SSR に含まれる SNP で、連鎖解析が行われ、連鎖群上に SNP-SSR マーカーが構築される

1.1.5. 遺伝子モデリング

ゲノム上に存在する遺伝子（機能性タンパク質または RNA 分子の合成に必要な全 DNA 配列[75]）の配列位置の推定を行う。遺伝子配列推定では、計算による遺伝子モデル予測、mRNA 配列、もしくは遺伝的特徴を含む様々なソースを使用し、遺伝子モデル[76]を作成する。

ゲノム構築後、遺伝子モデルにアノテーションを付与するために使用される遺伝子予測プログラムは、その多くが ORF を予測するプログラムである。得られた遺伝子モデルから得られたプライマを用い、シークエンスや逆転写ポリメラーゼ連鎖反応（RT-PCR）を使用することにより実際の配列を得ることができる[77]。遺伝子モデルを得る方法には、EST、完全長 cDNA や遺伝子とゲノムの相同性を利用し推定する方法、RNAseq をゲノムにマップし遺伝子モデルを推定する方法、ゲノム配列から *De novo* 遺伝子予測プログラムを使用し遺伝子モデルを作成する方法、RNAseq 配列をアセンブルし遺伝子モデルを作成する方法などがある。

EST、完全長 cDNA や遺伝子とゲノムの相同性を利用し推定する方法は、BLAST[78]、Smith-Waterman[79]などの相同性検索ソフトウェアなどを使用して、ゲノム上の領域を限定し遺伝子モデルを構築する。RNAseq をゲノムにマップし遺伝子モデルを推定する方法は、ゲノム上へマップするための Bowtie[80]、スプライスジャンクションを予測するための TopHat[81,82]、遺伝子構造予測するための Cufflinks[83]を用い、遺伝子モデルを構築する。ゲノム配列から *De novo* 遺伝子予測プログラムを使用し遺伝子モデルを作成する方法は、隠れマルコフモデル（HMM）のいくつかの変種に基づいて作成されている[77]。Genscan[84]、

Fgenesh[85]、Augustus[86]などのソフトウェアがあり、遺伝子モデルを構築することができる。RNAseq 配列をアセンブルし遺伝子モデルを作成する方法は、RNAseq でコンティグを作成する Inchworm プロセス、コンティグをクラスタリングし、de Bruijn グラフを作成する Chrysalis プロセス、de Bruijn グラフのコンポーネントからすべての可能性のあるシークエンスを抽出する Butterfly プロセスを持った Trinity により RNAseq から遺伝子モデルを構築する[87,88]。

1.1.6. データベース開発

解析されたゲノム情報を公開する上で、インターネットで利用できる様々な道具立てが作り出されてきた。その中でも、1995 年に *C. elegans* の AceDB[89]が遺伝的地図と物理的地図をもったゲノム情報を表示する道具立てのパイオニアである。この遺伝的地図と物理地図を表示する方法は、2000 年に INE: INtegrated rice genome Explorer[90]、2002 年に NCBI map viewer[91,92]、2009 年に Cmap[93]などで見ることができる。また、これとは別に、メガベース単位のゲノム情報を表示する道具立てとして、2002 年に Ensembl genome browser (Ensembl contigview)[94,95]、UCSC browser[96]、GBrowse[97]、2008 年に UTGB[98]などが開発された。1) NCBI map viewer は NCBI リソースの中の一つで、ゲノムアノテーションの簡単なテキストベースの検索を実行して遺伝子のゲノムテキストの表示、染色体に沿って移動、ズームイン/ズームアウト、表示されたマップを表示/非表示の切り替えができる機能を持つ。Map viewer は BLAST などの NCBI の塩基解析ツールにリンクされている[91,92]。NCBI map viewer は 2017 年 NCBI genome data viewer(GDV)にアップデートされた[99]。2) Ensembl genome browser (Ensembl contigview) は大規模なゲノムの配列を中心とする生物学を構成するためのバイオインフォマティックスのフレームワークを提供する Ensembl データベースプロジェクト[94]から生み出された。このプロジェクトはシークエンス解析からデータの保存や可視化まで関連する要件を処理できるポータルシステムを開発するオープンソースの開発プロジェクトである。Ensembl サイトはヒトゲノム配列のアノテーションを供給する主要なサイトの一つであり、国際的なヒトゲノムプロジェクトによる分析の多くを提供した。Ensembl プロジェクトはこのヒトゲノムアノテーションのデータベースが提供されているため、マウス、ラット、ゼブラフィッシュなどの脊椎動物ゲ

ノム配列を利用した比較ゲノム閲覧システムとしての使用が可能である[94,95]。3) UCSC browser は、ヒトゲノム用のゲノムブラウザとして開発され、その後脊椎動物やモデル生物用のゲノムブラウザとして利用されている。UCSC browser の特徴は、アノテーションの豊富さ、速度、安定性、拡張性、ユーザインタフェースの一貫性である[96]。4) GBrowse はショウジョウバエのゲノム配列のブラウザとして使用された。ゲノムの任意の領域をスクロールやズームする機能、ランドマークの検索、全文検索でゲノムの領域に入る機能、トラックを有効/無効にする機能、相対的な順序と外観を変更する機能などがある。ブラウザソフトウェア機能には、容易に利用できるオープンソースコンポーネント、簡単なインストール、柔軟な構成などがある[97]。5) UTGB（東京大学ゲノムブラウザ）は、日本のメダカのために開発された[98]。最小限の労力で簡単にシステムをインストールし、ローカルに保存されたデータをブラウスし、個々のニーズに合わせた Web インタフェースで迅速なインタラクティブな設計を満たすように設計されている[100]。

配列検索では、BLAST[78]、BLAT[101]、Smith-Waterman[79]などのツール、二次元電気泳動の結果を表示する ExPASy の The Make2D-DB II Package[102]などが開発された。また、2005 年に、各ユーザからのクエリーの格納、データの共通項/和/差などの操作の実行、他の計算ツールへのリンクなどの機能を持つ柔軟な履歴システム Galaxy[103]が開発された。Galaxy では UCSC browser がゲノムブラウザとして使用されている。

1.2. 農業生物ゲノム研究の課題

1.2.1. ゲノム構築・解析

遺伝子を網羅的に獲得し、機能解析をする上で、ゲノム構築は有効な方法であり[5]、1.1.1 節ゲノム解読の歴史に示したように、多種多様な生物でゲノム構築・解析が実施されている。農業生物、特に作物に関するゲノム構築・解析は、イネ、ブドウ、組換えパパイヤ、ソルガム、トウモロコシ、ダイズ、キュウリ、ハクサイ、トウゴマ、リンゴ、キマメ、イチゴ、カカオ、タルウマゴヤシ、パームヤシ、ジャガイモ、モモ、トマト、メロン、バナナ、ダイコンなどがある。

世界的な生産量では、ダイズは、イネ、小麦、トウモロコシからなる 3 大主要穀物の次に位置付けられており、食用タンパク質と植物油の主要な供給源で、世界で最も重要なマメ科作物の一つである。ダイズゲノムは、2010 年に米国の努力により、栽培品種である Williams 82 品種で構築された[29]。大まかな遺伝子獲得や機能解析であれば、Williams 82 ゲノムの使用で十分である。しかし、日本のダイズ育種では品種間固有の DNA マーカーを使用した育種（元来、表現系を指標に、その形質は染色体上の 1 箇所に起因するものとして遺伝解析をしていたものを、染色体上の直接の印「DNA マーカー」を使って遺伝解析を行い、目的の形質を集積した品種を作り上げる育種法。）が行われており、日本での育種は日本産品種同士の掛け合わせになることが多い。Williams 82 ゲノムと日本ダイズは同じダイズ種ではあるが、系統が離れているため[104]、Williams 82 から得られた DNA マーカーが使用できないケースがある。このため、国産ダイズ品種エンレイのゲノム構築・解析が必要となった。

1.2.2. データベース開発

ゲノム構築・解析を実施した場合、それらのデータを管理・閲覧するためのデータベースが必要となる。1.1.6 節で示したように、ゲノム閲覧のためのブラウザ（NCBI Map viewer、Ensembl genome browser、UCSC browser）は、ヒトを含む脊椎動物やモデル生物に焦点が当てられ、広範囲なデータリソースで構築されている。また、様々な解析ツールのワークフローを構築した後、繰り返し操作を簡便にする Galaxy、そのゲノムブラウザには UCSC browser が使用されている。Ensembl も広範囲なデータリソースで構築されている[105]。WormBase[106]や FlyBase[107]では閲覧システムに GBrowse が使用されている。カイコゲノムは、BAC エンド配列解析で構築された高密度 SNP 遺伝地図と FPC プログラムを使用した BAC フィンガプリンティングマップ[108]、様々な組織や異なる発育段階から得られた EST データが集められた SilkBase[109]、様々な組織や異なる発育段階のプロテオームデータベース[110]、レポータ発現パターンおよび遺伝子トラップ系統やエンハンサトラップ系統のミューテータの挿入された位置を提供するための *Bombyx* trap データベースなどが、カイコゲノムプロジェクト内の個

別研究やこれと並行した個別研究で作成されており、これらを効率的に統合するためのGBrowseを中核とした独自のデータベースが必要となった。

1.3. 研究目標と方法

本論文は、ゲノム構築・解析とそれらデータの閲覧システム（データベース）に亘る一連の流れに沿って、国産ダイズ品種エンレイゲノムの構築・解析[45]とカイコゲノム構築・解析から得られた情報を統合するためのデータベース開発[111]を目標とする。前者はゲノム構築・解析に力点を置き、後者は閲覧システムとその統合（データベース）に力点を置いている。

国産ダイズ品種エンレイゲノムの構築・解析では、エンレイの葉から核を調製し、DNA抽出、シークエンス、アセンブル、マーカー情報をもとにスキャフォールド、コンディグをゲノムへ整列させたゲノム（G.max_Enrei1）の構築、Williams 82ゲノムへのレファランスマッピングでゲノム構築を実施し、マーカー情報と*De novo*アセンブルから作成されたG.max_Enrei1で、レファランスマッピングで得られたゲノムを再構築する。得られたゲノム（G.max_Enrei2）から遺伝子モデルを作成する。解析において、アントシアニン・フラボノイド生合成系で、ゲノム上にある遺伝子を明確化する。加えて、プロテオーム解析の有用性を示すため、登熟期ダイズ種子の子葉部分のプロテオーム解析を実施し、ダイズで重要である貯蔵タンパクがどの染色体に座乗しているかを明確にする。さらに、RNAseqをアセンブルし、遺伝子モデルを作成する。作成された遺伝子モデルとゲノム配列から作成した遺伝子モデルの共通遺伝子モデルとWilliams 82の遺伝子モデル、シロイヌナズナの遺伝子モデル、シロイヌナズナの一種のミヤマハタザオ等の遺伝子モデルを用い、系統解析を実施する。

カイコゲノム構築・解析から得られた情報を統合するための閲覧システム・データベース構築では、カイコゲノムプロジェクトや個別研究で得られたゲノム情報（スキャフォールド、BAC、BACエンド配列、Fosmidエンド配列を含む）、DNAマーカー情報、遺伝子モデル情報、組織別・発育段階別のトランスクリプトーム情報、組織別・発育段階別のプロテオーム情報、レポータ発現パ

ターンおよび遺伝子トラップ系統やエンハンサトラップ系統のミューテータの挿入された位置情報をカイコゲノム統合データベース KAIKObase に統合する。KAIKObase では、ゲノム配列とマーカー情報を俯瞰するための遺伝地図と物理地図を併せ持つ PGmap、中程度の遺伝地図と物理地図を併せ持つ UnifiedMap、詳細なゲノム情報、遺伝子モデル情報、マーカー情報を閲覧できる GBrowse や UTGB、遺伝子モデルを閲覧するための GeneViewer、組織別・発育段階別のトランスクリプトーム情報を閲覧できるデータベース、組織別・発育段階別のプロテオーム情報を閲覧できるデータベース、レポータ発現パターンおよび遺伝子トラップ系統やエンハンサトラップ系統のミューテータの挿入された位置情報を閲覧できるデータベース、配列検索を行う BLAST サーチ、キーワードサーチを遺伝子モデル名や塩基配列情報などで統合する。

1.4. 本文の構成

第 1 章では、ゲノム研究の動向として、ゲノムは何故必要になったか、どのようなゲノムがいつ頃解読されたか、ゲノム解析を支える塩基配列解読技術はどのように進歩したのか、ゲノムアセンブルなどで利用できる DNA マーカーの種類、ゲノムを構築するためのアセンブル技術の種類、得られたゲノムや RNAseq から遺伝子モデルを構築するための方法、ゲノム情報を閲覧さえるための閲覧システムを示す。次に、ゲノム構築する上での課題として、ゲノム構築・解析とデータベース開発を示した後、研究目標と方法、本書の構成へと続く。

第 2 章では、国産ダイズ品種エンレイゲノムの構築・解析を示す。材料と方法として、ゲノムシークエンシング、アセンブルとレファランスマッピング、遺伝子モデル、系統解析、アントシアニン・フラボノイド生合成系、プロテオーム解析を示す。結果と考察として、ゲノムシークエンシングとレファランスマッピング、一塩基多型、挿入・欠失、遺伝子モデル、系統解析、アントシアニン・フラボノイド生合成系、子葉におけるタンパク質、エンレイゲノムデータベースを示し、最後に結論を示す。

第 3 章では、カイコゲノム構築・解析から得られた情報を統合するための閲覧システム・データベース開発を示す。データセット内容として、ゲノム配列情

報、ゲノム配列にマップされる情報、プロテオーム情報、エンハンサトラップ情報を示す。データベース構成として、KAIKObase に含まれるプロテオームデータベース、*Bombyx* trap データベース、配列検索システム、キーワード検索システムを示す。さらに、遺伝地図と物理地図のビューア（PGmap、UnifiedMap、UTGB、GBrowse）、遺伝子モデル情報を表示する GeneViewer、KAIKObase で使用しているソフトウェアを示す。使用方法と考察として、ユーザインタフェース、キーワードサーチ、ポジションサーチ、シークエンスサーチを示し、最後に、結論を示す。

第 4 章では、結言として本研究の成果について纏める。

第 2 章　国産ダイズゲノム構築・解析

2.1.　概要

ダイズ（*Glycine max*）の栽培品種エンレイゲノムを解明した。これは日本のダイズ栽培品種の特性評価のための参考情報を提供することができる。次世代シークエンサを用いて得られた全ゲノム配列を栽培品種 Williams 82 ゲノムにレファランスマッピングして、エンレイゲノムを決定した。決定されたゲノムは約 928Mb の塩基と 60,838 の遺伝子モデルを有するデータセットが得られた。系統解析では、エンレイおよび Williams 82 品種双方の系統関係、および野生ダイズを含む複合体からのそれらの相違に一瞥を与えた。遺伝子モデルは、アントシアニン・フラボノイド生合成に関連する形質およびプロテオームの全体的なプロファイルに関連して分析した。配列データは、DAIZUbase で利用可能となり、日本の広範なダイズ品種の比較ゲノミクスの包括的な情報資源と、国内外のダイズ品種の改良のための有効な参考情報となる。

2.2.　はじめに

ダイズ *Glycine max* は、食用タンパク質と植物油の主要な供給源として世界で最も重要なマメ科作物の一つである。世界的な生産量では、ダイズは、米、小麦、トウモロコシなどの主要穀物の次に位置付けられる。また、サポニン、イソフラボン、ファイトステロール、およびトコフェロールなどの生理活性物質の主要な供給源である。食品としてのダイズの消費量は、主にアジア地域に集中している。ダイズは日本人の食習慣の一部となっており、発酵食品である味噌、醤油、納豆などや発酵していない食品である枝豆、きな粉、豆乳、豆腐などのダイズやダイズ加工食品が古来より食される。他の主要作物のように、日本におけるダイズ育種の主なターゲットは、輸入ダイズに打ち勝つための高い収量、高い品質（種皮亀裂がないこと、へその色や種子の大きさの均一性、および食品加工適性）であり、安定生産のための生物的/非生物的ストレスに対する抵抗性がある。加えて、タンパク質が多いこと、貯蔵タンパク質の修飾、リポキシゲナーゼおよびサポニンが無いこと、イソフラボンが多いこと、およ

びスクロースが多いことなどの種子の化学成分について、多くのダイズ育種プログラムで検討されている[112]。

ツルマメ、つまり野生ダイズ種は、栽培ダイズの祖先であり、中国北部、日本、韓国、ロシアの東部で見つかっている[113]。考古学の研究において、ダイズという単語が、約 3,700 年前の中国の骨碑文に最初に登場し、約 2,600 年前の殷王朝の遺物から、炭化ダイズ種子が発見された[113]。考古学的な推定は、9,000-8,600 年前の中国北部で、7,000 年前の日本で、小粒ダイズの初期の広がりが示された[114]。炭化ダイズ種子の放射性炭素年代測定では、大粒ダイズ選択が 5,000 年前日本で、3,000 年前の韓国であったことが示された[114]。大規模なゲノム解析によるダイズの祖先と野生種ダイズの分化年代は、0.27Mya[115]もしくは 0.8Mya[116]と推定された。最近の研究で、在来種、外来種、栽培種、および野生種ダイズの 1,603 種間の遺伝的変異と集団構造の遺伝的分化が明確にされた[104]。

