

博士論文

論文題目 植物の完全長 cDNA 収集および DNA 多型
探索によるゲノム研究

氏 名 櫻井 哲也

目次

第1章 緒論.....	1
第2章 植物の完全長 cDNA 解析とその情報基盤整備	10
(1) シロイヌナズナのゲノム注釈情報の改善とデータベース RARGE の構築.....	10
序論.....	10
方法.....	12
結果と考察	15
図表.....	18
(2) ポプラ完全長 cDNA の収集とその配列解析	26
序論.....	26
方法.....	28
結果と考察	34
図表.....	39
(3) キャッサバ完全長 cDNA の収集とその配列解析	56
序論.....	56
方法.....	58
結果と考察	64
図表.....	68

第3章 ゲノム情報を活用した有用作物キャッサバの DNA 多型探索および DNA 多型と 遺伝子機能の関連性の解析	86
序論.....	86
方法.....	88
結果と考察	92
図表.....	97
第4章 結論.....	114
(1) シロイヌナズナのゲノム解析とゲノム統合データベース RARGE の構築.....	114
(2) 有用植物の完全長 cDNA 収集と草本植物シロイヌナズナとの比較解析	118
(3) 有用植物の DNA 多型解析と分子育種基盤整備の推進	120
(4) まとめ	123
参考文献	124
謝辞.....	133

第1章 緒論

ゲノム(genome)という語は、遺伝子 ”gene” + 染色体 “chromosome” あるいは ”gene” + 総体 “-ome”を併せて造られた生物学用語であり、1920年にドイツのハンブルク大学の植物学者 Hans Winkler に定義された。当時は、このゲノムという言葉は、卵と精子に含まれる染色体の1組という意味で使用されていたが、DNA という分子が遺伝情報の本体であることが明らかになった今日においては、生物が持つ全染色体を構成する全塩基配列および全遺伝子セットという定義がされている。真核生物においては、各細胞内に、核と呼ばれる小器官が存在し、生命活動を司る遺伝情報は、この核の中に存在する染色体に保存されている。染色体には、遺伝情報の本体である DNA (deoxyribonucleic acid; デオキシリボ核酸) が存在し、4種類の塩基 A (アデニン)、T (チミン)、G (グアニン)、C (シトシン) から構成されている。DNA は、A と T、G と C が相補的なペアとなっており、この1ペアを1塩基対と呼ぶ。1990年代に塩基配列決定の自動化技術が確立すると、ゲノム科学は急速に発展し、90年代半ばには様々な生物種においてゲノム塩基配列決定プロジェクトが進行した。同じ時期に、コンピュータを用いて大量データから生物学的知識の発見を行うバイオインフォマティクスという研究領域が生まれ、ゲノム科学の発展との関係は密接である。植物のゲノム塩基配列決定プロジェクトについては、2000年に顕花植物として初めてシロイヌナズナのゲノム塩基配列が決定され、2013年12月現在、40種以上の植物のゲノム塩基配列が公表されている。このようなゲノム解読の目的は生物の持つ遺伝情報の有限性を認識することであり、その有限性すなわち全体を理解し、「なぜ環境耐性を示すのか」、「なぜ特定の化合物を高生産するのか」といった様々な生命現象を明らかにするために、A、T、G、C という文字の羅列ともいえるゲノム配列の解析を行う。例えば、RNA 転写領域、翻訳産物(タンパク質)の予測、同定を行い、遺伝子概観を把握する。この解析で得られる情報をゲノム注釈またはゲノムアノテーションと呼び、ゲノム研究のみならず、広く利用されるため、ゲノム注釈は高品質であることが求められる。植物の基礎科学の領域においては、環境応答性の転写因子、プロモータの同定などといった RNA 転写メカニズムの理解にゲノム研究の成果が強く関与している。農業分野では、薬剤や病害耐性に富む植物種、系統の探索、収量などの量的表現形質形質と DNA 多型との関連解析、遺伝子組換え作物の研究や実証実験が活発に行われており、また環境・材料分野

では、アブラヤシやキャッサバ、ヤトロファ（ナンヨウアブラギリ）などの作物のゲノム研究が推進され、生産性の向上や生育に不適と考えられる環境での適応性を示す系統の開発による緑地化など様々な応用研究が取り組まれている。

ゲノム研究から得られる多様な遺伝情報は、現代の生命科学研究にとって必須である。生物種間でゲノムを比較することにより、生物種間における共通性や生物種固有の表現形質との関係性、進化過程の理解が可能になる。したがって、医療、創薬、育種、環境保全、物質生産など様々な応用研究の推進においても、ゲノム情報基盤の構築は有効な手法の1つといえる。