ゲノムの視点で、ダイズは、根の根粒形成、油糧種子生産、および二次代謝の点で豆類の比較研究のためのモデル植物として使用される。ダイズは、多くの品種の遺伝資源が利用できるため、ゲノム研究のための価値のある材料でもある。2010 年米国での大きな努力によって、ダイズ栽培種 Williams 82 の複二倍体のダイズゲノム配列が公開された（３つのバージョン Gmax109、Gmax189、Gmax275 のゲノム配列と遺伝子モデルが存在する）[29]。この品種は 1906 年に中国、北京から導入された品種 Peking から 1921 年に選抜された供与親 Kingwa が選抜され、疫病菌 Phytophthora の根腐れ耐性遺伝子座を戻し交配して作られた[117]。

日本では、国内の栽培条件と日本の生産者が持つ様々な用途に合わせてダイズ品種が開発されてきた。Williams 82 ゲノム配列は、多くの品種間の多様性を理解するのに有用であるが、日本のダイズ栽培に使用することができるゲノムリソースを有することが必要である。ここでは、長野県農業試験場桔梗ヶ原分場（現長野県野菜花き試験場）で 1971 年に開発された農林 2 号と東山 6 号（シロメユタカ）を親とする日本のダイズ品種エンレイ[118]のドラフトゲノム、系統解析およびアントシアニン・フラボノイド生合成およびプロテオーム

プロファイルを含むダイズ育種のための主要な特性に焦点を当て、日本のダイズ品種エンレイのゲノム配列の解析を示す。

2.3. 材料と方法

2.3.1. ゲノムシークエンシング

植物材料は農業生物資源研究所(以降、NIAS と呼ぶ)(現国立研究開発法人農業・食品産業技術総合研究機構)のジーンバンクより提供された。オルガネラ DNA を減らした高品質の核 DNA は、BAC DNA ライブラリ作成のゲノム DNA 抽出のために設計されたプロトコルを変更し使用し、若い葉から抽出した[119]。

配列決定はオペロンバイオテクノロジー社 (Eurofins ゲノミクス) で Illumina HiSeq2000 を使用して得られた。スタンダードショートリードライブラリと 8 kbp インサートのメイトペアーライブラリは、配列決定のため TruSeq SBS の V5 を使用して構築された。配列決定の後、ベースコールのため、HiSeq コントロールソフトウェア v.1.4.8 と CASAVA 1.8.1 (Illumina) を使用した。GS FLX Titanium General Library Preparation Kit and Rapid Library Preparation Kit (Roche)を用いて、シングルエンドライブラリと 3 kbp のメイトペアーライブラリを構築した。構築したライブラリは、NIAS の Roche 454 FLX Titanium で配列を読み出し、Roche 454 FLX Titanium のベースコーラで、配列を決定した。

2.3.2. アセンブルとレファランスマッピング

ゲノムの包括的な分析を容易にするために、*De novo* ゲノムアセンブリ (G.max_Enrei1)とレファランスゲノムアセンブリ(G.max_Enrei2)を構築した。G.max_Enrei1 アセンブリは、Roche 454 FLX Titanium でシークエンスしたシングルエンド配列と３kbps のメイトペアー配列、Illumina HiSeq2000 でシークエンスした 300bps のペアードエンド配列と 8kbps のメイトペアー配列、ABI 3730XL でシークエンスした約 100kbps の BAC エンド配列を Roche Newbler 2.7 を使用してアセンブルした。

G.max_Enrei2 アセンブリは、Roche シークエンサから得られたシングルエンド配列と Illumina HiSeq2000 シークエンサから得られたペアードエンド配列を BWA 0.7.5a[120]で Williams 82 のバージョン Gmax275（以降、Gmax275）ゲノム配列にマップし、SAMtools 0.1.19[121]でインデルを呼び出した後、NIG script[122]で、レファランスゲノムを作成した。

DNA マーカーは、Williams 82 ゲノム構築時に使用された SSR マーカー、EST-SSR マーカーなどの配列、エンレイの SNP-SSR から作成されたマーカー等を使用して作成された。

BLASTn[123]を使用して G.max_Enrei2 シュードモレキュルとスキャフォールドに DNA マーカーをマップし、DNA マーカーの順序を確認した。DNA マーカー配列はクリアシークエンス領域、ギャップ領域、BAC エンド配列のヒット位置もしくはヒット位置から推定される領域、*De novo* アセンブル由来のスキャフォールドのヒット位置にマップされ、これらの情報を使い、DNA マーカー順を入れ替えるための切断点が決定され、レファランスマッピングで作られたシュードモレキュルを再構築した。

2.3.3. 遺伝子モデリング

リピート配列をマスクした Gmax275 ゲノムの領域［１６番染色体、30,000,000-37,887,014 bps］を使い、Augustus[124]で、Augustus で使用するパラメータファイルを構築した。RepeatMasker[125]で、G.max_Enrei2 のシュードモレキュルやスキャフォールドからトランスポゾンを除去した配列を作成し、augustus-3.0.2[124]で遺伝子モデルを構築した。RepeatMasker[125]で遺伝子モデルからトランスポゾンを除去し、更に、この遺伝子モデルをクエリーとし、soyTE データベース[126]をデータベースとした BLASTn サーチを行い、ビットスコア 100 以上の遺伝子モデルを除去した。これとは別に、Trinity version 2014-07-17[87]で、RNAseq (PRJDB3582) をアセンブルし、172,753 の遺伝子モデルを構築した。この遺伝子モデルは、EMBOSS getorf [127]を使用して、各最長の ORF を持つものとした。

2.3.4. 系統解析

シロイヌナズナ[128]、ミヤマハタザオ[129]、タルウマゴヤシ[36]、およびイネ[130]の遺伝子モデルのアミノ酸配列と Gmax275 と G.max_Enrei2 の遺伝子モデルのアミノ酸配列を用い、OrthoMCL v2.0.7[131]でクラスタリングした。不完全な遺伝子モデルを除き、さらに、ゲノムから作られた遺伝子モデルと RNAseq から作られた遺伝子モデルが一致する遺伝子モデルから作られたシングルコピー遺伝子（オルソログ）のセットを作成した。シングルコピー遺伝子のセットの塩基（コドンの 3 塩基目が、A/T/G/C の何でも同じアミノ酸になる塩基）で構築された各種の配列を Clustal Omega 1.2.0[132]を使ってアラインした。アラインされた配列を MEGA 6.06[133]を使用して、基礎となる系統樹を作成し、PAML 4.8a[134]、Multidivtime[135]、および FigTree1.4.2[136]を使用して、系統樹を作成した。

2.3.5. アントシアニン・フラボノイド生合成系

アントシアニン・フラボノイド生合成に関連した Gmax275 と G.max_Enrei2 の遺伝子モデルを OrthoMCL[131]でクラスタリングした。これらの遺伝子モデルを BLASTn で関連付けた。

2.3.6. プロテオーム解析

エンレイ品種のプロテオーム解析は、登熟したダイズ種子を用いた。10 個の種子子葉を液体窒素中で砕き、標準的な手順[137]を使って相分離で精製した。精製タンパク質をトリプシンで消化した。質量分析のために、溶出されたペプチドは、タンパク質同定のために使用した MS スペクトルとナノスプレーLTQ XL Orbitrap 質量分析計で分析した。タンパク質の同定は、Williams 82 バージョン Gmax189（以降 Gmax189）のダイズペプチド配列 54,175[29]に対して Mascot 検索エンジンのバージョン2.4.1(Matrix Science, London, UK)およびProteome Discoverer のソフトウェアバージョン 1.4.0.288 (Thermo Fisher Scientific) を用いた。

Mascot の結果は、ペプチド同定の精度と感度向上のために Mascot Percolator ソフトウェアを使用してフィルタされた[138]。篠田ら[139]の記載のようにタンパク質の存在量は、emPAI 値を使用して分析した。Gmax275-Gmax189 の遺伝子対応リスト[29]を使用して、遺伝子モデル Gmax189 で作成された結果を Gmax275 の遺伝子モデルに変換した。OrthoMCL[131]を使って Gmax275 と G.max_Enrei2 の遺伝子モデルをクラスタリングした後、クラスタリングされた Gmax275 と G.max_Enrei2 の遺伝子モデルを BLASTn で関連付けた。

2.4. 結果と考察

2.4.1. ゲノムシークエンシングとレファランスマッピング

シークエンスされた配列情報を *De novo* アセンブリとレファランスマッピングで使用した配列を表 2-1 に示す。

表 2-1 *De novo* アセンブリとレファランスマッピングで使用した配列

シークエンサ	読取方法	インサート	カバー率	塩基数［bp］	アクセッション番号	リード数	平均リード長[bp]	*De novo* アセンブリで使用	レファランスマッピングで使用	備考
Illumina HiSeq2000	PE	300bp	12.1x	12,132,771,213	DRR021742	131,238,300	92	o	o	
	MP	8Kbp	7.7x	7,697,646,089	DRR021743	98374888	78	o		
Roche FLX titanium	SE	-	10.1x	10,170,746,555	DRR021740	25,690,322	396	o	o	
	MP	3Kbp	0.43x	429,217,443	DRR021741	1232922	348	o		
Life Technologies ABI 3730xl	BES	110/178 kbp	0.097x	97,235,392	LB000001-LB184894	184,902	526	o		登録時に8配列削除

DDBJ BioProject ID PRJDB3582

30.4 倍のカバレッジの配列を用いて、エンレイゲノムの *De novo* アセンブル配列 G.max_Enrei1:アクセッション番号 BBNX01000001-BBNX01092182 (92,182 エントリ)が得られた。22.2 倍のカバレッジの配列を用いて、エンレイゲノムのレファランスマッピング配列が得られた。G.max_Enrei2 に DNA マーカーをマップし、連鎖群との差がある 8 箇所の部位（補足表 2-1）を得た。ゲノムのギャ

ップ位置、BAC エンド配列のマップ位置、DNA マーカーのマップ位置、G.max_Enrei1 のマップ位置を基に決定された切断点を使い、修正されたレファランスマッピング配列 G.max_Enrei2:アクセッション番号 BBNX02000001-BBNX02108601（10,8601 エントリ）とシュードモレキュル（表 2-2）を作成した。シュードモレキュルと Gmax275 ゲノム（ギャップあり 978,495,272bps、ギャップなし 955,380,172bps）[29]との長さの比較では、シュードモレキュルのギャップありでは 501,501bps 短く、ギャップなしでは 27,675,438bps 短かった。また、エンレイゲノムに Gmax275 遺伝子モデル、DNA マーカー、BES エンド配列をマップし、表 2-3 エンレイゲノムにマップされた数・割合を得た。

表 2-2 エンレイゲノムアセンブルと遺伝子アノテーション

レファランスマッピング	ギャップあり[bp]	ギャップなし[bp]	割合[%]
染色体配列長	946,877,581	904,901,085	95.6
スキャフォールド配列長	31,116,190	22,803,649	73.3
合計長	977,993,771	927,704,734	94.9
遺伝子モデル			
遺伝子モデル数	60,838		
平均コード配列長	1455.3[bp]		
平均エクソン数	4.5		
平均エクソン長	323.4[bp]		

表 2-3 エンレイゲノムにマップされた数・割合

項目	配列数	マップされた数	割合(%)	備考
Gmax275遺伝子モデル	56,264	56,043	99.6	
エンレイDNAマーカー	1,860	1,773	98.8	別冊表2-2
エンレイBACエンドペアー配列	92,451	70,551	76.3	

2.4.2. 一塩基多型、挿入・欠失

max275 ゲノムに G.max_Enrei2 ゲノム構築で使用したシークエンス配列をマップすることで、合計 1,659,041 の一塩基多型（SNP）と 344,418 の挿入・欠失（INDEL）を同定した（表 2-4）。Gmax275 と G.max_Enrei2 ゲノム間で、一塩基多型（SNP）と挿入・欠失（INDEL）は存在し、表 2-4 は二品種のゲノム構造の違

いを示している。主なところは、SNP および INDEL の両方が１８番染色体で多く、１１番染色体で少なかった。Gmax275 に対する SNP 間の平均距離は、589.8 bp/SNP、最小距離は１８番染色体の 320.8 bp/SNP、最大距離は５番染色体の 984.9 bp/SNP であった。

表 2-4 エンレイゲノムの一塩基多型と挿入・欠失

Chr / Scaffold	Total	SNPs	INDELs	Gmax275 length (bp)	Ave. distance between SNPs (bp/SNP)
Chr01	100,579	83,446	17,133	56,831,624	681.1
Chr02	74,948	59,609	15,339	48,577,505	814.9
Chr03	141,828	119,939	21,889	45,779,781	381.7
Chr04	145,472	124,527	20,945	52,389,146	420.7
Chr05	55,513	42,883	12,630	42,234,498	984.9
Chr06	120,485	100,191	20,294	51,416,486	513.2
Chr07	80,162	65,291	14,871	44,630,646	683.6
Chr08	70,402	55,030	15,372	47,837,940	869.3
Chr09	83,562	66,788	16,774	50,189,764	751.5
Chr10	86,648	70,730	15,918	51,566,898	729.1
Chr11	48,787	38,151	10,636	34,766,867	911.3
Chr12	68,258	55,465	12,793	40,091,314	722.8
Chr13	112,150	90,966	21,184	45,874,162	504.3
Chr14	68,399	55,182	13,217	49,042,192	888.7
Chr15	130,143	111,062	19,081	51,756,343	466.0
Chr16	91,740	75,716	16,024	37,887,014	500.4
Chr17	88,611	73,878	14,733	41,641,366	563.7
Chr18	209,015	180,878	28,137	58,018,742	320.8
Chr19	138,041	118,469	19,572	50,746,916	428.4
Chr20	70,129	55,818	14,311	47,904,181	858.2
Scaffolds	18,587	15,022	3,565	29,311,887	1951.3
Total	2,003,459	1,659,041	344,418	978,495,272	589.8

2.4.3. 遺伝子モデル

Augustus で予測された遺伝子モデル数は、107,423 個となった。この遺伝子モデルを RepeatMasker にかけリピート配列を除いた遺伝子モデル数は、80,519 個、更に、ダイズ固有のトランスポゾンを除くため、soyTE データベースでヒットした遺伝子モデルを除き、最終的に、60,838 個のスプライスバリアント（DNA から 3 個以上のエクソンが切り出される場合、mRNA が生成される過程で、エクソンが選択的に使用されることで、異なる活性、構造を持つタンパクが生成されること）がない遺伝子モデルを得た（表 2-2）。Gmax275 の遺伝子モデル 56,044 個 [29]（スプライスバリアントなし）との比較では、コーディング平均配列長がエンレイで 1,455bps、Gmax275 で 1,168bps となり、エンレイの方が長く、平均エクソン長がエンレイで 323bps、Gmax275 で 231bps となり、エンレイの方が長かった。平均エクソン数では、Gmax275 で 5 個、エンレイで 4.5 個であり、エンレ

イの方が少なかった。若葉由来の RNAseq（表 2-5）から作られた最も長いオープンリーディングフレーム(ORF)を持つ遺伝子モデルが 172,753 個あり、ゲノムから作られた遺伝子モデルにマップした。50%以上のカバーした遺伝子モデルは 20,542 個、90%以上のカバーした遺伝子モデルは 5,950 個、完全に一致した遺伝子モデルは 2,269 個であった。

表 2-5 若葉から抽出した cDNA の配列とアセンブル

Library name	EW1_ATCACG_L001..EW1_ATCACG_L007
No. of libraries	7
Insert size	300 bp
No. of entries	71,134,481
No. of contigs	754,548
No. of ORFs	395,330
No. of longest ORFs	172,753

Gmax275 と G.max_Enrei2 間の遺伝子モデルの違いは、二品種間の SNP だけでなく、遺伝子モデルの構築に使用されるいくつかのパラメータに起因する可能性があると考えられる。

2.4.4. 系統解析

OrthoMCL[131]で、シロイヌナズナ[128]、シロイヌナズナの近縁種ミヤマハタザオ[129]、ダイズ品種 Williams 82[29]、エンレイ、メタルウマゴヤシ[36]、イネ(Os-Nipponbare-Reference-IRGSP-1.0)をクラスタリングし、整列させた。これらの種間の系統関係と分岐年代推定のため、RNAseq から作られた遺伝子モデルと完全一致するエンレイの遺伝子モデルのシングルコピー遺伝子のセットが、選択された（別冊表 2-3）。

シロイヌナズナとミヤマハタザオの分岐年代は、約 13Mya[140]と推定される。この値を基準にすると計算されたシロイヌナズナとミヤマハタザオの分岐年代（95PD： 19.30 から 8.52Mya）、Williams 82 とエンレイの分岐年代は、0.34Mya（95PD: 0.78 から 0.10Mya）と推定された（図 2-1）。