モデル植物シロイヌナズナとゲノム研究

シロイヌナズナは、草丈が約 20cm 程度と小さく、生活環も 2 か月程度である上に、形質転換技術が確立していることなどから、植物研究において最も幅広く使用されている生物種の1つである(Meinke et al., 1998)。シロイヌナズナの染色体数は 5 本で、ゲノムサイズは約 1 億 3000 万塩基対であり、顕花植物としては最小の部類である。この特性から、日本、ヨーロッパ、アメリカの研究機関による国際ゲノム配列決定プロジェクトが発足し、2000 年 12 月に顕花植物として初めて、全ゲノム塩基配列が決定された(Arabidopsis Genome Initiative, 2000)。

その後、シロイヌナズナの包括的研究が盛んに推進され、公共データバンク DNA data bank of Japan)(DDBJ)(Ogasawara et al., 2013)、GenBank (Benson et al., 2013)、現 European Nucleotide Archive(ENA)の European Molecular Biology Laboratory (EMBL) (Cochrane et al., 2013)へ登録される cDNA 等の配列データは増大し、シロイヌナズナのモデル植物としての地位が確立された。ゲノム概観の獲得、遺伝子機能、構造の理解に有効である発現配列タグ(expressed sequence tag; EST)の収集に関しては、フランス、アメリカ、日本による大規模な EST 収集プロジェクトが実施され(Cooke et al., 1996; Meinke et al., 1998; Asamizu et al., 2000)、DDBJ などの公共データバンクでのシロイヌナズナ EST の登録数は 2002 年時点で 20 万配列に達した。しかし、これらプロジェクトで収集された cDNA の多くは完全長ではなく、5'末端側のタンパク質コード領域などの配列情報が欠けているため、転写領域の予測などのゲノム注釈情報に不正確な情報を含むことが懸念されていた。そこで、遺伝子の転写制御領域や詳細な遺伝子構造の理解を推進するため、発現遺伝子の完全長 cDNA の大規模収集および配列決定が行われた(Seki et al.,

2002)。このシロイヌナズナ完全長 cDNA の配列情報を活用することで、シロイヌナズナのゲノムワイドな選択的スプライシング解析や転写因子ファミリータンパク質の識別といった研究が推進された(Iida et al., 2004, 2005)。

プロモータ中には、転写制御に重要な数々の DNA 配列モチーフが含まれており、転写調節メカニズムの解明にはプロモータの同定が重要である。この点において、完全長 cDNA 配列を基に転写開始点を決定する手法は、正確な転写開始点を同定することに適している。また、形質転換体の作出や合成されるタンパク質の機能や構造解析に関する研究においても、完全長 cDNA は、タンパク質を合成するための設計情報をすべて有し、タンパク質そのものを合成することができることから有用な研究基盤といえる。

ゲノム塩基配列の決定、完全長 cDNA の大規模収集およびその配列決定により、遺伝子領域が同定されると研究フェーズは遺伝子機能の同定に移る。*Ds* トランスポゾンや T-DNA の挿入による遺伝子破壊システムの作出は、遺伝子機能同定のための逆遺伝学的なアプローチであり、大規模なシロイヌナズナ変異体システムの開発に基づく研究プロジェクトが推進され、作製された各々のタグ挿入変異体のゲノム上のタグ挿入位置が同定された(Martienssen, 1998; Ito et al., 2002; Alonso et al., 2003; Kuromori et al., 2004)。これらのシロイヌナズナ変異体資源を活用することで、遺伝子機能の同定や遺伝子と表現形質との関係性に関する様々な解析が推進された(Kuromori et al., 2006; Myouga et al., 2010)。

上述のように、シロイヌナズナの cDNA やタグ挿入変異体といった研究資源は、大規模かつ急速に整備されていった。これらの研究資源やその研究から生産された情報は、シロイヌナズナ情報資源データベース(The Arabidopsis Information Resource; TAIR <http://arabidopsis.org>)(Lamesch et al., 2012)などのようなインターネットを通じて閲覧できるオンラインデータベースとして編纂されていった。収録されているシロイヌナズナ遺伝子 ID などのゲノム注釈情報をキーとして、推定される遺伝子機能、DNA 多型などの関連情報が構成されているが、cDNA などの生物遺伝資源への関連付けについては十分とはいえない。転写開始位置や転写制御領域解析などの包括的なシロイヌナズナ研究のより一層の推進のためには、これらの更なる注釈付けが重要である。ゲノム注釈情報の品質に関しても、高品質な完全長 cDNA 配列データなどの活用による改善が肝要である。また、機能ゲノム研究の推進においては、画一的な

品質かつ大量なタグ挿入変異体の情報を用いた解析が有効であり、特にタグ挿入位置の同定のためにはタグ挿入位置近傍ゲノム塩基配列データを獲得できることが必須である。しかし、これら集積された遺伝子破壊系統変異体の情報を検索する手段はゲノム上の物理位置情報によるものでしかないため、タグ挿入位置やゲノム注釈情報を活用した検索機能をもつデータベース構築は、シロイヌナズナの機能ゲノム研究の推進に有意義と考えられる。また、米国科学財団によるデータベース TAIR の運営助成の打ち切りが発表されており、研究目的に応じたデータベースの構築、分散化について議論されている。したがって、この問題への対応する意味でもデータベース構築は重要といえる。