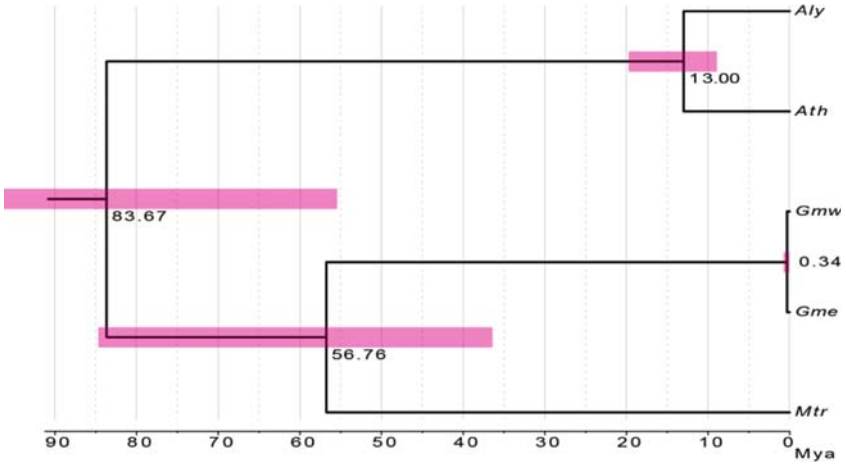


図 2-1 分岐年代。　Gmw:ダイズ Williams 82、Gme:エンレイ、Ath:シロイヌナズナ、Aly:ミヤマハタザオ、Mtr:タルウマゴヤシ、ピンク色のバーは、95%の確率密度、Mya は百万年の単位である。

また、ダイズ属とタルウマゴヤシの分岐年代は 56.76Mya（95PD：84.54 から 36.99Mya）と推定され、ダイズ/タルウマゴヤシ属とシロイヌナズナ属の分岐年度は、83.67Mya（95PD：122.51 から 55.57Mya）であった。以前の研究では、ダイズ属とタルウマゴヤシの分岐年代が 54Mya と推定される[141]。58Mya 頃発生した全ゲノム重複（WGD）がタルウマゴヤシを形作る主要な要因である[36]。

ダイズの祖先とダイズの野生種は、0.27Mya[115]もしくは 0.8Mya[116]に最も近い共通祖先から分岐した。分岐が 0.8Mya だとすると、Williams 82 とエンレイの分岐年代は 0.34Mya とより新しかった。Li らの指摘にあるように、分岐選択は、ダイズ種の祖先と野生種ダイズの分化に寄与した可能性がある[116]。異なる環境への適応としての分岐選択は、最も近い共通祖先から Williams 82 とエンレイ双方の分化に寄与したと考えられる。

2.4.5.　アントシアニン・フラボノイド生合成系

ダイズのいくつかのカルコン合成(CHS)遺伝子 *CHS3*（ P19168 ）、*CHS1*（ P24826 ）、*CHS7*（ P30081 ）、*CHS4*（ Q6XON0 ）、および *CHS8*（ AY237728 ）は、種皮の色素沈着に関連付けられている[142]。これら *CHS* 遺伝子の物理的位置は、RNA サイレンシングに関連付けられている遺伝子座を持つ BAC アセンブリ[143,144]や

WGS アセンブリ[29]を使用して決定された。Gmax275 ゲノムアセンブリに対応する遺伝子は、以下のとおりである。

CHS1（Glyma.08G109400）、*CHS2*（Glyma.05G153200）、*CHS3*（Glyma.08G110300 と Glyma.08G110900）、*CHS4*（Glyma.08G110500 と Glyma.08G110700）、*CHS5*（Glyma.08G109200、Glyma.08G109300、および Glyma.08G110400）、*CHS6*（Glyma.09G075200）、*CHS7*（Glyma.01G228700）、*CHS8*（Glyma.11G011500、および *CHS9*（Glyma.08G109500）

このパスウェイにおける大多数の遺伝子は、Williams 82 とエンレイの双方にある（図 2-2）。

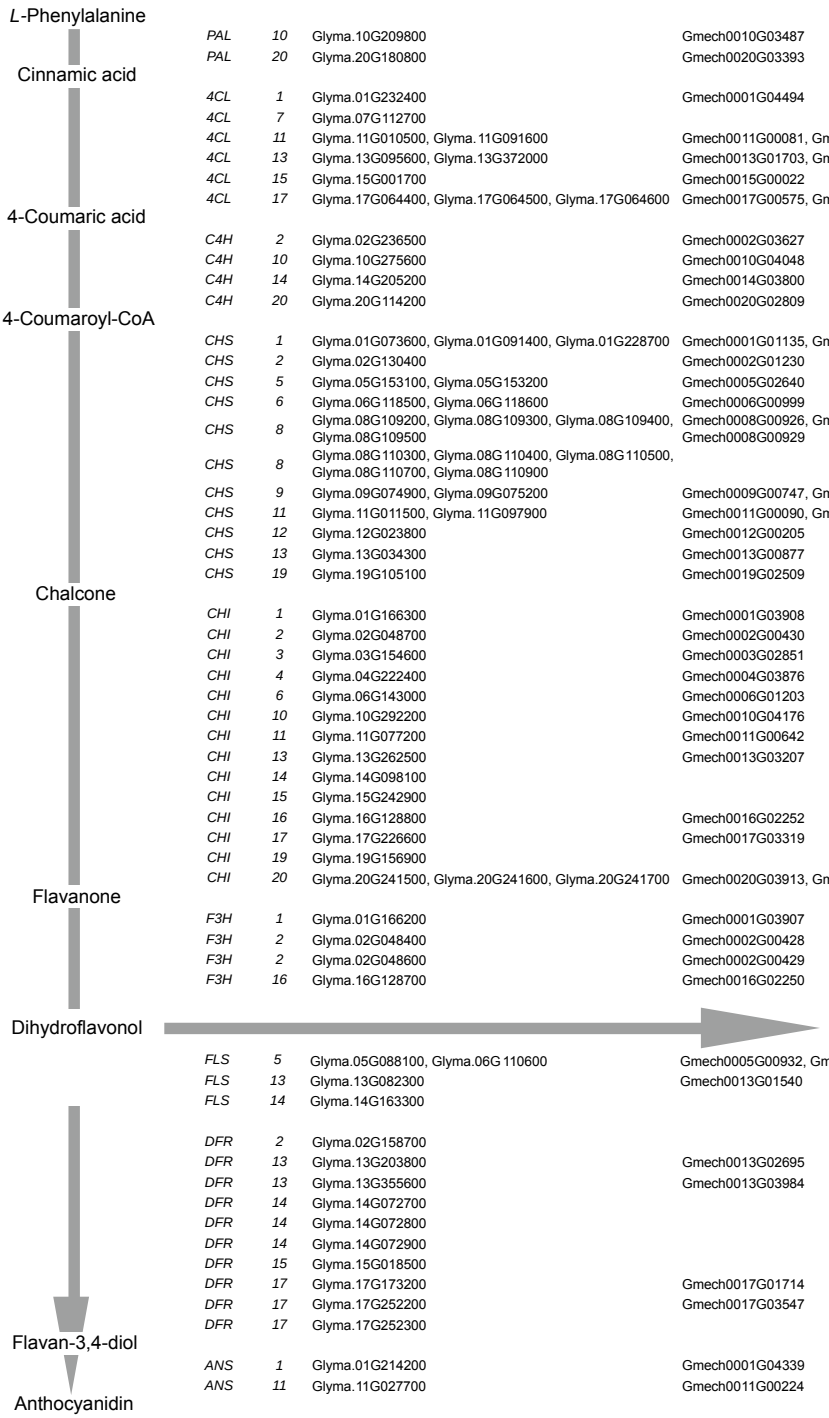

Intermediate
Enzyme
Chr.no.
Williams82 gene name (Gmax275)
ENREI gene name
L-Phenylalanine
PAL 10 Glyma.10G209800 Gmech0010G03487
PAL 20 Glyma.20G180800 Gmech0020G03393
Cinnamic acid
4CL 1 Glyma.01G232400 Gmech0001G04494
4CL 7 Glyma.07G112700
4CL 11 Glyma.11G010500, Glyma.11G091600 Gmech0011G00081, Gmech0011G00767
4CL 13 Glyma.13G095600, Glyma.13G372000 Gmech0013G01703, Gmech0013G04132
4CL 15 Glyma.15G001700 Gmech0015G00022
4CL 17 Glyma.17G064400, Glyma.17G064500, Glyma.17G064600 Gmech0017G00575, Gmech0017G00576, Gmech0017G00577
4-Coumaric acid
C4H 2 Glyma.02G236500 Gmech0002G03627
C4H 10 Glyma.10G275600 Gmech0010G04048
C4H 14 Glyma.14G205200 Gmech0014G03800
C4H 20 Glyma.20G114200 Gmech0020G02809
4-Coumaroyl-CoA
CHS 1 Glyma.01G073600, Glyma.01G091400, Glyma.01G228700 Gmech0001G01135, Gmech0001G02138, Gmech0001G04461
CHS 2 Glyma.02G130400 Gmech0002G01230
CHS 5 Glyma.05G153100, Glyma.05G153200 Gmech0005G02640
CHS 6 Glyma.06G118500, Glyma.06G118600 Gmech0006G00999
CHS 8 Glyma.08G109200, Glyma.08G109300, Glyma.08G109400, Glyma.08G109500 Gmech0008G00926, Gmech0008G00927, Gmech0008G00928, Gmech0008G00929
CHS 8 Glyma.08G110300, Glyma.08G110400, Glyma.08G110500, Glyma.08G110700, Glyma.08G110900
CHS 9 Glyma.09G074900, Glyma.09G075200 Gmech0009G00747, Gmech0009G00751
CHS 11 Glyma.11G011500, Glyma.11G097900 Gmech0011G00090, Gmech0011G00823
CHS 12 Glyma.12G023800 Gmech0012G00205
CHS 13 Glyma.13G034300 Gmech0013G00877
CHS 19 Glyma.19G105100 Gmech0019G02509
Chalcone
CHI 1 Glyma.01G166300 Gmech0001G03908
CHI 2 Glyma.02G048700 Gmech0002G00430
CHI 3 Glyma.03G154600 Gmech0003G02851
CHI 4 Glyma.04G222400 Gmech0004G03876
CHI 6 Glyma.06G143000 Gmech0006G01203
CHI 10 Glyma.10G292200 Gmech0010G04176
CHI 11 Glyma.11G077200 Gmech0011G00642
CHI 13 Glyma.13G262500 Gmech0013G03207
CHI 14 Glyma.14G098100
CHI 15 Glyma.15G242900
CHI 16 Glyma.16G128800 Gmech0016G02252
CHI 17 Glyma.17G226600 Gmech0017G03319
CHI 19 Glyma.19G156900
CHI 20 Glyma.20G241500, Glyma.20G241600, Glyma.20G241700 Gmech0020G03913, Gmech0020G03914, Gmech0020G03915
Flavanone
F3H 1 Glyma.01G166200 Gmech0001G03907
F3H 2 Glyma.02G048400 Gmech0002G00428
F3H 2 Glyma.02G048600 Gmech0002G00429
F3H 16 Glyma.16G128700 Gmech0016G02250
Dihydroflavonol
Flavanol
FLS 5 Glyma.05G088100, Glyma.06G110600 Gmech0005G00932, Gmech0006G00935
FLS 13 Glyma.13G082300 Gmech0013G01540
FLS 14 Glyma.14G163300
DFR 2 Glyma.02G158700
DFR 13 Glyma.13G203800 Gmech0013G02695
DFR 13 Glyma.13G355600 Gmech0013G03984
DFR 14 Glyma.14G072700
DFR 14 Glyma.14G072800
DFR 14 Glyma.14G072900
DFR 15 Glyma.15G018500
DFR 17 Glyma.17G173200 Gmech0017G01714
DFR 17 Glyma.17G252200 Gmech0017G03547
DFR 17 Glyma.17G252300
Flavan-3,4-diol
ANS 1 Glyma.01G214200 Gmech0001G04339
ANS 11 Glyma.11G027700 Gmech0011G00224
Anthocyanidin

図 2-2 アントシアニン・フラボノイド生合成のための主要なパスウェイに関与する酵素、Gmax275 とエンレイの対応する遺伝子。 *PAL* (Phenylalanine ammonia-lyase)、*4CL* (4-coumaroyl-CoA-ligase)、*C4H* (cinnamate-4-hydroxylase)、*CHS* (chalcone synthase)、*CHI* (chalcone reductase)、*F3H* (flavanone 3-hydroxylase)、*FLS* (flavonol synthase)、*DFR* (dihydroflavonol 4-reductase) and *ANS* (anthocyanidin synthase).

しかし、７番染色体の一つの *4CL* 遺伝子、８番染色体の５つの *CHS* 遺伝子、１４番/１５番/１９番染色体の３つの *CHI* 遺伝子、１４番染色体の一つの *FLS* 遺伝子、２番/１４番/１５番/１７番染色体の６つの *DFR* 遺伝子が、エンレイゲノムでは存在が認められなかった。UniProt (The Universal Protein Resource) でアノテーションが付けられ位置が決定された両種の *CHS* 遺伝子以外の遺伝子は、断片化された配列であったため、エンレイでは存在が認められなかった（図 2-3）。

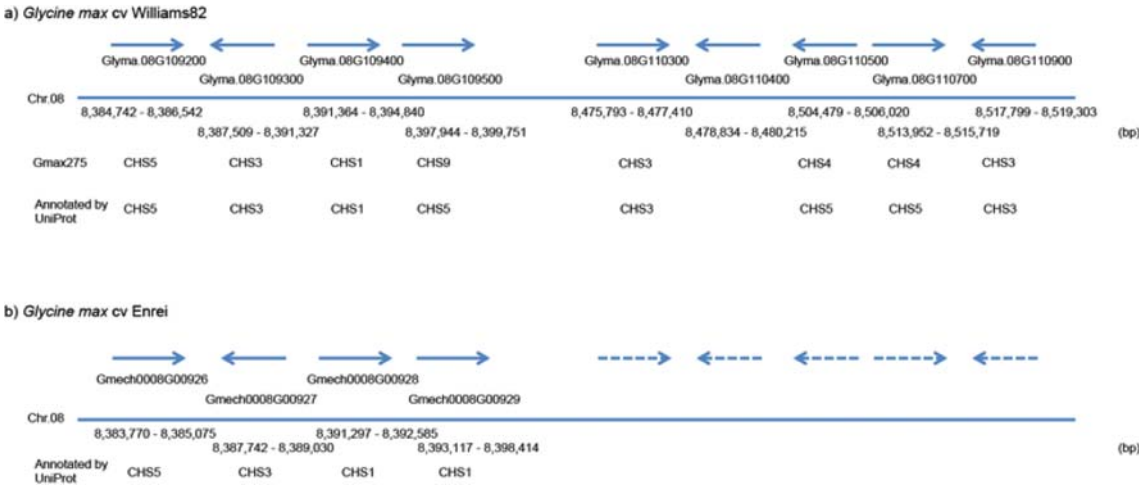


図 2-3 ダイズ８番染色体の *CHS* 遺伝子クラスタの位置を示す領域。 *CHS* 遺伝子クラスタによって特徴付けられた Gmax275 の領域[８番染色体: 8.3-8.5 Mb](a) に対応するエンレイ(b)。上流側の *CHS* 遺伝子は両者間で関係付けられたが、下流側の *CHS* 遺伝子は断片化配列のためエンレイでは存在が確認できなかった。

これら遺伝子の殆どは、豆の皮色やヘソの色に関するものであり、アントシアニン・フラボノイド生合成パスウェイのエンレイにおけるこれらの欠落は、siRNA活性[145,146]に関係しているかも知れない。

2.4.6. 子葉におけるタンパク質

種子プロテオームデータに関連する遺伝子モデルを使用して、貯蔵タンパク質、脂質合成/分解酵素、タンパク質輸送/折り畳み、LEA（Late embryogenesis abundant）タンパク質、解糖経路の酵素、プロテアーゼ/プロテアーゼ阻害剤などに対応する 164 の子葉におけるタンパク質の遺伝子モデル（別冊表 2-4）が同定された。

乾燥種子のタンパク質含有量は乾物重の 35%から 42%であり[147,148]、子葉タンパク質の 70%は cupin スーパーファミリ[149]の一部であるβ-コングリシニンに相当する 7S グロブリンとグリシニンに相当する 11S グロブリン[147,150,151]である。登熟したダイズ種子の関連する子葉タンパク質を同定するために、子葉のプロテオーム解析を行い、160 個の Gmax189 のタンパク質遺伝子モデルは範囲 7.87mol%から 0.03mol% （別冊表 2-4）のエンレイのタンパク質遺伝子のモデルに対応する。これらのタンパク質のほとんどは、貯蔵タンパク質やβ-コングリシニンやグリシニンを含む cupin であり、総 mol%の約 42%、総重量の約 55%（質量*モルの和）である（表 2-6）。