有用植物の完全長 cDNA 収集と草本植物シロイヌナズナとの比較解析

モデル植物シロイヌナズナの研究成果を活用し、環境保全や持続可能な炭素資源としての利用が期待される樹木、デンプンやタンパク質源として食用とされる作物などの他の有用植物の研究の加速化を図ることが重要である。例えば、森林は地表面のおよそ 30%を覆い、生物多様性の維持に貢献している。また、大気や水の浄化、木材、繊維、燃料といった人類にとっての利益を提供し、全世界の産業における供給原料の 25%は森林資源と深く関係している (Food and Agricultural Organization of the United Nations, 2003)。体高が大きく、長い生活環を示す樹木は、一年生草本植物とは異なる選択圧の下で進化し、特有の表現形質を得るに至った。樹木の持つ有用な生物学的特徴を理解することは、植物の分子機構のさらなる解明につながると期待されている。化石資源の枯渇、および環境悪化の深刻化などから、樹木の利用は重要性を増している。いくつかの樹木で、その生物的特徴理解のための研究が進められているが、他の樹木に比べゲノムサイズが小さい(約 4 億 8 千万塩基対)ことから、樹木のモデル植物として、ポプラ(genus *Populus*; ハコヤナギ属)のゲノム研究資源の整備が推進されている。例えば、国際連携研究プロジェクトによって、ポプラ(*Populus trichocarpa*; コットンウッド)の全ゲノム塩基配列が決定された(Tuskan et al., 2006)。樹木であるポプラの生殖期間は長く、交雑に基づく遺伝学的な研究方法には適さないため、機能ゲノム研究に基づく逆遺伝学的アプローチを欠くことはできない。そこで、ポプラの遺伝子概観の獲得や機能ゲノム研究を推進するための重要なツールとして、cDNA の収集およびその配列決定が行われ(Sterky et al., 1998; Hertzberg et al., 2001; Wullschlegler et

al., 2002; Sterky et al., 2004; Ralph et al., 2006)、現在では 30 万以上の EST が利用可能になった。

同様にデンプン作物については、キャッサバのゲノム研究を取り上げる。熱帯性低木キャッサバ(*Manihot esculenta* Crantz)は、熱帯地域における重要作物の一つであり(Cock, 1982)、年間 2 億トン以上ものキャッサバが収穫され、5 億もの人々のカロリー源となっている(Food and Agricultural Organization of the United Nations, 2010)。キャッサバの根から抽出されるデンプンは、食品加工、製紙、繊維、合板などの原料として活用され(Tonukari, 2004)、キャッサバを原料としたデンプン生産は他のデンプン作物と比較しても安価であることから注目を集めている(Amutha and Gunasekaran, 2001)。またキャッサバは、高温、多湿、貧栄養、酸性土壌といった過酷な環境下での栽培が可能であり、度々干ばつが生じるような降水量が非常に少ない地域でもキャッサバ栽培が行われている(El-Sharkawy and Cadavid, 2002)。このような様々な環境に対する高い適応力を示すことに加え、キャッサバはイネやトウモロコシなど他のデンプン作物に比べて栽培が容易であり、デンプンを貯蔵する根塊は土中で最低 2 年間は保持することができるため、飢饉への対応にも優れる一面も持つ(Raheem and Chukwuma, 2001)。上記のような悪条件での生育が可能であるが、最適環境下における生育時に比べ生産性は低下し不安定になることや、細菌またはウイルスによる病気による被害も問題になっており、克服すべき課題といえる。また高デンプン含量を示すキャッサバ有用系統の多くは、食用するにあたって有毒であるシアン化合物を多く含む(Andersen et al., 2000)。以前より行われてきた育種方法により、収量などの幾つかの問題点の改善に成功したが、多くのキャッサバ系統が持つ遺伝的ヘテロ接合性や長い生活環のため、従来法による育種の進展は遅い(Fauquet and Tohme, 2004)。遺伝子地図の作成と量的表現形質遺伝子座の同定に関する研究の報告があるが、より詳細なマッピングや遺伝子同定を行うためには、より多くのゲノムまたは転写産物の配列情報が必要になる。キャッサバの遺伝子概観の獲得や機能ゲノム研究の推進のための重要なツールの 1 つとして cDNA の収集およびその配列決定がある。しかし、その重要性にもかかわらず、公共データバンク GenBank などから利用できるキャッサバの cDNA 配列データは、3 万程度に過ぎない。

cDNA 配列は、ゲノム研究推進に重要な情報資源であるが、上述のように、ポプラの EST は 30 万、キャッサバにいたっては 3 万に過ぎず、他の有用植物トウモロコシの 300 万配列データ、