表 2-6 エンレイにおける貯蔵タンパクおよび cupin 成分

Chromosome	Related no. of gene	Weight % (mass * mol)	mol %
Chr10	6	19.8	14.27
Chr20	4	15.7	14.58
Chr03	1	6.6	4.31
Chr13	1	4.6	3.06
Chr19	1	2.8	1.88
Chr04	1	2.4	1.99
Chr02	1	2.1	1.36
Chr11	1	1.2	0.85
Chr01	1	0.1	0.07
total	17	55.4	42.4

種子貯蔵タンパク質の含有量を制御する遺伝子は多く存在する[152,153]。リポキシゲナーゼ 1、パーオキシゲナーゼ 2、およびオレオシンファミリータンパク質遺伝子[154]のような脂質代謝に関連した遺伝子、HSP20 のようなシャペロン、PDI-like、SNF7 ファミリ、および液胞ソーティング受容体タンパク質のようなソーティング/折り畳み関連タンパク質、凝集での他のたんぱく質からの保護に重要であるかも知れない LEA タンパク質遺伝子などが、エンレイゲノムの中で多く現れた。また、発芽段階で重要な役割を果たし得る解糖パスウェイ、酵素、およびプロテアーゼ/プロテアーゼ阻害剤に含まれるいくつかの遺伝子が見つかった。このプロテオームプロフィールは、栽培条件への適応や品種多様性を理解するための基盤を提供する。

2.4.7.　エンレイゲノムデータベース

ゲノム情報表示には DAIZUbase[155]を用いた。エンレイより BAC ライブラリを作成し、この BAC エンド配列をサンガー法シークエンサでリードしたエンレイ BAC エンド配列を Williams 82 ゲノムにマップすることにより BAC エンド配列の物理的位置を決定し、DAIZUbase で公開した。すべての配列データは日本のダイズ品種エンレイに焦点をあてたダイズゲノムのインフォマッティックスリソースとして、DAIZUbase[156]に加えた。このデータベースには、エンレイゲノム配列とエンレイ BAC クローンとそれに付随するすべての注釈を表示するためのインタラクティブなページ付きの GBrowse[97]を用意した。また、DAIZUbase には、遺伝地図とエンレイ品種の物理的地図との関係を示す地図が含まれる。

2.5.　結論

国内栽培に合うダイズの品種改良のための様々な情報を国産ダイズ品種エンレイゲノム配列は提供する。ゲノム配列は、日本のダイズゲノム構造解析、農業的に重要な形質のランドマークとなる DNA マーカーの開発、ダイズにおける重要な遺伝子の同定のための研究資源の開発、病害および病害虫抵抗性、生産性および地域適応性のような重要な形質を制御する遺伝子の単離を通じて、効果的なダイズ育種のための新しい戦略を補完する。特定の形質を制御する遺伝子

の詳細な知識は、より効果的なダイズの改良を可能にし、研究者に様々な環境条件に適応する植物型の開発を許容する。

第3章　カイコゲノム統合データベース開発

3.1.　概要

大規模な絹生産を行うためのカイコ *Bombyx mori* は、発展途上国で経済的に最も重要な昆虫の1つである。ゲノムやバイオテクノロジツールの進歩により、カイコは生物医学的に興味のある様々な組換タンパク質の生産のためのバイオリアクタ（生化学反応を行う装置）にもなっている。2004年にカイコゲノムに関する2つのゲノム塩基配列決定プロジェクトが日本および中国のチームによって報告された。しかし、両者のゲノムは、明確なゲノムアノテーションに不可欠な長いゲノムスキャフォールドを構築するにはデータセットが不十分であった。2008年に日中のデータセットが、日中の共同作業によって統合され、カイコゲノムが構築された。日本および中国のグループが、カイコホールゲノムショットガン配列決定のために準備したデータセットに、新たに日本側からfosmidエンド配列およびBACエンド配列が加えられた。構築された配列は、昆虫ゲノムの中で最良の連続性：N50（配列を降順にならべ、全長の50%のときの配列長）スキャフォールドサイズで～3.7Mb、28本の染色体すべてに対する塩基カバー率は88%となった。シュードモレキュル作成では、BACクローンのフィンガプリンティングによって構築されたBACコンティグの物理マップおよびBACエンド配列を用いて構築されたSNPリンケージマップを利用した。

並行して、様々な組織および発育段階における二次元ポリアクリルアミドゲル電気泳動のプロテオームデータを、カイコプロテオームデータベースKAIKO2DDBにまとめた。最後に、トランスポゾン挿入系統の挿入位置および発現データを記録するための *Bombyx* trapデータベースを構築した。機能解析研究におけるゲノム情報の効率的利用のために、ゲノム配列、物理的および遺伝的地図情報およびESTデータをKAIKObaseに組込んだ。

カイコゲノム統合データベースKIAKObaseは、4種類のマップビュア、ジーンビュア、および配列、キーワード、位置の検索システムからなり、ゲノム配列、遺伝子、スキャフォールドおよび染色体レベルで結果およびデータを表示

する。更に、カイコプロテオームデータベース KAIKO2DDB と *Bombyx* trap データベースを KAIKObase と統合することにより、包括的で効率的なカイコのゲノムデータベースを構築した。

3.2. はじめに

カイコ *Bombyx mori* は、クワコ（野生のカイコ）*Bombyx mandarina* から約 5,000 年前に絹生産のために家畜化された。カイコは飼いならされた唯一の昆虫であり、カイコの生存と繁殖は人間に完全に依存している。現在、カイコは多くの開発途上国において大規模な絹生産のために利用され、経済的に最も重要な昆虫の一つである。ゲノムレベルでのクワコとの比較は、家畜化に至る人工的選択の影響を調べる機会を提供する。更に、カイコは最も有害な農業害虫を含む昆虫の中で 2 番目に種類が多い鱗翅目に属する昆虫で、モデル生物でもある。バイオテクノロジの発展により、カイコは組換えタンパク質の生産のための重要なバイオリアクタとして使用されるようになった[157,158]。カイコのゲノム情報は、養蚕の改善に強い影響を与えるだけでなく、害虫駆除のための新しい方法の開発を容易にするものである。

昆虫は地球上で最も多様な種であり、それらの特徴的な生物学的現象は基礎科学および産業にとって重要な資源であるため、昆虫のゲノム解析は近年急速に進められた。ショウジョウバエ[15]、ハマダラカ[16]、ミツバチ[17]およびコキヌモドキ[18]のドラフトゲノム配列が公表されている。2004 年にはカイコのホールゲノムショットガン（WGS）配列が日本[159]と中国[160]で報告されたが、これら両者のゲノム情報は、他の種の解析と比較してシークエンス量が少ないためゲノム配列情報が不十分であった。その後、2 つの WGS データセットに新たに得られた fosmid エンド配列および BAC エンド配列を加え、ゲノムの再構築が実施された。これらの 2 つのゲノムシークエンスは 2 つの異なるカイコの系統に由来するが、両者の配列比較の結果、塩基レベルでわずか 0.2%の差しか示さなかった。また、日本の WGS に用いられた p50T 近交系は、中国系の Dazao 系統と同じ起源に由来していた。RAMEN アセンブラは、高感度領域のシード文字列のルックアップテーブル生成と、効率的なダイナミックプログラミングによるオーバーラップリードの迅速な検出と高速アライメントが使われ

ている。さらに、RAMEN には、2 つのユニークなサブコンティグに隣接するリピートサブコンティグを 1 つのユニークなコンティグに変換するためのリピートのもつれを解く方法が含まれているため、カイコゲノムの高密度トランスポゾンに関連する問題を回避することができる。配列決定された農業上重要な昆虫ゲノムの中で、カイコゲノムアセンブリ（432Mb）は最良の連続性（N50 スキャフォールドサイズで約 3.7Mb）を有し、28 本の染色体すべてに対する塩基カバー率は 88%を得た。これは、BAC エンド配列の解析で構築された統合された高密度 SNP リンケージマップの使用、FPC プログラムを用いた BAC フィンガプリンティングで確立されたコンティグの物理マップの統合で可能となった[108]。関連するプロジェクトでは、SilkBase[109]にさまざまな組織や異なる発育段階から得られた EST データが集められた。二次元ポリアクリルアミドゲル電気泳動と質量分析[110]から異なる発育段階の異なる組織のプロテオームデータが得られた。レポータ発現パターンおよび遺伝子トラップ系統やエンハンサトラップ系統[161]のミューテータの挿入された位置を提供するための *Bombyx* trap データベースが構築された。

様々な生物のゲノム情報の膨大な蓄積により、データと結果を視覚化するための広範なツールが開発された。遺伝地図と物理地図を表示できる AceDB[162]、INE[90]、NCBI Map Viewer[163]、Cmap[93,164]などのシステムで広く使用される。続いて、染色体またはスキャフォールドなどの Mb オーダーの広範なゲノム情報を表示する Ensembl[94]、GBrowse[97]、UCSC browser[96]、UTGB[98]などゲノムブラウザが開発された。

ユーザフレンドリで効率的なカイコのゲノムデータベースを構築するため、ゲノム配列、地図情報、EST、プロテオームデータ、エンハンサトラップの情報を統合して、データのアクセス性を向上させたカイコゲノム統合データベース KAIKObase というデータベースを構築した。塩基配列、スキャフォールドおよび染色体は、GBrowse および UTGB によって表示される。

3.3. データセット内容

3.3.1. ゲノム配列情報

KAIKObase にはスキャフォールド（アクセッション番号 DF090316-DF092116）およびコンティグ（アクセッション番号 BABH01000001-BABH01088672）のうち、スキャフォールドで使用されていないコンティグを含む合計 43,462 個の配列で、総ゲノムサイズは 482Mb（403Mb ギャップなし）が含まれる。そのうち 192 のスキャフォールドが 28 本の染色体にマッピングされた。スキャフォールド間に 500Kb の人為的なギャップを割り当て、全長が 503Mb（ギャップなしで 393Mb）に相当するシュードモレキュルを作成した。さらに、81,705 個の BAC エンド配列（アクセッション番号 DE283657-DE378560、DE378561-DE420875）、174,222 の fosmid エンド配列(アクセッション番号 DE143284-DE189151、DE246947-DE248527、DE420876-DE647768）および 166,757 の EST が含まれる[19]。EST のアクセッション番号をもった cDNA ライブラリーリストを別冊表 3-1 に示す。

3.3.2. ゲノム配列にマップされる情報

遺伝地図と物理地図は、カイコゲノムの基本情報を提供する。組み合わされたマップは、16,209 の遺伝子モデル、1,532 の SNP マーカー、770 の形質マーカー、および 5,419 の FPC コンティグを含む。14,622 の遺伝子モデルは、GLEAN に基づくアルゴリズム[165]を用いて中国のグループが作成した。さらに、GPCR[19]、OBP、CSP[19]、キューティクルタンパク質[19]および tRNA 遺伝子を含む 1,587 の遺伝子[19]は、日本のグループが自動注釈および手動注釈によって作成した。SNP マーカーは、BAC エンド配列から同定された。形質マーカーは、トランスポゾン挿入系統（エンハンサトラップ系統または遺伝子トラップ系統）におけるトランスポゾンベクタ（ミューテータ）の位置を表す。FPC コンティグは、BAC フィンガプリンティングによってアセンブルされた BAC コンティグを表す。

3.3.3. プロテオーム情報

カイコムプロテオームに関する情報は KAIKO2DDB[110]によって提供される。異なる発育段階（例えば、第 4 齢および第 5 齢幼虫、紡糸および蛹化段階）および様々な組織（例えば、中腸、脂肪体、中部絹糸線、後部絹糸線、マルピギ細管、卵巣、および血リンパ）を含む 2 次元ポリアクリルアミドゲル電気泳動は 116 の画像で構成される。2 次元ポリアクリルアミドゲル電気泳動のスポットは、分子

量等電点などの情報を提供する。このスポットに対応する EST および選ばれた遺伝子モデルもゲル画像上に表示される。

3.3.4. 発現遺伝子可視化情報

新規遺伝子の網羅的な探索と導入遺伝子の人為的な発現制御のため、エンハンサトラップ系統(補足 3-1 エンハンサトラップ系統参照)[161]が作出された。作出された系統のプロファイルをまとめるため、*Bombyx* trap データベースが構築された。エンハンサトラップ系統における EGFP-reporter 発現とミューテータの挿入部位に関する情報を含む *Bombyx* trap データベースには、288 個のトランスポゾン挿入系統、例えばエンハンサトラップ系統および遺伝子トラップ系統の情報、ゲノム配列中の挿入位置、種々の発育段階、器官および組織の遺伝子発現プロファイル、および関連する写真(遺伝子発現可視化)が含まれる。

3.4. データベース KAIKObase の構成

KAIKObase は、カイコのゲノム情報への入口として、4 つの地図ブラウザ(PGmap、UnifiedMap、GBrowse、UTGB)、遺伝子ビューア (GeneViewer)、2 つの独立データベース (KAIKO2DDB と *Bombyx* trap データベース)、配列検索、位置検索システムを持つ (図 3-1)。PGmap と UnifiedMap は、各染色体の利用可能な情報の全体像を提供する。UTGB および GBrowse は、同様な情報 (染色体毎のスキャフォールドから生成された領域に基づいた塩基レベル) を提供する。GeneViewer は各染色体上の遺伝子を含む遺伝子モデルの説明を提供する。

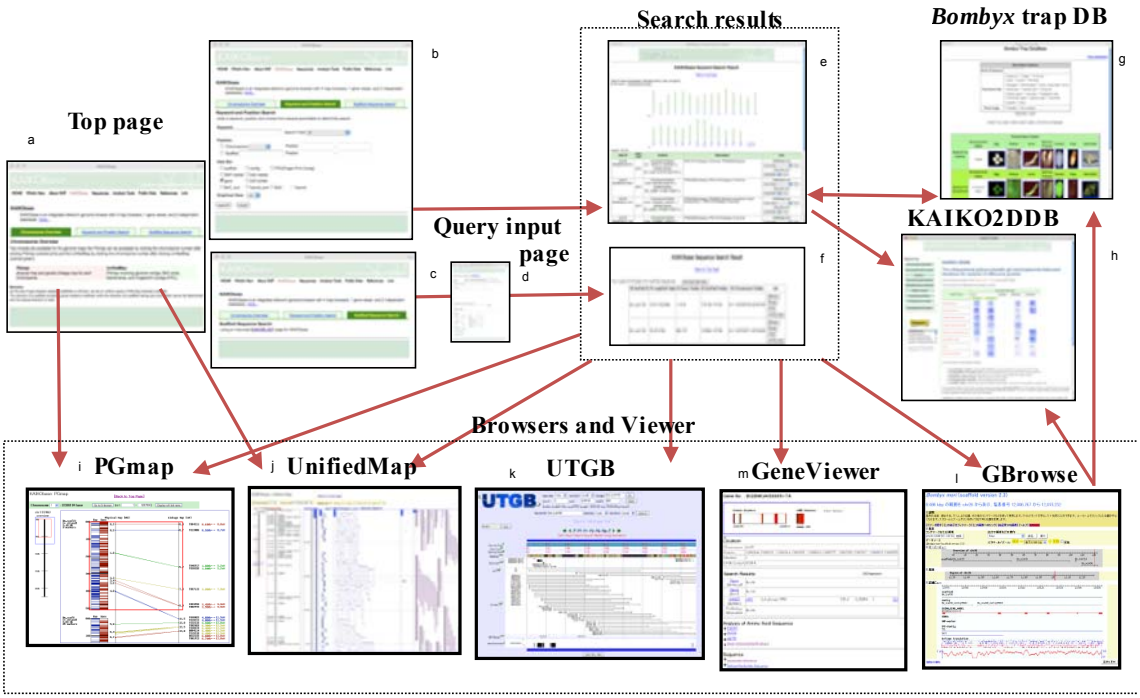


図 3-1 KAIKObase のフローチャート。 a）KAIKObase のトップページに PGmap と UnifiedMap へのリンクを設けた；b）キーワードおよび位置検索機能；c）BLAST を用いた配列検索機能；d）fasta シークエンスおよび設定パラメータの入力；e）キーワードおよび位置検索の結果；f）配列検索の結果；g）*Bombyx* trap データベースのトップページ；h）Proteome データベースのトップページ；i）PGmap 遺伝地図と物理地図の画像を示す；j）UnifiedMap 遺伝地図と様々な選択可能な物理的地図を示す；k）UTGB 様々な選択可能な物理的マップ機能を示す；l）GBrowse 様々な選択可能な物理的マップ機能を示；m）GeneViewer 遺伝子プロフィールを示す。

KAIKO2DDB（Proteome データベース）は、カイコの様々な発育段階と組織から生成されたプロテオームデータで構成される。このシステムは、ExPASy の make2DDB II（ver. 2. 50. 1）パッケージ[102, 166]を使用して開発した[110]。

Bombyx trap データベースは、トランスポゾン挿入系統（エンハンサトラップ系統または遺伝子トラップ系統）におけるトランスポゾンベクタ（突然変異体）の

レポータ発現および位置に関する情報を 2 つのデータマイニングアプローチ、すなわち「キーワード検索」および「図検索」で提供する。

配列検索機能は、NCBI BLAST ソフトウェア[123]を用いてクエリー配列のゲノム内での位置情報を提供する。

キーワードおよび位置検索機能は、クエリーキーワードのゲノム配列中の位置に関する情報およびその範囲を区切るための情報を提供する。

１）PGmap

PGmap は、遺伝地図と物理地図の外観を閲覧できる地図である。これは、SNP マーカー、*Bombyx* trap データベースで使用する形質マーカー、反復配列のバーチャート、および染色体範囲の遺伝子配列のバーチャートからなる。ゲノム中の領域または位置を指定することによって、染色体全体または特定の染色体領域の遺伝的および物理的長さを視覚的に比較することができる。特に、選択された領域の配列は、GBrowse に連結される。PGmap の Web インタフェースは、Javascript で記述された通信を使用する（図 3-2）。

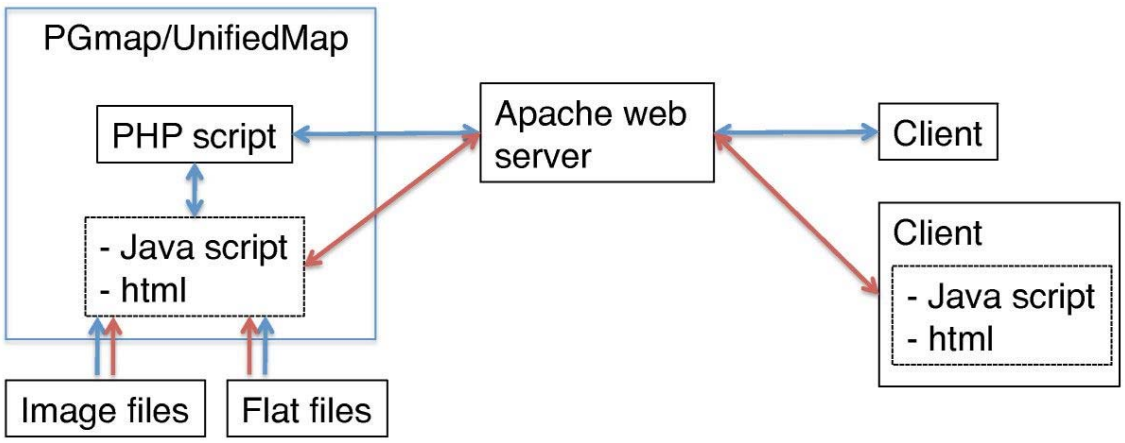


図 3-2 PGmap と UnifiedMap の通信。　青い線は、PGmap と UnifiedMap にアクセスして染色体番号を選択し、地図の縮尺を変更するときの情報の流れを表

す。赤い線は、カーソルを配置してUnifiedMapからの詳細情報を選択したときの情報の流れを表す。

２）UnifiedMap

UnifiedMapは、遺伝地図と物理地図の概要を提供するPGmapと同様の機能を持つ中程度の外観を閲覧するための地図である。目的の染色体範囲内のスキャフォールド、コンティグ、FPCコンティグ、BACエンド、fosmidエンド、SNPマーカー、および形質マーカーに関する詳細情報を提供する。チェックボックスからのオン・オフにより、物理的な地図アイテムを表示または非表示にする。4段階の表示スケーリングを持つ。遺伝地図内のマーカーは物理地図上のマーカーにリンクされ、物理地図上のマーカーはGBrowseおよび配列情報にリンクされる。Web インタフェースは、JavaScript で記述された通信を使用する（図 3-2）。

３）UTGB

UTGBは、PGmapおよびUnifiedMapによって提供される染色体範囲よりも小さいが、GBrowse カバレッジよりも大きい中間カバレージを提供する。表示項目は、FPCコンティグ、BACエンド、fosmidエンド、および遺伝子モデルである。UTGBで使用している情報を表示するトラックは、従来のゲノムブラウザの表現力と拡張性を高めた独立したWebアプリケーションである。

４）GBrowse

GBrowseは、制限酵素部位、FPCコンティグ、6フレーム翻訳、DNA / GC含量、コンティグ、EST、転写プロファイル、BAC、BACエンド、fosmidエンド、遺伝子モデルおよび遺伝子、SNPマーカーおよび形質マーカー情報をトラックとして提供する。遺伝子モデルトラックのポップアップバルーンは、1）GBrowse 機能によって表示される配列情報、2）各遺伝子の詳細情報を有するGeneViewer、および3）プロテオームデータベースへのリンクを示す。形質マーカーに関連するポップアップバルーンは、1）GBrowse 機能によって表示される配列情報、および

2) *Bombyx* trap データベースへのリンクを示す。さらに、スキャフォールドに EST をマッピングするために、フィルタオプションのない BLASTn 検索を使用した。クエリーは EST、データベースとしてスキャフォールドを使用した。BLAST e-value が 0.01 未満で BLAST のトップスコアを持った EST がマップされた。

５）GeneViewer

GeneViewer は遺伝子モデルの全体像を提供する。表示項目は、1）塩基およびスプライシングされた塩基の画像および Pfam データベースにおけるドメイン検索の結果；2）KAIKO2DDB へのリンク；3）染色体番号、エキソンの位置および GC 含量を含む予測された遺伝子に関する詳細情報；4）BLASTn（上位 3EST）、BLASTp（上位 10 タンパク質）、HMMER および ProfileScan の配列アラインメントを持った相同性検索の結果；５） PSORT、SOSUI、MOTIF および Gene ontology(InterProScan 結果から GO へのマップによる)におけるアミノ酸分析の結果、InterProScan のグラフ表示、および 6）予測遺伝子の塩基配列、スプライシングされた塩基配列および翻訳されたタンパク質配列にリンクする。

６）ソフトウェア

KAIKObase で使用されるソフトウェアはパブリックドメインから入手し、特定のデータに合わせて修正した。2002 年 Stein らによって開発された GBrowse の改定版は、GMOD (Generic Model Organism Database Project) [167]から利用可能であり KAIKObase ではこれを利用した。UTGB バージョン 1.0 は、UTGB (UT Genome Browser) [168]のゲノムブラウザである。BLAST 検索エンジンは、配列検索のために NCBI BLAST[123]バージョン 2.2.17 を使用する。データベースエンジンは、キーワードおよび位置検索で PostgreSQL[169]バージョン 8.2.1 を使用する。HMMER[170]バージョン 2.1.1、ProfileScan[171]バージョン 2.2、PSORT[172]バージョン 6.4、SOSUI[173]バージョン 1.0、MOTIF[174]、および InterProScan[175]バージョン 4.3.1（データバージョン 14.0）は GeneViewer で使用される。

3.5.　使用方法と考察

3.5.1. ユーザインタフェース

データベースの相互運用性を図 3-3 に示す。赤い矢印はユーザインタフェースを表す。ユーザは、PGmap と UnifiedMap を使用してゲノム領域を指定し、関連するすべての情報を取得し、GBrowse、UTGB、および GeneViewer へのリンクを介してデータマイニングを実行することで、詳細な情報を獲得することができる。ユーザは、*Bombyx* trap データベースを使用して発現部位、強度および発育段階を指定することができ、インバース PCR 結果の位置情報は GBrowse にリンクされ、データマイニングを行うことができる。ユーザは、遺伝子モデルおよび関連遺伝子に対する検索を指定して、プロテオームデータベースから発現サイトや発育段階の情報を得ることができる。

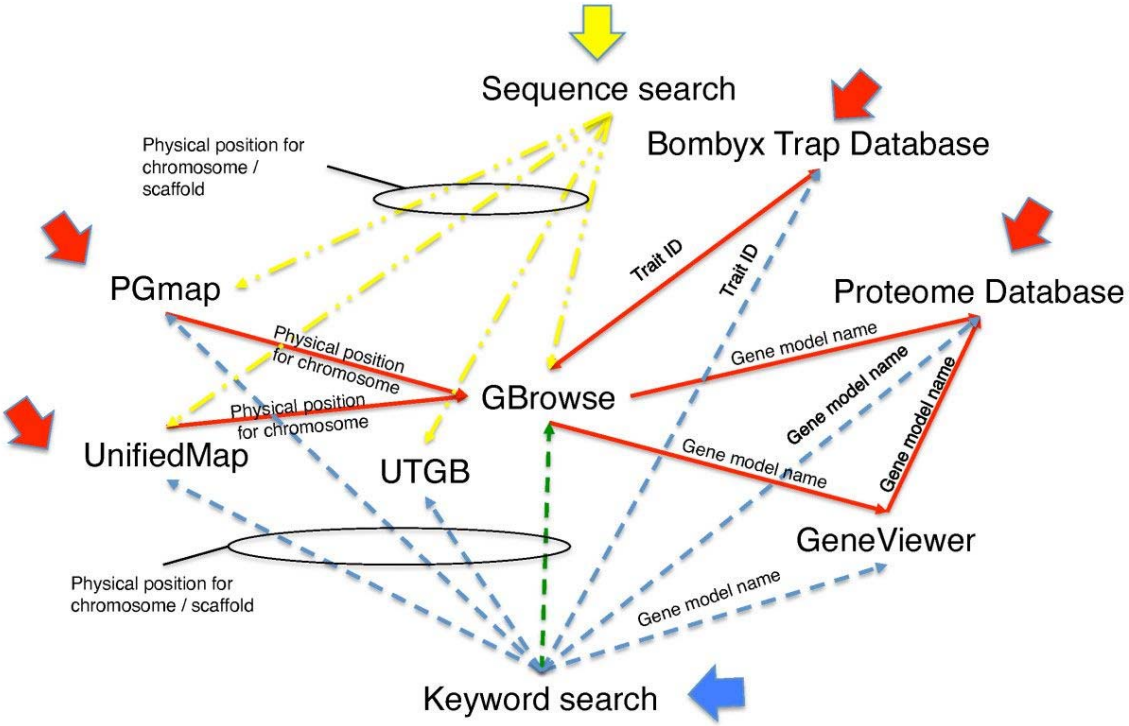


図 3-3 ブラウザ、ビューア、独立したデータベース間のリンク。　大きな赤い矢印は、ブラウザおよび関連するブラウザとデータベースからマイニングを表す。大きな青い矢印は、キーワード検索と関連するブラウザとデータベースからのマイニングを表す。大きな黄色の矢印は、配列検索および関連するブラウザおよびデータベースからのマイニングを表す。実線、破線、一点鎖線は、1

つのデータベース、ブラウザ、およびビューアから別のデータベース、ブラウザ、およびビューアへの情報の流れを表している。

3.5.2. キーワードサーチ、ポジションサーチ

キーワードとポジションサーチによるマイニング機能は、図 3-3 の青い矢印で示される。スキャフォールド、コンティグ、FPC コンティグ、SNP マーカー、形質マーカー、遺伝子モデル、EST / cDNA、BAC、BAC エンド、fosmid エンド、スキャフォールド上の位置などの情報を KAIKObase から直接検索することができる。さらに、このシステムは、スキャフォールドまたは染色体上の特定の領域の検索を提供する。検索結果は、染色体画像に検索された位置がマークされた形で表示され、4 つのブラウザ、1 つのビューア、および 2 つの独立したデータベースに対応するリンクとともにリスト表示される。

3.5.3. シークエンスサーチ

このクラスのデータマイニングへの入力は、図 3-3 の黄色の矢印で示される。クエリーは、核酸またはアミノ酸配列で、BLAST を使用したツールはスキャフォールドまたは染色体の位置を見出すことを可能にする。検索結果は、4 つのブラウザにリンクするためのボタンが付いたビットスコアの降順でソートされたデータとともにリスト表示される。表示結果から更に塩基位置範囲を狭めるもしくは広げるための開始位置および終了塩基位置入力ボックスを配列検索ページの上部に設置した。

3.6. 結論

効果的なデータマイニングと包括的なゲノム応用のためのカイコゲノム情報を提供するカイコゲノム統合データベース KAIKObase を開発した。KAIKObase に、カイコゲノム配列、ゲノム地図情報および EST データを統合した。
KAIKObase は、塩基配列、遺伝子、スキャフォルド、染色体の各段階のデータを 4 種類の MapViewer（PGmap、UnifiedMap、UTGB、GBrowse）、GeneViewer、配列検索、キーワード・位置検索を表示する。さらに、プロテオームデータ用

の KAIKO2DDB と遺伝子導入およびレポータデータ用の *Bombyx* trap データベースの統合により、KAIKObase の機能がさらに強化された。カイコの研究には、包括的なカイコゲノムデータベースが不可欠である。さまざまな最先端の視覚化ツールやマイニングツールを導入することで、KAIKObase は強力なデータリソースになり、鱗翅目の研究だけでなく、養蚕の改善や新しい害虫駆除の開発を容易にする。

第 4 章 結言

本論文は、ゲノム構築から解析とそれらのデータの閲覧システムに亘る一連の流れに沿って、国産ダイズ品種エンレイゲノムの構築・解析とカイコゲノム構築・解析から得られた情報を統合するためのデータベース構築を行った。

国産ダイズ品種エンレイゲノムの構築・解析では、栽培品種 Williams 82 に基づく基準ゲノム情報が入手可能であるが、国内品種との系統の違いにより国内品種の育種および改良に効率的に使用することができない。それゆえ、国内のダイズ品種エンレイのゲノム配列を構築・解析し、遺伝子モデルを構築した。その結果、アントシアニン・フラボノイド生合成系、系統発生解析などので、Williams 82 およびエンレイの遺伝子モデルの違いからゲノムの違いを明らかにした。エンレイゲノム配列データはダイズゲノムデータベースである DAIZUbase に統合し、ゲノム情報を利用したダイズ育種などに利用できるようになった。

カイコゲノム構築・解析から得られた情報を統合するための閲覧システム・データベース開発では、大きく多様なカイコのゲノム情報を効率的に利用するためのプラットフォームを構築した。カイコゲノム統合データベース KAIKObase には、ゲノム配列、遺伝子地図情報、発現配列タグデータ、遺伝子情報、エンハンサートラップ系統情報、タンパク質情報が統合されている。BAC と fosmid エンド、FPC、SNP マーカーと形質マーカー、ゲノムアノテーション、遺伝子モデルなどを連鎖地図と物理地図、スキャフォールドとコンティグ、配列とキーワード機能表示をまとめてブラウジングできるフレームワークを構築した。その結果、様々なゲノミクス情報を統合したゲノムインフラストラクチャーが開発され、養蚕の改善と新しい害虫駆除法の開発に利用できるようなった。

謝辞

ゲノムの研究をご教授頂いた特定非営利活動法人近畿アグリバイオ 北村實彬副理事長、東京農業大学 佐々木卓治教授、松本隆教授、西南大学 三田和英教授、国立研究開発法人農業・食品産業技術総合研究機構 片寄裕一研究管理役、山本公子ユニット長、門野敬子領域長、瀬筒秀樹ユニット長、長村吉晃博士、Baltazar A. Antonio 主席研究員、呉健忠博士、旧国立研究開発法人農業生物資源研究所 味村正博博士、故末次克行博士、University of Rhode Island Marian R Goldsmith 名誉博士、The Centre for DNA Fingerprinting and Diagnostics 故 Javaregowda Nagaraju 博士、西南大学 Qingyou Xia 教授、進化系統をご教授頂いた国立研究開発法人農業・食品産業技術総合研究機構 三中信宏博士、解析やプログラムなどのオペレーションをサポートして頂いた向井喜之氏、三菱スペース・ソフトウエア株式会社 並木信和氏、南博氏、伊川浩司氏、佐藤親忠氏、釜付香氏、元三菱スペース・ソフトウエア株式会社 大柳一博士、シークエンシングを担って頂いた国立研究開発法人農業・食品産業技術総合研究機構 金森裕之博士、栗田加奈子氏、矢野亮一博士、プロテオーム解析をご教授頂いた福井工業大学 小松節子教授、国立研究開発法人農業・食品産業技術総合研究機構 梶原英之主席研究員、ダイズ研究をご指導頂いた国立研究開発法人農業・食品産業技術総合研究機構 石本政男領域長、加賀秋人ユニット長、三菱スペース・ソフトウエア株式会社在職中に後期博士課程進学を許可頂いた三菱スペース・ソフトウエア株式会社 渡辺克昭氏、清水裕司氏、博士論文参考文献整理にご助力頂いた前橋工科大学 吉田綾子氏、最後に後期博士過程のご指導を賜った坂田克己教授に感謝する次第です。

参考文献

1. Kevles DJ, Hood LE: **The Code of codes : scientific and social issues in the Human Genome Project**. Cambridge, Mass.: Harvard University Press; 1992.
2. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications**. *Cell Res* 2011, **21**(3):381-395.
3. Cech TR, Steitz JA: **The noncoding RNA revolution-trashing old rules to forge new ones**. *Cell* 2014, **157**(1):77-94.
4. Parada LA, McQueen PG, Misteli T: **Tissue-specific spatial organization of genomes**. *Genome Biol* 2004, **5**(7):R44.
5. Dulbecco R: **A turning point in cancer research: sequencing the human genome**. *Science* 1986, **231**(4742):1055-1056.
6. Wheeler DA, Wang L: **From human genome to cancer genome: the first decade**. *Genome Res* 2013, **23**(7):1054-1062.
7. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, FitzHugh W, Fields C, Gocyne JD, Scott J, Shirley R, Liu L, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO and Venter JC: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**(5223):496-512.
8. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC: **The minimal gene complement of *Mycoplasma genitalium***. *Science* 1995, **270**(5235):397-403.
9. Ankeny RA, Leonelli S: **What's so special about model organisms?** *Studies in History and Philosophy of Science Part A* 2011, **42**(2):313-323.
10. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick

HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of *Escherichia coli* K-12**. *Science* 1997, **277**(5331):1453-1462.