イネの 200 万配列データなどに比べ、あまりにも小規模といえる (2008 年現在)。また、2012 年に公開されたキャッサバゲノム概要塩基配列は、想定されるゲノムの 60%程度の解読であり (Prochnik et al., 2012)、EST の質、量ともに不十分であるため、RNA 転写領域予測も、不正確な領域定義を多く含むことが考えられる。様々な生物種において EST は着々と増加していったが、そのほとんどは部分長 cDNA 由来の配列データであるため、これらによる遺伝子機能の同定は十分とは言えず、ゲノム塩基配列からの RNA 転写領域予測に際しても、不正確な領域定義を少なからず含む可能性がある。

完全長 cDNA は、転写産物の機能解析だけでなく、転写開始位置の把握やゲノム塩基配列の修正、遺伝子領域の同定に非常に有効である。5'末端のキャップ構造までの全長メッセンジャーRNA を優先的に単離するための様々な方法が開発され (Kristiansen and Pandey, 2002)、これらの完全長 cDNA 単離法は、ヒト、マウス、ショウジョウバエ、シロイヌナズナ、イネ、ヒメツリガネゴケ (Konno et al., 2001; Seki et al., 2002; Stapleton et al., 2002; Suzuki et al., 2002; Kikuchi et al., 2003; Nishiyama et al., 2003) といった様々な完全長 cDNA 大規模収集プロジェクトで活用された。本研究においても、高効率に全長 RNA を収集することができるビオチン化キャップトラッパー法 (Carninci et al., 2000) を用いて、ポプラおよびキャッサバの完全長 cDNA ライブラリを作製した。完全長 cDNA 収集はゲノム注釈情報の品質向上に有効な手段の 1 つであり、このポプラとキャッサバの cDNA 配列データと、既に公開されているポプラとキャッサバ各々の概要ゲノム塩基配列由来の遺伝子予測情報との比較により、新規転写単位 (遺伝子) および遺伝子モデル (転写産物) の同定などのゲノム注釈情報の改善を図った。複合的な遺伝子機能予測や適切な遺伝子機能情報の確認による遺伝子機能注釈は、実際の分子機能の理解に不可欠である。したがって、様々なタンパク質データセットや生物種間共通遺伝子 (オルソログ)、遺伝子オントロジ (Ashburner et al., 2000) を用い、収集したポプラおよびキャッサバの完全長 cDNA および改善した予測転写物に注釈付けを行い、またシロイヌナズナの遺伝子情報と比較することで、各生物種固有の遺伝子を推定し、その遺伝子機能の傾向把握を試みた。

熱帯作物キャッサバの DNA 多型解析 ～ 集積したゲノム情報の活用 ～

ゲノム塩基配列や EST などのゲノム情報が豊富になることで実現することの 1 つとして、生

物種内における比較解析がある。例えば、個体間、系統間における DNA 多型を検出することが可能になり、個体、系統を識別する分子マーカーの整備が推進される。検出した DNA 多型と各個体、系統の表現形質との間の関係性の研究は従来から行われているが、高密度かつゲノムワイドな分子マーカーの整備が可能になることで、より詳細な DNA 多型と表現形質との関連解析が行われるようになる。ゲノム情報を活用することで開発される分子マーカーは、ヒトの病気の原因、投薬効果の傾向の把握や有用作物の品種改良などの研究に直結する。

近年、植物の核酸配列データの蓄積は、遺伝子探索の効果的な研究手法として推進されている (Mochida and Shinozaki, 2010)。実際、幾つかの植物種に対して行われた cDNA 収集とその配列決定は、機能ゲノム研究の推進に貢献した (Kikuchi et al., 2003; Nanjo et al., 2007; Taji et al., 2008; Umezawa et al., 2008; Soderlund et al., 2009)。熱帯性デンプン作物キャッサバにおいても、幾つかの研究グループによって cDNA 収集が行われ (Anderson et al., 2004; Lopez et al., 2004; Lokko et al., 2007; Sakurai et al., 2007)、その成果は網羅的な発現遺伝子を実装したマイクロアレイ設計に活用され、トランスクリプトーム研究を実現させた (Sojikul et al., 2010; An et al., 2012; Utsumi et al., 2012)。現在、キャッサバのゲノム概要塩基配列が公開され、推定ゲノムサイズ 770Mbp の 54%に相当する 419.5Mbp のゲノム塩基配列データが利用可能であり、このキャッサバゲノム概要塩基配列から 30,666 のタンパク質コード遺伝子が予測されている (Prochnik et al., 2012)。

上述したように、分子マーカー開発は育種研究推進に有用であり、マーカー支援選抜 (marker-assisted selection; MAS) を通じ、個体選抜の効率化に活用されている。植物遺伝学研究において、DNA 多型の検出は、集団構造解析、進化関連研究に応用され、モデル植物シロイヌナズナでは、ゲノムワイドな遺伝構造解析が行われている (Cao et al., 2011)。特に近年、一塩基多型 (single nucleotide polymorphism; SNP) マーカーは注目されており、動物やヒトのゲノム解析において、何万もの SNP マーカーを用いたハプロタイプ解析が行われている (International HapMap Consortium et al., 2007)。このような SNP 解析は、作物の品種改良においてもその応用が期待され、植物育種コミュニティでの関心が高まっている。