11. Jacq C, Alt-Morbe J, Andre B, Arnold W, Bahr A, Ballesta JP, Bargues M, Baron L, Becker A, Biteau N, Blocker H, Blugeon C, Boskovic J, Brandt P, Bruckner M, Buitrago MJ, Coster F, Delaveau T, del Rey F, Dujon B, Eide LG, Garcia-Cantalejo JM, Goffeau A, Gomez-Peris A, Zaccaria P, et al.: **The nucleotide sequence of *Saccharomyces cerevisiae* chromosome IV**. *Nature* 1997, **387**(6632 Suppl):75-78.
12. C. elegans Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science* 1998, **282**(5396):2012-2018.
13. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. *Nature* 2004, **431**(7011):931-945.
14. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner

TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**(6915):520-562.

15. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy

M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, WoodageT, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**(5461):2185-2195.

16. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL: **The genome sequence of the malaria mosquito *Anopheles gambiae***. *Science* 2002, **298**(5591):129-149.

17. Honeybee Genome Sequencing Consortium, Collaborators: Weinstock GM RG, Gibbs RA, Weinstock GM, Robinson GE,, Worley KC EJ, Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P, Elsik, CG EJ, Hartfelder K, Hunt GJ,

Robertson HM, Robinson GE, Maleszka R,, Weinstock GM WK, Zdobnov EM, Hartfelder K, Amdam GV, Bitondi MM, Collins , AM CA, Evans JD, Lattorff MG, Lobo CH, Moritz RF, Nunes FM, Page RE Jr,, Simões ZL WD, Carninci P, Fukuda S, Hayashizaki Y, Kai C, Kawai J,, Sakazume N SD, Tagami M, Maleszka R, Amdam GV, Albert S, Baggerman G,, Beggs KT BG, Cazzamali G, Cohen M, Drapeau MD, Eisenhardt D, Emore C, Ewing, MA FS, Forêt S, Grimmelikhuijzen CJ, Hauser F, Hummon AB, Hunt GJ,, Huybrechts J JA, Kadowaki T, Kaplan N, Kucharski R, Leboulle G, Linial M, , Littleton JT MA, Page RE Jr, Robertson HM, Robinson GE, Richmond TA,, Rodriguez-Zas SL RE, Sattelle DB, Schlipalius D, Schoofs L, Shemesh Y,, Sweedler JV VR, Verleyen P, Vierstraete E, Williamson MR, Beye M, Ament, SA BS, Corona M, Dearden PK, Dunn WA, Elekonich MM, Elsik CG, Forêt S,, Fujiyuki T GE, Gempe T, Hasselmann M, Kadowaki T, Kage E, Kamikouchi , A KT, Kucharski R, Kunieda T, Lorenzen M, Maleszka R, Milshina NV, Morioka, M OK, Overbeek R, Page RE Jr, Robertson HM, Robinson GE, Ross CA, Schioett, M ST, Takeuchi H, Toth AL, Willis JH, Wilson MJ, Robertson HM, Zdobnov EM,, Bork P EC, Gordon KH, Letunic I, Hackett K, Peterson J, Felsenfeld A,, Guyer M SM, Agarwala R, Cornuet JM, Elsik CG, Emore C, Hunt GJ, Monnerot, M MF, Reese JT, Schlipalius D, Vautrin D, Weaver DB, Gillespie JJ, Cannone, JJ GR, Johnston JS, Elsik CG, Cazzamali G, Eisen MB, Grimmelikhuijzen CJ,, Hauser F HA, Iyer VN, Iyer V, Kosarev P, Mackey AJ, Maleszka R, Reese JT,, Richmond TA RH, Solovyev V, Souvorov A, Sweedler JV, Weinstock GM,, Willliamson MR ZE, Evans JD, Aronstein KA, Bilikova K, Chen YP, Clark, AG DL, Gelbart WM, Hetru C, Hultmark D, Imler JL, Jiang H, Kanost M,, Kimura K LB, Lopez DL, Simuth J, Thompson GJ, Zou Z, De Jong P,, Sodergren E CM, Milosavljevic A, Johnston JS, Osoegawa K, Richards S, Shu , CL WG, Elsik CG, Duret L, Elhaik E, Graur D, Reese JT, Robertson HM,, Robertson HM EC, Maleszka R, Weaver DB, Amdam GV, Anzola JM, Campbell KS, , Childs KL CD, Crosby MA, Dickens CM, Elsik CG, Gordon KH, Grametes LS,, Grozinger CM JP, Jorda M, Ling X, Matthews BB, Miller J, Milshina NV,, Mizzen C PM, Reese JT, Reid JG, Robertson HM, Robinson GE, Russo SM,, Schroeder AJ SPS, Wang Y, Zhou P, Robertson HM, Agarwala R, Elsik CG,, Milshina NV RJ, Weaver DB, Worley KC, Childs KL, Dickens CM, Elsik CG,, Gelbart WM JH, Kitts P, Milshina NV, Reese JT, Ruef

B, Russo SM,, Venkatraman A WG, Zhang L, Zhou P, Johnston JS, Aquino-Perez G,, Cornuet JM MM, Solignac M, Vautrin D, Whitfield CW, Behura S, Berlocher , SH CA, Gibbs RA, Johnston JS, Sheppard WS, Smith DR, Suarez AV, Tsutsui, ND WD, Wei X, Wheeler D, Weinstock GM, Worley KC, Havlak P, Li B, Liu Y, , Sodergren E ZL, Beye M, Hasselmann M, Jolivet A, Lee S, Nazareth LV, Pu LL,, Thorn R WG, Stolc V, Robinson GE, Maleszka R, Newman T, Samanta M,, Tongprasit WA AK, Claudianos C, Berenbaum MR, Biswas S, de Graaf DC,, Feyereisen R JR, Oakeshott JG, Ranson H, Schuler MA, Muzny D, Gibbs RA, , Weinstock GM CJ, Davis C, Dinh H, Gill R, Hernandez J, Hines S, Hume J,, Jackson L KC, Lewis L, Miner G, Morgan M, Nazareth LV, Nguyen N, Okwuonu G,, Paul H RS, Santibanez J, Savery G, Sodergren E, Svatek A, Villasana D,, R. W: **Insights into social insects from the genome of the honeybee *Apis mellifera***. *Nature* 2006, **443**(7114):931-949.

18. Tribolium Genome Sequencing Consortium, Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Beeman RW, Brown SJ, Bucher G, Friedrich M, Grimmelikhuijzen CJ, Klingler M, Lorenzen M, Richards S, Roth S, Schroder R, Tautz D, Zdobnov EM, Muzny D, Gibbs RA, Weinstock GM, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Richards S, Ruiz SJ, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Tautz D, Vattahil S, Villasana D, White CS, Wright R, Park Y, Beeman RW, Lord J, Oppert B, Lorenzen M, Brown S, Wang L, Savard J, Tautz D, Richards S, Weinstock G, Gibbs RA, Liu Y, Worley K, Weinstock G, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beeman RW, Beidler J, Brown SJ, Demuth JP, Drury DW, Du YZ, Fujiwara H, Lorenzen M, Maselli V, Osanai M, Park Y, Robertson HM, Tu Z, Wang JJ, Wang S, Richards S, Song H, Zhang L, Sodergren E, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Pruitt K, Sapojnikov V, Souvorov A, Mackey AJ, Waterhouse RM, Wyder S, Zdobnov EM, Zdobnov EM, Wyder S, Kriventseva

EV, Kadowaki T, Bork P, Aranda M, Bao R, Beermann A, Berns N, Bolognesi R, Bonneton F, Bopp D, Brown SJ, Bucher G, Butts T, Chaumot A, Denell RE, Ferrier DE, Friedrich M, Gordon CM, Jindra M, Klingler M, Lan Q, Lattorff HM, Laudet V, von Levetsow C, Liu Z, Lutz R, Lynch JA, da Fonseca RN, Posnien N, Reuter R, Roth S, Savard J, Schinko JB, Schmitt C, Schoppmeier M, Schroder R, Shippy TD, Simonnet F, Marques-Souza H, Tautz D, Tomoyasu Y, Trauner J, Van der Zee M, Vervoort M, Wittkopp N, Wimmer EA, Yang X, Jones AK, Sattelle DB, Ebert PR, Nelson D, Scott JG, Beeman RW, Muthukrishnan S, Kramer KJ, Arakane Y, Beeman RW, Zhu Q, Hogenkamp D, Dixit R, Oppert B, Jiang H, Zou Z, Marshall J, Elpidina E, Vinokurov K, Oppert C, Zou Z, Evans J, Lu Z, Zhao P, Sumathipala N, Altincicek B, Vilcinskas A, Williams M, Hultmark D, Hetru C, Jiang H, Grimmelikhuijzen CJ, Hauser F, Cazzamali G, Williamson M, Park Y, Li B, Tanaka Y, Predel R, Neupert S, Schachtner J, Verleyen P, Raible F, Bork P, Friedrich M, Walden KK, Robertson HM, Angeli S, Foret S, Bucher G, Schuetz S, Maleszka R, Wimmer EA, Beeman RW, Lorenzen M, Tomoyasu Y, Miller SC, Grossmann D, Bucher G: **The genome of the model beetle and pest *Tribolium castaneum***. *Nature* 2008, **452**(7190):949-955.

19. International Silkworm Genome Consortium: **The genome of a lepidopteran model insect, the silkworm *Bombyx mori***. *Insect Biochem Mol Biol* 2008, **38**(12):1036-1045.

20. Theologis A, Ecker JR, Palm CJ, Federspiel NA, Kaul S, White O, Alonso J, Altafi H, Araujo R, Bowman CL, Brooks SY, Buehler E, Chan A, Chao Q, Chen H, Cheuk RF, Chin CW, Chung MK, Conn L, Conway AB, Conway AR, Creasy TH, Dewar K, Dunn P, Etgu P, Feldblyum TV, Feng J, Fong B, Fujii CY, Gill JE, Goldsmith AD, Haas B, Hansen NF, Hughes B, Huizar L, Hunter JL, Jenkins J, Johnson-Hopson C, Khan S, Khaykin E, Kim CJ, Koo HL, Kremenetskaia I, Kurtz DB, Kwan A, Lam B, Langin-Hooper S, Lee A, Lee JM, Lenz CA, Li JH, Li Y, Lin X, Liu SX, Liu ZA, Luros JS, Maiti R, Marziali A, Militscher J, Miranda M, Nguyen M, Nierman WC, Osborne BI, Pai G, Peterson J, Pham PK, Rizzo M, Rooney T, Rowley D, Sakano H, Salzberg SL, Schwartz JR, Shinn P, Southwick AM, Sun H, Tallon LJ, Tambunga G, Toriumi MJ, Town CD, Utterback T, Van Aken S, Vaysberg M, Vysotskaia VS, Walker M, Wu D, Yu G, Fraser CM, Venter

JC, Davis RW: **Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana***. *Nature* 2000, **408**(6814):816-820.

21. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.

22. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.

23. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P, French-Italian Public Consortium for Grapevine Genome C: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**(7161):463-467.

24. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S: **Genome structure of the legume, *Lotus japonicus***. *DNA Res* 2008, **15**(4):227-239.

25. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)**. *Nature* 2008, **452**(7190):991-996.

26. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS: **The *Sorghum bicolor* genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.

27. Vielle-Calzada JP, Martinez de la Vega O, Hernandez-Guzman G, Ibarra-Laclette E, Alvarez-Mejia C, Vega-Arreguin JC, Jimenez-Moraila B, Fernandez-Cortes A, Corona-Armenta G, Herrera-Estrella L, Herrera-Estrella A: **The Palomero genome suggests metal effects on domestication**. *Science* 2009, **326**(5956):1078.

28. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EA, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan, Wu Z, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S: **The genome of the cucumber, *Cucumis sativus* L**. *Nat Genet* 2009, **41**(12):1275-1281.

29. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA: **Genome sequence of the palaeopolyploid soybean**. *Nature* 2010, **463**(7278):178-183.

30. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD: **Draft genome sequence of the oilseed species *Ricinus communis***. *Nat Biotechnol* 2010, **28**(9):951-956.

31. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S,

Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R: **The genome of the domesticated apple (*Malus* x *domestica* Borkh.)**. *Nat Genet* 2010, **42**(10):833-839.

32. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Wang J, Wang X, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Liu B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, Lin C, Edwards D, Mu D, Shen D, Soumpourou E, Li F, Fraser F, Conant G, Lassalle G, King GJ, Bonnema G, Tang H, Wang H, Belcram H, Zhou H, Hirakawa H, Abe H, Guo H, Wang H, Jin H, Parkin IA, Batley J, Kim JS, Just J, Li J, Xu J, Deng J, Kim JA, Li J, Yu J, Meng J, Wang J, Min J, Poulain J, Wang J, Hatakeyama K, Wu K, Wang L, Fang L, Trick M, Links MG, Zhao M, Jin M, Ramchiary N, Drou N, Berkman PJ, Cai Q, Huang Q, Li R, Tabata S, Cheng S, Zhang S, Zhang S, Huang S, Sato S, Sun S, Kwon SJ, Choi SR, Lee TH, Fan W, Zhao X, Tan X, Xu X, Wang Y, Qiu Y, Yin Y, Li Y, Du Y, Liao Y, Lim Y, Narusaka Y, Wang Y, Wang Z, Li Z, Wang Z, Xiong Z, Zhang Z, Brassica rapa Genome Sequencing Project C: **The genome of the mesopolyploid crop species *Brassica rapa***. *Nat Genet* 2011, **43**(10):1035-1039.

33. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang SP, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, Jackson SA: **Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers**. *Nat Biotechnol* 2011, **30**(1):83-89.

34. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA,

Troggio M, Viola R, Ashman TL, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW, Jr., Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Lopez Girona E, Zdepski A, Wang W, Kerstetter RA, Schwab W, Korban SS, Davik J, Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM: **The genome of woodland strawberry (*Fragaria vesca*)**. *Nat Genet* 2011, **43**(2):109-116.

35. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Berard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C: **The genome of *Theobroma cacao***. *Nat Genet* 2011, **43**(2):101-108.

36. Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, Spannagl M, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Berges H, Bidwell S, Bisseling T, Choisne N, Couloux A, Denny R, Deshpande S, Dai X, Doyle JJ, Dudez AM, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, Gonzalez AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong DH, Jing Y, Jocker A, Kenton SM, Kim DJ, Klee K, Lai H, Lang C, Lin S, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun JH, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin B, Qu C, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi R, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang BB, Wang K, Wang M, Wang X, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing Y, Yang L, Yao Z, Ying F,

Zhai J, Zhou L, Zuber A, Denarie J, Dixon RA, May GD, Schwartz DC, Rogers J, Quetier F, Town CD, Roe BA: **The *Medicago* genome provides insight into the evolution of rhizobial symbioses**. *Nature* 2011, **480**(7378):520-524.

37. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, Al-Azwani EK, Chaluvadi S, Pontaroli AC, DeBarry J, Arondel V, Ohlrogge J, Saie IJ, Suliman-Elmeer KM, Bennetzen JL, Kruegger RR, Malek JA: **De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*)**. *Nat Biotechnol* 2011, **29**(6):521-527.

38. Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DM, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo B, Mejia N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, Herrera Mdel R, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SE, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa AN, Geoffroy M, Lundback S, DellaPenna D, Buell CR, Sharma SK, Marshall DF, Waugh R, Bryan GJ, Destefanis M, Nagy I, Milbourne D, Thomson SJ, Fiers M, Jacobs JM, Nielsen KL, Sonderkaer M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CW, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Hekkert B, Goverse A, van Ham RC, Visser RG: **Genome sequence and analysis of the tuber crop potato**. *Nature* 2011, **475**(7355):189-195.

39. Tomato Genome Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**(7400):635-641.

40. Garcia-Mas J, Benjak A, Sanseverino W, Bourgeois M, Mir G, Gonzalez VM, Henaff E, Camara F, Cozzuto L, Lowy E, Alioto T, Capella-Gutierrez S, Blanca J, Canizares J, Ziarsolo P, Gonzalez-Ibeas D, Rodriguez-Moreno L, Droege M, Du L, Alvarez-Tejado M, Lorente-Galdos B, Mele M, Yang L, Weng Y, Navarro A, Marques-Bonet T, Aranda MA, Nuez F, Pico B, Gabaldon T, Roma G, Guigo R, Casacuberta JM, Arus P, Puigdomenech P: **The genome of melon (*Cucumis melo* L.)**. *Proc Natl Acad Sci U S A* 2012, **109**(29):11872-11877.

41. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, Da Silva C, Jabbari K, Cardi C, Poulain J, Souquet M, Labadie K, Jourda C, Lengelle J, Rodier-Goud M, Alberti A, Bernard M, Correa M, Ayyampalayam S, McKain MR, Leebens-Mack J, Burgess D, Freeling M, Mbeguie AMD, Chabannes M, Wicker T, Panaud O, Barbosa J, Hribova E, Heslop-Harrison P, Habas R, Rivallan R, Francois P, Poiron C, Kilian A, Burthia D, Jenny C, Bakry F, Brown S, Guignon V, Kema G, Dita M, Waalwijk C, Joseph S, Dievart A, Jaillon O, Leclercq J, Argout X, Lyons E, Almeida A, Jeridi M, Dolezel J, Roux N, Risterucci AM, Weissenbach J, Ruiz M, Glaszmann JC, Quetier F, Yahiaoui N, Wincker P: **The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants**. *Nature* 2012, **488**(7410):213-217.

42. International Peach Genome Initiative, Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel LA, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, Goodstein DM, Xuan P, Del Fabbro C, Aramini V, Copetti D, Gonzalez S, Horner DS, Falchi R, Lucas S, Mica E, Maldonado J, Lazzari B, Bielenberg D, Pirona R, Miculan M, Barakat A, Testolin R, Stella A, Tartarini S, Tonutti P, Arus P, Orellana A, Wells C, Main D, Vizzotto G, Silva H, Salamini F, Schmutz J, Morgante M, Rokhsar DS: **The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution**. *Nat Genet* 2013, **45**(5):487-494.

43. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, Tonosaki K, Shirasawa S, Fukushima A, Yokoi S, Takahata Y, Kakizaki T, Ishida M, Okamoto S, Sakamoto K, Shirasawa K, Tabata S, Nishio T: **Draft sequences of the radish (*Raphanus sativus* L.) genome**. *DNA Res* 2014, **21**(5):481-490.

44. Mitsui Y, Shimomura M, Komatsu K, Namiki N, Shibata-Hatta M, Imai M, Katayose Y, Mukai Y, Kanamori H, Kurita K, Kagami T, Wakatsuki A, Ohyanagi H, Ikawa H, Minaka N, Nakagawa K, Shiwa Y, Sasaki T: **The radish genome and comprehensive gene expression profile of tuberous root formation and development**. *Sci Rep* 2015, **5**:10835.

45. Shimomura M, Kanamori H, Komatsu S, Namiki N, Mukai Y, Kurita K, Kamatsuki K, Ikawa H, Yano R, Ishimoto M, Kaga A, Katayose Y: **The *Glycine***

max **cv. Enrei Genome for Improvement of Japanese Soybean Cultivars**. *Int J Genomics* 2015, **2015**:358127.

46. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA**. *Nature* 1977, **265**(5596):687-695.
47. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**(12):5463-5467.
48. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H: **Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction**. *Cold Spring Harb Symp Quant Biol* 1986, **51 Pt 1**:263-273.
49. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE: **Fluorescence detection in automated DNA sequence analysis**. *Nature* 1986, **321**(6071):674-679.
50. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK: **The origin of the Haitian cholera outbreak strain**. *N Engl J Med* 2011, **364**(1):33-42.
51. Mikheyev AS, Tin MM: **A first look at the Oxford Nanopore MinION sequencer**. *Mol Ecol Resour* 2014, **14**(6):1097-1102.
52. Schadt EE, Turner S, Kasarskis A: **A window into third-generation sequencing**. *Hum Mol Genet* 2010, **19**(R2):R227-240.
53. Green P: **PHRAP documentation**. *http://www.phrap.org* 1994.
54. Sutton GG, White O, Adams MD, Kerlavage AR: **TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects**. *Genome Science and Technology* 1995, **1**(1):9-19.
55. Nagarajan N, Pop M: **Sequence assembly demystified**. *Nat Rev Genet* 2013, **14**(3):157-167.
56. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC: **A whole-genome assembly of *Drosophila***. *Science* 2000, **287**(5461):2196-2204.

57. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler**. *Genome Res* 2002, **12**(1):177-189.

58. Roche: **Technical Bulletin GS FLX+ System & GS FLX System, Installation of 454 Sequencing System Software v2.8**. *Roche* 2012.

59. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.

60. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome Res* 2009, **19**(6):1117-1123.

61. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome Res* 2010, **20**(2):265-272.

62. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE**. *Bioinformatics* 2011, **27**(4):578-579.

63. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

64. Kan YW, Dozy AM: **Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation**. *Proc Natl Acad Sci U S A* 1978, **75**(11):5631-5635.

65. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers**. *Nucleic Acids Res* 1990, **18**(22):6531-6535.

66. Ligtenberg MJ, Gennissen AM, Vos HL, Hilkens J: **A single nucleotide polymorphism in an exon dictates allele dependent differential splicing of episialin mRNA**. *Nucleic Acids Res* 1991, **19**(2):297-301.

67. Zietkiewicz E, Rafalski A, Labuda D: **Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification**. *Genomics* 1994, **20**(2):176-183.

68. Urquhart A, Kimpton CP, Downes TJ, Gill P: **Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers**. *Int J Legal Med* 1994, **107**(1):13-20.

69. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al.: **AFLP: a new technique for DNA fingerprinting**. *Nucleic Acids Res* 1995, **23**(21):4407-4414.
70. Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W: **Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat**. *Theor Appl Genet* 2002, **104**(2-3):399-407.
71. Weil MM, Pershad R, Wang R, Zhao S: **Use of BAC end sequences for SNP discovery**. *Methods Mol Biol* 2004, **256**:1-6.
72. Yamamoto K, Narukawa J, Kadokuda K, Nohata J, Sasanuma M, Suetsugu Y, Banno Y, Fujii H, Goldsmith MR, Mita K: **Construction of a single nucleotide polymorphism linkage map for the silkworm, *Bombyx mori*, based on bacterial artificial chromosome end sequences**. *Genetics* 2006, **173**(1):151-161.
73. Clevenger J, Chavarro C, Pearl SA, Ozias-Akins P, Jackson SA: **Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations**. *Mol Plant* 2015, **8**(6):831-846.
74. Lee M: **DNA Markers and Plant Breeding Programs**. In: *Advances in Agronomy*. Edited by Sparks DL, vol. 55: Academic Press; 1995: 265-344.
75. NCBI: **Section 9.1 Molecular Definition of a Gene.** *https://www.ncbi.nlm.nih.gov/books/NBK21640/*
76. ABRC: **Gene model in Using the Gene Search Results**. *https://www.arabidopsis.org/help/helppages/generesu.jsp*.
77. Brent MR: **How does eukaryotic gene prediction work?** *Nat Biotechnol* 2007, **25**(8):883-885.
78. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
79. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *J Mol Biol* 1981, **147**(1):195-197.
80. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
81. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions**. *Genome Biol* 2013, **14**(4):R36.

82. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

83. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-515.

84. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**(1):78-94.

85. Salamov AA, Solovyev VV: **Ab initio gene finding in *Drosophila* genomic DNA**. *Genome Res* 2000, **10**(4):516-522.

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

87. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**(7):644-652.

88. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A: **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis**. *Nat Protoc* 2013, **8**(8):1494-1512.

89. Eeckman FH, Durbin R: **ACeDB and macace**. *Methods Cell Biol* 1995, **48**:583-605.

90. Sakata K, Antonio BA, Mukai Y, Nagasaki H, Sakai Y, Makino K, Sasaki T: **INE: a rice genome database with an integrated map view**. *Nucleic Acids Res* 2000, **28**(1):97-101.

91. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank**. *Nucleic Acids Res* 2002, **30**(1):17-20.

92. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology**. *Nucleic Acids Res* 2003, **31**(1):28-33.

93. Youens-Clark K, Faga B, Yap IV, Stein L, Ware D: **CMap 1.01: a comparative mapping application for the Internet**. *Bioinformatics* 2009, **25**(22):3040-3042.
94. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**(1):38-41.
95. Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T: **Ensembl 2004**. *Nucleic Acids Res* 2004, **32**(Database issue):D468-470.
96. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.
97. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database**. *Genome Res* 2002, **12**(10):1599-1610.
98. Ahsan B, Kobayashi D, Yamada T, Kasahara M, Sasaki S, Saito TL, Nagayasu Y, Doi K, Nakatani Y, Qu W, Jindo T, Shimada A, Naruse K, Toyoda A, Kuroki Y, Fujiyama A, Sasaki T, Shimizu A, Asakawa S, Shimizu N, Hashimoto S, Yang J, Lee Y, Matsushima K, Sugano S, Sakaizumi M, Narita T, Ohishi K, Haga S, Ohta F, Nomoto H, Nogata K, Morishita T, Endo T, Shin IT, Takeda H, Kohara Y, Morishita S: **UTGB/medaka: genomic resource database for medaka biology**. *Nucleic Acids Res* 2008, **36**(Database issue):D747-752.
99. NCBI: **NCBI's Genome Data Viewer (GDV) to replace Map Viewer**. *https://ncbiinsights.ncbi.nlm.nih.gov/2017/10/24/ncbis-genome-data-viewer-gdv-to-replace-map-viewer/*.

100. Saito TL, Yoshimura J, Sasaki S, Ahsan B, Sasaki A, Kuroshu R, Morishita S: **UTGB toolkit for personalized genome browsers**. *Bioinformatics* 2009, **25**(15):1856-1861.

101. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.

102. Mostaguir K, Hoogland C, Binz PA, Appel RD: **The Make 2D-DB II package: conversion of federated two-dimensional gel electrophoresis databases into a relational format and interconnection of distributed databases**. *Proteomics* 2003, **3**(8):1441-1444.

103. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome Res* 2005, **15**(10):1451-1455.

104. Kaga A, Shimizu T, Watanabe S, Tsubokura Y, Katayose Y, Harada K, Vaughan DA, Tomooka N: **Evaluation of soybean germplasm conserved in NIAS genebank and development of mini core collections**. *Breed Sci* 2012, **61**(5):566-592.

105. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser**. *Genome Res* 2004, **14**(5):951-955.

106. Harris TW, Lee R, Schwarz E, Bradnam K, Lawson D, Chen W, Blasier D, Kenny E, Cunningham F, Kishore R, Chan J, Muller HM, Petcherski A, Thorisson G, Day A, Bieri T, Rogers A, Chen CK, Spieth J, Sternberg P, Durbin R, Stein LD: **WormBase: a cross-species database for comparative genomics**. *Nucleic Acids Res* 2003, **31**(1):133-137.

107. FlyBase Consortium: **The FlyBase database of the Drosophila genome projects and community literature**. *Nucleic Acids Res* 2003, **31**(1):172-175.

108. Yamamoto K, Nohata J, Kadono-Okuda K, Narukawa J, Sasanuma M, Sasanuma S, Minami H, Shimomura M, Suetsugu Y, Banno Y, Osoegawa K, de Jong PJ, Goldsmith MR, Mita K: **A BAC-based integrated linkage map of the silkworm *Bombyx mori***. *Genome Biol* 2008, **9**(1):R21.

109. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, Kawasaki H, Kadono-Okuda K, Yamamoto K, Suzuki MG, Shimada T, Goldsmith MR, Maeda S: **The**

construction of an EST database for *Bombyx mori* and its application. *Proc Natl Acad Sci U S A* 2003, **100**(24):14121-14126.

110. Kajiwara H, Nakane K, Piyang J, Imamaki A, Ito Y, Togasaki F, Kotake T, Murai H, Nakamura M, Mita K, Nomura R, Shimizu Y, Shimomura M, Ishizaka M: **Draft of silkworm proteome database**. *Journal of Electrophoresis* 2006, **50**(3,4):39-41.

111. Shimomura M, Minami H, Suetsugu Y, Ohyanagi H, Satoh C, Antonio B, Nagamura Y, Kadono-Okuda K, Kajiwara H, Sezutsu H, Nagaraju J, Goldsmith MR, Xia Q, Yamamoto K, Mita K: **KAIKObase: an integrated silkworm genome database and data mining tool**. *BMC Genomics* 2009, **10**:486.

112. Hajika M: **Present state and prospect of soybean production and soybean breeding in Japan**. In: *Proceedings of the 14th NIAS International Workshop on Genetic Resources and Comparative Genomics of Legumes (Glycine and Vigna): 2011*: National Institute of Agrobiological Sciences; 2011: 49-52.

113. Qiu L, Chang R: **The origin and history of soybean**. In. Wallingford: CABI; 2010: 1-23.

114. Lee GA, Crawford GW, Liu L, Sasaki Y, Chen X: **Archaeological soybean (*Glycine max*) in East Asia: does size matter?** *PLoS One* 2011, **6**(11):e26720.

115. Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, Kim DS, Lee YS, Park D, Ma J, Kim WY, Kim BC, Park S, Lee KA, Kim DH, Kim KH, Shin JH, Jang YE, Kim KD, Liu WX, Chaisan T, Kang YJ, Lee YH, Kim KH, Moon JK, Schmutz J, Jackson SA, Bhak J, Lee SH: **Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome**. *Proc Natl Acad Sci U S A* 2010, **107**(51):22032-22037.

116. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang SS, Zuo Q, Shi XH, Li YF, Zhang WK, Hu Y, Kong G, Hong HL, Tan B, Song J, Liu ZX, Wang Y, Ruan H, Yeung CK, Liu J, Wang H, Zhang LJ, Guan RX, Wang KJ, Li WB, Chen SY, Chang RZ, Jiang Z, Jackson SA, Li R, Qiu LJ: **De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits**. *Nat Biotechnol* 2014, **32**(10):1045-1052.

117. Bernard R, Cremeens C: **Registration of 'Williams 82'soybean**. *Crop Science* 1988, **28**(6):1027-1028.

118. 農林水産省生産局: **水陸稲･麦類･大豆奨励品種特性表 平成 14 年 3 月**. *農業技術協会*, 2002.

119. Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH: **Construction of plant bacterial artificial chromosome (BAC) libraries: an illustrated guide**. *Journal of Agricultural genomics* 2000, **5**:1-100.

120. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

121. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

122. NIG: **NGS Surfer's wiki**. *http://cell-innovation.nig.ac.jp/wiki/tiki-indexphp?page=samtools#mpileup*.

123. NCBI: **BLAST ftp site**. *ftp://ftp.ncbi.nih.gov/blast/*.

124. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W435-439.

125. Smit A, Hubley, R & Green, P.: **RepeatMasker Open-3.0.** *http://www.repeatmasker.org* 1996-2010.

126. Du J, Grant D, Tian Z, Nelson RT, Zhu L, Shoemaker RC, Ma J: **SoyTEdb: a comprehensive database of transposable elements in the soybean genome**. *BMC Genomics* 2010, **11**:113.

127. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.

128. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E: **The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res* 2012, **40**(Database issue):D1202-1210.

129. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL: **The *Arabidopsis lyrata* genome**

sequence and the basis of rapid genome size change. *Nat Genet* 2011, **43**(5):476-481.

130. NARO: **Os-Nipponbare-Reference-IRGSP-1.0**. *http://rapdb.dna.affrc.go.jp/download/archive/irgsp1/IRGSP-10_protein_2014-06-25fastagz*.

131. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.

132. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega**. *Mol Syst Biol* 2011, **7**:539.

133. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0**. *Mol Biol Evol* 2013, **30**(12):2725-2729.

134. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood**. *Mol Biol Evol* 2007, **24**(8):1586-1591.

135. Thorne JL, Kishino H: **Divergence time and evolutionary rate estimation with multilocus data**. *Syst Biol* 2002, **51**(5):689-702.

136. FigTree: *http://tree.bio.ed.ac.uk/software/figtree/*.

137. Komatsu S, Han C, Nanjo Y, Altaf-Un-Nahar M, Wang K, He D, Yang P: **Label-free quantitative proteomic analysis of abscisic acid effect in early-stage soybean under flooding**. *J Proteome Res* 2013, **12**(11):4769-4784.

138. Brosch M, Yu L, Hubbard T, Choudhary J: **Accurate and sensitive peptide identification with Mascot Percolator**. *J Proteome Res* 2009, **8**(6):3176-3181.

139. Shinoda K, Tomita M, Ishihama Y: **emPAI Calc--for the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry**. *Bioinformatics* 2010, **26**(4):576-577.

140. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S: **Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana***. *Proc Natl Acad Sci U S A* 2010, **107**(43):18724-18728.

141. Lavin M, Herendeen PS, Wojciechowski MF: **Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary**. *Syst Biol* 2005, **54**(4):575-594.