キャッサバの遺伝子マッピングや分子マーカーの整備が進められ (Akano et al., 2002; Okogbenin and Fregene, 2002; Rabbi et al., 2012)、単純反復配列 (simple sequence repeat; SSR)

の検出による分子マーカー整備やそれを用いた量的形質遺伝子座マッピング(quantitative trait loci mapping; QTL mapping)についての報告がされている(Raji et al., 2009; Sraphet et al., 2011)。さらに遺伝学研究と育種を推進するためには、SNP マーカーのようなより高密度なマーカー整備が求められる。SNP やゲノム挿入/欠損(insertion and deletion; InDel)は、集団内で生じる突然変異であり、キャッサバでもこれらに関して報告されている(Lopez et al., 2005; Ferguson et al., 2012)。また、SNP やゲノム挿入/欠損によって生じた DNA 多型と遺伝子機能との関係性についても報告されており、分子育種の推進だけでなくゲノム構造や遺伝子機能の理解にも重要である(Yamaguchi-Kabata et al., 2008)。

情報資源の活用とデータベース構築

ゲノムや cDNA の大量な配列データが収集され、比較ゲノム解析などの研究に使用されることで DNA 多型などの詳細なゲノム注釈情報が生産されている。これらのゲノム情報は統合され、インターネット上に公開されることにより、効率的な研究推進に活用されている。例えば、The Arabidopsis Information Resource (TAIR) (Lamesch et al., 2012)は、シロイヌナズナ情報のポータルデータベースであり、シロイヌナズナ研究者を始め多くの植物研究者に利用されている。同様に Gramene は、単子葉植物の比較ゲノム研究に関する情報を提供している(Liang et al., 2008)。ムギ類やトウモロコシなどのその他の植物についても情報を統合した有用なデータベースが整備されつつある。したがって、各種ゲノム解析などによって新たに生産された大量の情報をデータベースとして統合し、インターネットに公開することは、更なる植物研究の促進に貢献すると考えられる。

本研究について

以上のようにゲノム科学は、対象生物のゲノム塩基配列解読および転写領域の予測等のゲノム注釈付けによる遺伝情報の有限性の確保を起点に、遺伝子機能の理解からヒトの病気の原因や投薬効果の傾向の把握や有用作物の品種改良など広範囲に応用されている研究領域である。その研究推進に有用である完全長 cDNA の収集とその配列データの生産は、対象生物種のゲノム注釈情報を改善し、生物種間や生物種内でのゲノム比較解析による生物種固有の遺伝子機能の解明や

DNA 多型の検出などを可能にする。また、収集した完全長 cDNA 自体が形質転換体の作出等の生物遺伝資源としての価値を持つ。

本研究では、双子葉草本モデル植物シロイヌナズナの完全長 cDNA 配列を用いたゲノム注釈情報の改善および遺伝子機能解析に有用な *Ds* トランスポゾン挿入変異体の注釈付けを行い、解析に使用した完全長 cDNA と *Ds* トランスポゾン挿入変異体を主としたシロイヌナズナの遺伝子機能情報の統合データベースを構築した。また、樹木ポプラとデンプン作物キャッサバの完全長 cDNA を収集し、cDNA 配列データを獲得した。モデル植物シロイヌナズナでのゲノム注釈改善の手法を活用し、ポプラとキャッサバの cDNA 配列データを用い、各々のゲノム注釈の改善を行った。モデル植物シロイヌナズナのゲノム情報との比較解析を行うことで、各対象生物種とシロイヌナズナとの共通性、対象生物種固有の遺伝子機能の抽出等ゲノム概観の把握を行った。さらに、大規模に収集したキャッサバの cDNA 配列データおよびキャッサバのゲノム概要塩基配列を用い、系統間における DNA 多型の探索および DNA 多型と遺伝子機能との関係性について解析を行った。また、DNA 多型の同定にとどまらず、各 DNA 多型を検出するための PCR プライマーペア配列を設計するなど、分子マーカー開発を支援し、育種研究の推進に貢献した。これらキャッサバのゲノム解析の結果を編纂し、データベースとしてインターネット上に公開することで周辺研究の推進に貢献した。

このように、モデル植物のゲノム研究に始まり、その成果を活用した樹木や作物の完全長 cDNA 収集および配列解析、集積したゲノム情報を用いた DNA 多型の同定といった一連のゲノム研究を行った。さらに、これらの研究成果を編纂し、データベースを構築するといった情報基盤整備を行った。本研究の詳細について次章より述べる。

第2章 植物の完全長 cDNA 解析とその情報基盤整備

(1) シロイヌナズナのゲノム注釈情報の改善とデータベース RARGE の構築

【序論】

シロイヌナズナは、草丈が約 20cm 程度と小さく、生活環も 2 か月程度である上に、形質転換技術が確立していることなどから、植物研究において最も幅広く使用されている生物種の 1 つである(Meinke et al., 1998)。ゲノムサイズは約 1 億 3000 万塩基対であり、顕花植物としては最小の部類であることから、日本、ヨーロッパ、アメリカの研究機関による国際ゲノム配列決定プロジェクトが発足し、2000 年 12 月に顕花植物として初めて、全ゲノム塩基配列が決定された(Arabidopsis Genome Initiative, 2000)。その後、シロイヌナズナの包括的研究が盛んに推進され、公共データバンク DNA data bank of Japan(DDBJ)(Ogasawara et al., 2013)、GenBank(Benson et al., 2013)、現 European Nucleotide Archive(ENA)の European Molecular Biology Laboratory (EMBL) (Cochrane et al., 2013)へ登録される DNA 等の配列データは増大した。ゲノム概観の獲得、遺伝子機能、構造の理解に有効である発現配列タグ(expressed sequence tag; EST)の収集に関しては、フランス、アメリカ、日本による大規模な EST 収集プロジェクトが実施され(Cooke et al., 1996; Meinke et al., 1998; Asamizu et al., 2000)、GenBank での EST 情報の登録数は 2002 年時点でも 20 万配列に達した。しかし、これらプロジェクトで収集された cDNA の多くは完全長ではなく、タンパク質コード領域などの配列情報が欠けていた。そこで、遺伝子の転写制御領域や詳細な遺伝子構造の理解を推進するため、発現遺伝子の完全長 cDNA の大規模収集および配列決定が行われた(Seki et al., 2002)。このシロイヌナズナ完全長 cDNA の配列情報を活用することで、シロイヌナズナのゲノムワイドな選択的スプライシング解析や転写因子ファミリータンパク質の識別といった研究が推進された(Iida et al., 2004, 2005)。

Ds トランスポゾンや T-DNA の挿入による遺伝子破壊系統の開発は、遺伝子機能同定のための効果的なアプローチの 1 つであり、シロイヌナズナ研究において、cDNA 収集と同様に大規模な変異体系統の開発に基づく遺伝子機能解析プロジェクトが推進され、多くのタグ挿入変異体が作出された。(Martienssen, 1998; Ito et al., 2002; Alonso et al., 2003; Kuromori et al., 2004)。こ

これらのシロイヌナズナ変異体資源を活用することで、遺伝子機能の同定、遺伝子と表現形質との関係性に関する様々な解析が推進された(Kuromori et al., 2006; Myouga et al., 2010)。

上述のようにシロイヌナズナの cDNA やタグ挿入変異体といった研究資源は、大規模かつ急速に整備されていった。それら資源やその研究から生産された情報は、シロイヌナズナ情報資源データベース(The Arabidopsis Information Resource; TAIR <http://arabidopsis.org/>) (Lamesch et al., 2012)などのようなインターネットを通じて閲覧できるオンラインデータベースとして編纂されている。2000年12月に解読されたシロイヌナズナの全ゲノム塩基配列データは、ゲノム配列決定が一般的になった今日においても高い品質といわれている。このゲノム塩基配列データに加え、シロイヌナズナの cDNA 配列データは、2012年9月現在でも200万配列に達しようとしており、これら cDNA 配列データを使用したゲノム注釈情報は、植物ゲノムの内で最高水準といえる。しかし、転写領域の同定に使用した cDNA 配列データは、完全長 cDNA 由来または部分長 cDNA 由来であるのかを正しく編纂せずに行われており、不正確なゲノム注釈を含むことが考えられる。また、このシロイヌナズナ情報資源データベース TAIR は、収録されているゲノム注釈情報をキーとして、推定される遺伝子機能、DNA 多型などの関連情報が構成されているが、cDNA などの生物資源への関連付けについては十分とはいえない。RNA 転写開始位置や転写制御領域解析といったトランスクリプトーム研究などの包括的な研究の一層の推進のためには、これら情報の更なる対応付けが必要である。同様に、多くのタグ挿入変異体が作出されたにもかかわらず、各変異体におけるゲノム上のタグ挿入位置など注釈情報の整備は限定的であるため、これらタグ挿入変異体に関する情報整備は機能ゲノム研究の推進に有意義である。したがって、cDNA 情報としては特に重要である完全長 cDNA の関連情報、および画一的な品質かつ大量にタグ挿入位置が同定されている理化学研究所作製の *Ds* トランスポゾンタグ挿入変異体を対象とし、各生物資源についてのゲノム解析、注釈情報生産結果を編纂したシロイヌナズナのゲノム統合情報データベース RARGE (RIKEN Arabidopsis Genome Encyclopedia)を構築し、インターネット上に公開した(<http://rarge.psc.riken.jp/>)。

【方法】

シロイヌナズナ完全長 cDNA データの獲得

理化学研究所の研究チームが作製した完全長 cDNA の配列データを解析対象とした。該当のシロイヌナズナ完全長 cDNA 収集(Seki et al., 2002)、および全ゲノム選択的スプライシング解析 (Iida et al., 2004)で用いられた完全長 cDNA 配列データを公共データベース GenBank (Benson et al., 2012)より獲得した。