142. Cho YB, Jones SI, Vodkin L: **The transition from primary siRNAs to amplified secondary siRNAs that regulate chalcone synthase during development of *Glycine max* seed coats**. *PLoS One* 2013, **8**(10):e76954.

143. Clough SJ, Tuteja JH, Li M, Marek LF, Shoemaker RC, Vodkin LO: **Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of *CHS* genes comprising the I locus**. *Genome* 2004, **47**(5):819-831.

144. Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO: **Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats**. *Plant Cell* 2009, **21**(10):3063-3077.

145. Senda M, Masuta C, Ohnishi S, Goto K, Kasai A, Sano T, Hong JS, MacFarlane S: **Patterning of virus-infected *Glycine max* seed coat is associated with suppression of endogenous silencing of chalcone synthase genes**. *Plant Cell* 2004, **16**(4):807-818.

146. Tuteja JH, Clough SJ, Chan WC, Vodkin LO: **Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max***. *Plant Cell* 2004, **16**(4):819-835.

147. Tsukada Y, Kitamura K, Harada K, Kaizuma N: **Genetic analysis of subunits of two major storage proteins (β-conglycinin and glycinin) in soybean seeds**. *Japanese Journal of Breeding* 1986, **36**(4):390-400.

148. Krishnan HB, Natarajan SS, Mahmoud AA, Nelson RL: **Identification of glycinin and beta-conglycinin subunits that contribute to the increased protein content of high-protein soybean lines**. *J Agric Food Chem* 2007, **55**(5):1839-1845.

149. Dunwell JM: **Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins**. *Biotechnol Genet Eng Rev* 1998, **15**:1-32.

150. Yaklich RW: **beta-Conglycinin and glycinin in high-protein soybean seeds**. *J Agric Food Chem* 2001, **49**(2):729-735.

151. Shutov AD, Kakhovskaya IA, Bastrygina AS, Bulmaga VP, Horstmann C, Muntz K: **Limited proteolysis of beta-conglycinin and glycinin, the 7S and 11S storage globulins from soybean [*Glycine max* (L.) Merr.]. Structural and evolutionary implications**. *Eur J Biochem* 1996, **241**(1):221-228.

152. Yoshikawa T, Utsumi S, Fukuda T, Okumoto Y, Sayama T, Tanisaka T: **Identification of genes controlling the contents of seed storage proteins in soybean-identification and functional analysis of the quantitative trait locus qPro1**. *Soy Protein Research, Japan* 2009, **12**:27-32.

153. Tsubokura Y, Hajika M, Kanamori H, Xia Z, Watanabe S, Kaga A, Katayose Y, Ishimoto M, Harada K: **The beta-conglycinin deficiency in wild soybean is associated with the tail-to-tail inverted repeat of the alpha-subunit genes**. *Plant Mol Biol* 2012, **78**(3):301-309.

154. McGlew K, Shaw V, Zhang M, Kim RJ, Yang W, Shorrosh B, Suh MC, Ohlrogge J: **An annotated database of *Arabidopsis* mutants of acyl lipid metabolism**. *Plant Cell Rep* 2015, **34**(4):519-532.

155. Katayose Y, Kanamori H, Shimomura M, Ohyanagi H, Ikawa H, Minami H, Shibata M, Ito T, Kurita K, Ito K, Tsubokura Y, Kaga A, Wu J, Matsumoto T, Harada K, Sasaki T: **DaizuBase, an integrated soybean genome database including BAC-based physical maps**. *Breed Sci* 2012, **61**(5):661-664.

156. NARO: **DAIZUbase**. *https://daizubase.daizu.dna.affrc.go.jp/*.

157. Tamura T, Thibert C, Royer C, Kanda T, Abraham E, Kamba M, Komoto N, Thomas JL, Mauchamp B, Chavancy G, Shirk P, Fraser M, Prudhomme JC, Couble P: **Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector**. *Nat Biotechnol* 2000, **18**(1):81-84.

158. Tomita M, Munetsuna H, Sato T, Adachi T, Hino R, Hayashi M, Shimizu K, Nakamura N, Tamura T, Yoshizato K: **Transgenic silkworms produce recombinant human type III procollagen in cocoons**. *Nat Biotechnol* 2003, **21**(1):52-56.

159. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H, Namiki N, Kitagawa M, Yamashita H, Yasukochi Y, Kadono-Okuda K, Yamamoto K, Ajimura M, Ravikumar G, Shimomura M, Nagamura Y, Shin IT, Abe H, Shimada T, Morishita S, Sasaki T: **The genome sequence of silkworm, *Bombyx mori***. *DNA Res* 2004, **11**(1):27-35.

160. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y, Qian J, Hou Y, Wu Z, Li G, Pan M, Li C, Shen Y, Lan X, Yuan L, Li T, Xu H, Yang G, Wan Y, Zhu Y, Yu M, Shen W, Wu D, Xiang Z, Yu J, Wang J, Li R, Shi J, Li H, Li G, Su J, Wang X, Li G, Zhang Z, Wu Q, Li J,

Zhang Q, Wei N, Xu J, Sun H, Dong L, Liu D, Zhao S, Zhao X, Meng Q, Lan F, Huang X, Li Y, Fang L, Li C, Li D, Sun Y, Zhang Z, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Li J, Ye J, Chen H, Zhou Y, Liu B, Wang J, Ye J, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Li S, Wang J, Wong GK, Yang H, Biology Analysis G: **A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*)**. *Science* 2004, **306**(5703):1937-1940.

161. Uchino K, Sezutsu H, Imamura M, Kobayashi I, Tatematsu K, Iizuka T, Yonemura N, Mita K, Tamura T: **Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori***. *Insect Biochem Mol Biol* 2008, **38**(12):1165-1173.
162. Durbin R, Thierry-Mieg J: **The ACEDB Genome Database. In:Suhai S**. Computational Methods in Genome Research. Boston, MA.: Springer; 1994.
163. Dombrowski SM, Maglott D: **20. Using the Map Viewer to Explore Genomes**. 2002.
164. Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR: **Gramene, a tool for grass genomics**. *Plant Physiol* 2002, **130**(4):1606-1613.
165. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM: **Creating a honey bee consensus gene set**. *Genome Biol* 2007, **8**(1):R13.
166. SIB: **ExPASy, The Make2D-DB II Package site**. *https://world-2dpage.expasy.org/make2ddb/*.
167. GMOD: **GBrowse site**. *http://gmod.org/wiki/GBrowse*.
168. utgenome: **University of Tokyo Genome Browser site**. *http://utgenome.org/*.
169. PostgreSQL Global Development Group: **PostgreSQL site**. *https://www.postgresql.org/*.
170. HMMER: **HMMER site**. *http://hmmer.org/*.
171. UC Davis: **ProfileScan site**. *http://rothlab.ucdavis.edu/genhelp/profilescan.html*.
172. HGC: **PSORT site**. *https://psort.hgc.jp/*.
173. Mitaku Group: **SOSUI site**. *http://harrier.nagahama-i-bio.ac.jp/sosui/*.
174. Kyoto University Bioinformatics Center: **MOTIF site**. *https://www.genome.jp/tools/motif/*.

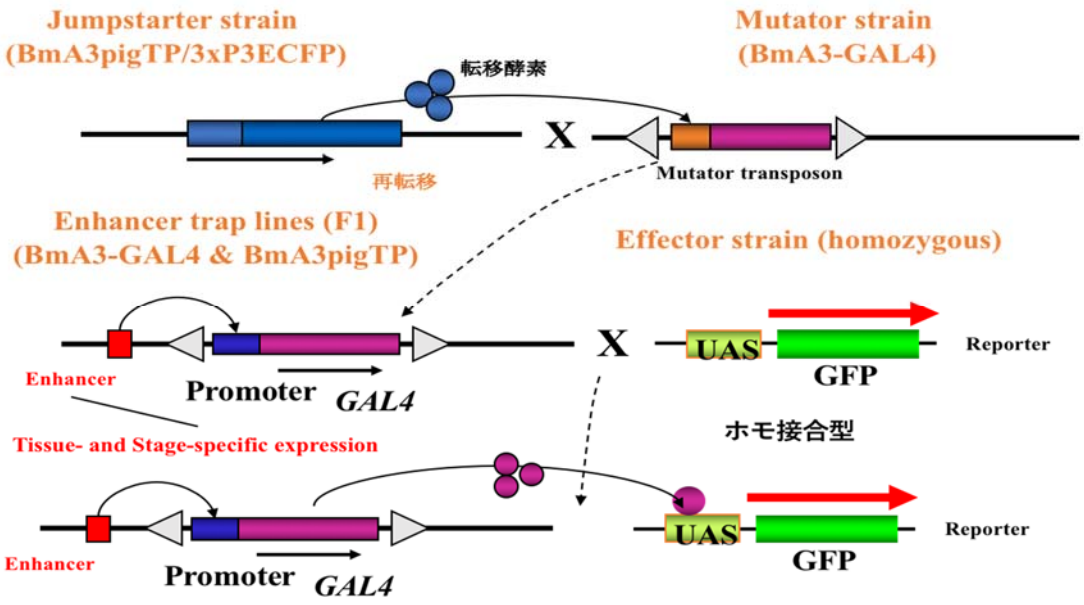
175. EMBL-EBI: **InterProScan site**. *http://www.ebi.ac.uk/interpro/search/sequence-search*.

補足・別冊

補足番号	タイトル
補足 3-1	エンハンサトラップ系統
補足表 2-1	連鎖距離順が一致しないマーカーと物理位置
別冊表 2-2	レファランスマッピングに使用したエンレイの DNA マーカーと物理位置
別冊表 2-3	系統解析で使用したフィルタされたシングルコピー遺伝子
別冊表 2-4	子葉タンパクデータに対応する Gmax275 とエンレイの遺伝子モデル
別冊表 3-1	ライブラリ由来のカイコ cDNA ライブラリと EST のアクセッション番号

補足 3-1 エンハンサトラップ系統

新規遺伝子の網羅的な探索と導入遺伝子の人為的な発現制御のため、トランスポゾンベクタを再可動化することにより、エンハンサトラップ系統が開発された。これらの系統は、GAL4 / UAS を使用して、発育段階および器官/組織特異的な標的導入遺伝子の発現を可視化して調べることに利用することができる。カイコエンハンサトラップ系統を構築するために、最初に、トランスポザーゼをコードする効率的なジャンプスターター系統を開発し、第 2 の piggyBac トランスポゾンベクタ（ミューテータ）を再可動させた。突然変異体トランスポゾンを再可動させるジャンプスターター系統の能力は、*Bombyx* 細胞質アクチン遺伝子(BmA3)プロモータを有する GAL4 構成を保有する突然変異系統と交配することによって試験された。高い再可動活性を有するジャンプスターター系統を確立し、突然変異系統と交配し、F1 子孫を UAS-EGFP（レポータ）系統とハイブリダイズさせることによってエンハンサトラップ系統の作出に使用した。補足図 3-1 に、GAL4-UAS によるカイコのエンハンサトラップを同定するための交配スキームを示す。胚、幼虫、蛹、および成体段階における特徴的な発現パターンを示すいくつかの BmA3-GAL4 エンハンサトラップ系統が、その後の世代において得られた。



補足図 3-1 GAL4-UAS によるカイコのエンハンサトラップを同定するための交配スキーム

補足表 2-1 連鎖距離順が一致しないマーカーと物理位置

Chromosome	Genetic dist. (cM)	Genetic marker	Hit chr/scaffold	Start position (bp)
Chr05	62	T0005064511	Chr05	8,161,203
Chr05	62.1	T000506933s	Chr05	8,645,607
Chr05	63.4	T000515966m	Chr05	9,951,801
Chr05	63.4	T0005150271	Chr05	10,893,519
Chr05	63.7	T0005101741	Chr05	11,852,853
Chr05	63.7	T0005094881	Chr05	12,550,191
Chr05	63.7	T000509183s	Chr05	12,845,189
Chr05	64.8	s000317639-2	Chr05	14,417,488
Chr05	64.8	s000318623-2	Chr05	15,403,493
Chr05	64.5	T000517293s	Chr05	15,988,144
Chr05	63	T0005138981	Chr05	17,235,590
Chr05	62.4	T000513220m	Chr05	17,903,379
Chr05	62.4	T000511990m	Chr05	19,153,689
Chr05	64.2	T000511342s	Chr05	19,796,928
Chr05	64	T000510747m	Chr05	20,402,455
Chr05	64.8	s000314909-2	Chr05	21,959,100
Chr05	64.8	s000311101-2	Chr05	25,741,073
Chr06	106	s004400621	Chr06	32,962,820
Chr08	107.9	Satt708	Chr06	40,461,455
Chr06	107.8	C06-BARC-029025-06052	Chr06	42,014,619
Chr06	108.4	s025900023	Chr06	42,217,040
Chr10	57.8	s032900010	Chr10	7,236,243
Chr10	59.6	s020900037-2	Chr10	7,394,813
Chr10	58.5	s018300465r	Chr10	7,975,136
Chr10	59.2	Sat_221	Chr10	9,932,592
Chr10	59.3	C10-BARC-042707-08378	Chr10	10,258,512
Chr10	59.6	s012201678-2	Chr10	10,293,934
Chr10	59.6	s019800659-2	Chr10	11,902,178
Chr10	106.5	GMES1511	Chr10	43,382,603
Chr10	108.2	s030100090	Chr10	43,531,072
Chr10	107.6	s030100085r	Chr10	43,536,136
Chr10	110.3	Satt331	Chr10	44,029,161

Chr11	71.7	s013701958-2	Chr11	11,055,924
Chr11	71.9	s008000014-2	scaffold_32	12,933
Chr11	74.9	T001111280m	scaffold_32	244,735
Chr11	80.1	s008003841-2	scaffold_21	14,824
Chr11	79.6	s008003458-2	scaffold_21	406,930
Chr11	79.6	Satt519	scaffold_21	886,506
Chr11	79.6	s008002161-2	scaffold_21	1,546,311
Chr11	79.6	s008000965-2	scaffold_21	2,877,395
Chr11	77.4	GMES0027	scaffold_21	3,388,081
Chr11	97.9	GMES2944	Chr11	13,718,116
Chr11	98.4	s019400004	Chr11	15,790,291
Chr11	98.4	s010300014	Chr11	15,812,389
Chr11	98.7	s011500007	Chr11	18,581,497
Chr11	99.5	s004503530-2	Chr11	21,603,186
Chr11	99.2	Satt332	Chr11	23,024,614
Chr11	81.8	T001115342m	Chr11	24,843,454
Chr11	85	T001115835m	Chr11	25,342,287
Chr11	88.6	GMES0675	Chr11	26,277,783
Chr11	90.8	T001116639s	Chr11	26,608,321
Chr11	93.8	Sat_348	Chr11	26,863,332
Chr11	94	Satt597	Chr11	27,034,965
Chr11	99.5	s004504766-2	scaffold_22	190,744
Chr11	99.5	s004505064-2	scaffold_22	490,444
Chr11	99.5	s004506014-2	Chr11	25,635,064
Chr11	100	s004507090-2	Chr11	28,009,064
Chr13	0.8	GMES1672	Chr13	888,124
Chr13	0.5	Satt146	Chr13	1,357,493
Chr13	0.5	T001313439l	Chr13	1,717,198
Chr13	0.5	T001313681l	Chr13	1,980,236
Chr13	0.5	T001314138m	Chr13	2,427,311
Chr13	0.5	T001314272m	Chr13	2,560,889
Chr13	0.5	T001316370s	Chr13	4,895,686
Chr13	0.5	T001317338m	Chr13	5,869,323
Chr13	0	T001320046m	Chr13	6,769,027
Chr13	0	T001319486m	Chr13	7,319,987
Chr13	0	T001319133l	Chr13	7,675,235
Chr13	0.3	T001318664l	Chr13	8,027,540
Chr13	0.5	Satt325	Chr13	8,587,247
Chr13	1.6	Satt343	Chr13	10,391,811

Chr14	63	Sat_355	Chr14	13,587,514
Chr14	70.7	T001435717m	Chr14	18,889,235
Chr14	63.8	T001418240m	Chr14	22,373,532
Chr14	64.9	GMES4127	Chr14	26,042,606
Chr14	64.9	GMES0016	Chr14	28,514,396
Chr14	64.9	GMES5996	Chr14	28,987,541
Chr14	65.2	T0014252711	Chr14	29,394,202
Chr14	65.4	s004007100	Chr14/scaffold_522	30,238,512
Chr14	65.4	Satt601	Chr14	31,286,161
Chr14	70.7	s004002046	Chr14	35,344,964
Chr14	70.7	Satt556	Chr14	38,861,141
Chr14	70.7	s003600119	Chr14	39,068,156
Chr18	59.5	s000403406-2	Chr18	9,335,021
Chr18	60.3	GMES0241	Chr18	9,953,214
Chr18	60	Satt394	Chr18	10,003,381
Chr18	61.6	T001811100m	Chr18	11,124,085