シロイヌナズナ完全長 cDNA 配列データの物理的位置の把握とクラスタリング

獲得した完全長 cDNA 配列データの冗長性を排除するため、配列類似性に基づきシロイヌナズナのゲノム塩基配列に対して整列させた。シロイヌナズナのゲノム塩基配列はシロイヌナズナ情報ウェブサイト TAIR (Lamesch et al., 2012)より獲得し、はじめに、核酸配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を用いて、完全長 cDNA 配列が対応するおおよそのゲノム領域の塩基配列データを獲得した。続いて、cDNA 配列整列ソフトウェア SIM4 (Florea et al., 1998)を用いて、完全長 cDNA 配列が対応する正確なゲノム位置、転写方向、遺伝子構造 (エクソン/イントロン構造) を得た。各完全長 cDNA のゲノム上における物理的位置に基づき、クラスタリングを行った。完全長 cDNA クラスタについては、タンパク質配列類似性検索ソフトウェア BLASTP を用い、米国国立生物工学情報センター (National Center for Biotechnology Information; NCBI) (Sayers et al., 2012)提供のタンパク質データセット (Non-Redundant protein sequence dataset; nr) に対する検索を行い、遺伝子機能の注釈付けを行った。BLASTP は、期待値が 10^{-5} 未満 ($-e 1e-5$)、単純配列のマスキング不使用 ($-F F$) の条件で行った。

シロイヌナズナ完全長 cDNA 転写開始位置に基づくプロモータ領域配列の獲得とシス因子の検索

上述の解析で得られた 5'末端読みおよび全長読み完全長 cDNA 配列の物理的位置と転写方向情報に基づきプロモータ領域を得た。本解析では、転写開始位置からその上流 1,000 塩基の領域をプロモータ領域と定義した。定義したプロモータ配列中のシス因子(cis-element)をシス因子検索ツール PLACE (Higo et al., 1998)を用いて検出した。

シロイヌナズナ *Ds* トランスポゾンタグ挿入変異体のタグ近傍ゲノム DNA 配列データの獲得と物理的位置の把握

理化学研究所の研究チームが作製したシロイヌナズナ *Ds* トランスポゾンタグ挿入変異体を解析対象とした(Ito et al., 2002; Kuromori et al., 2004; Kuromori et al., 2006)。核酸配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を用いて、*Ds* トランスポゾンタグ挿入変異体のゲノム近傍配列データを、シロイヌナズナのゲノム塩基配列へ整列させることで、ゲノム上における *Ds* トランスポゾンタグの物理的挿入位置を計算した。BLASTN は、期待値が 10^{-15} 未満(-e 1e-15)、単純配列のマスキング不使用(-F F)の条件で行った。続いて、シロイヌナズナのゲノム注釈情報と *Ds* トランスポゾンタグの物理的挿入位置との対応付けを独自に作成した Perl プログラムを用いて行い、*Ds* トランスポゾンタグ挿入位置と *Ds* トランスポゾンタグ挿入位置近傍に存在する遺伝子との距離(塩基数)を同定した。

シロイヌナズナ完全長 cDNA、*Ds* トランスポゾンタグ挿入変異体の情報統合とデータベース RARGE の構築

本解析で得られたシロイヌナズナ完全長 cDNA および *Ds* トランスポゾンタグ挿入変異体の情報をレコード化し、データベースとして統合した。データベース構築に際しては、レコード間の共通項目の関係モデルに基づく、関係型(リレーショナル)データベースエンジンである PostgreSQL を採用した。

シロイヌナズナ完全長 cDNA データについては、cDNA 名を主キーとし、関連するシロイヌナズナ遺伝子モデル名(AGI gene model ID)との関係を構築した。同様に *Ds* トランスポゾンタグ挿入変異体についても、変異体名を主キーとしたレコード化を行い、変異体が関連するシロイヌナズナ遺伝子モデル名を関連させた。完全長 cDNA と *Ds* トランスポゾンタグ挿入のゲノム上の物理的位置情報については、独自に開発したゲノムビューアを用いて統合した。同様に、シロイヌナズナ遺伝子モデル名によって、完全長 cDNA と *Ds* トランスポゾンタグ挿入変異体の情報を結び付けた。

データベース構築の対象とした完全長 cDNA および *Ds* トランスポゾンタグ挿入変異体に関係

する研究成果の統合や外部データベースとのリンクを実装し、シロイヌナズナ完全長 cDNA を用いた全ゲノム選択的スプライシング解析(Iida et al., 2004)、*Ds* トランスポゾンタグ挿入変異体の表現形質解析(Kuromori et al., 2006)、核コードの葉緑体関連遺伝子解析 (Myouga et al., 2010) を対象データとした。また、シロイヌナズナ転写因子の同定を行ったデータベース RARTF (Iida et al., 2005)とのリンクも実装した。

【結果と考察】

シロイヌナズナ完全長 cDNA データのクラスタリング、新規転写領域の探索

公共データバンク GenBank から理化学研究所が作製したシロイヌナズナ完全長 cDNA の末端読み(EST)および全長読み配列データをそれぞれ、265,107 配列、20,683 配列獲得した(表 2-1-1)。

獲得した完全長 cDNA 配列データのクラスタリング、注釈付けとレコード化についての流れを図 2-1-1 に示す。シロイヌナズナゲノム塩基配列(TAIR10)へマップしたところ、獲得した完全長 cDNA 配列データの 98.7%がマップし、ゲノム上の位置情報を獲得できた(表 2-1-1)。ゲノム配列へマップできた各完全長 cDNA 配列データの位置情報に基づき、既知の遺伝子注釈情報(AGI gene)へ対応させたところ、シロイヌナズナの核ゲノムにコードされている遺伝子の 62.2%である 16,911 のシロイヌナズナ遺伝子に対応した。既知の転写領域と対応しなかった 1,502 配列を対象にゲノム上の位置情報に基づくクラスタリングを行ったところ、503 クラスタを得た。この 503 クラスタは、新規シロイヌナズナ遺伝子の候補とみなせる。また、ゲノム注釈情報と対応した配列についても精査したところ、4,119 のシロイヌナズナ遺伝子が既存のゲノム注釈による転写領域を 500 塩基以上延長できる新規な遺伝子モデルを持つ可能性を示した。これらの配列データに関して、NCBI タンパク質データセット(nr)との配列類似性検索を行い遺伝子機能に関する注釈情報を付加した。最後に、上記のフローで生産したデータについて、データベース構築に際するレコード化を行った。以上のように、大量のシロイヌナズナ完全長 cDNA の配列データのクラスタリングと体系的な注釈付けを行い、レコード化することにより、完全長 cDNA 資源についての円滑な情報参照が可能になった。

シロイヌナズナ完全長 cDNA 転写開始位置に基づくプロモータ領域配列の獲得とシス因子の検索

完全長 cDNA の全長読みおよび 5'末端読み配列のゲノム上の物理的位置情報を用い、プロモータ領域配列を獲得した。本研究では、転写開始位置から上流 1,000 塩基をプロモータ領域と定義し、転写開始点各々の該当領域のゲノム塩基配列を獲得した。続いて、シス因子検索ツール PLACE (Higo et al., 1998)を用いて、獲得したプロモータ配列中のシス因子を検出し、レコード化した。

シロイヌナズナ *Ds* トランスポゾンタグ挿入変異体のタグ近傍ゲノム DNA 配列データの獲得と物理的位置の把握

理化学研究所の研究チームが作製したシロイヌナズナ *Ds* トランスポゾンタグ挿入変異体(Ito et al., 2002; Kuromori et al., 2004; Kuromori et al., 2006)の 17,671 系統について、挿入された *Ds* トランスポゾンタグの近傍ゲノム塩基配列を得た。この配列データをゲノム塩基配列へマップすることで、ゲノム上における *Ds* トランスポゾンタグ挿入位置を計算した。続いて、同定した各々の *Ds* トランスポゾン挿入位置情報に基づき、既知の遺伝子注釈情報(AGI gene)へ対応させ、プロモータ領域、タンパク質コード領域、5'側非翻訳領域、3'側非翻訳領域、非翻訳性 RNA、遺伝子間領域に分類した結果を表 2-1-2 に示す。本解析では、プロモータ領域を転写開始位置から上流 1,000 塩基と定義した。タグ挿入位置が転写領域内であった変異体は 11,570、遺伝子モデルは 9,524 存在し、これらは該当遺伝子モデルの転写を阻害している可能性が高いと思われる。以上のように、大量の *Ds* トランスポゾンタグ挿入変異体について、タグ挿入位置の推算、関連遺伝子の同定などの体系的な注釈付けを行い、レコード化することにより、変異体についての円滑な情報参照を可能にし、該当遺伝子の機能解析を支援する情報資源を整備した。

シロイヌナズナゲノム情報データベース RARGE の構築

上述の解析で生産した情報は、データベース RARGE として統合し、インターネット上に公開した(<http://rarge.psc.riken.jp/>)。各完全長 cDNA と *Ds* トランスポゾンタグ挿入変異体の詳細情報ページを作成するだけでなく、膨大な収録データの円滑な閲覧を可能にするため、各種検索機能を実装した。完全長 cDNA については、cDNA 名、AGI シロイヌナズナ遺伝子名、キーワードによる検索が可能であり(図 2-1-2)、検索結果から該当する cDNA の詳細情報ページに遷移することができる(図 2-1-3)。cDNA 配列データやゲノムビューアによる物理的位置情報の閲覧、TAIR や GenBank といった外部のデータベースへのリンクも備える。同様に、*Ds* トランスポゾンタグ挿入変異体についても、キーワード、変異体名、シロイヌナズナ遺伝子名(AGI 遺伝子 ID)による検索機能を実装し(図 2-1-4)、検索結果から該当する *Ds* トランスポゾンタグ挿入変異体の詳細情報ページに遷移することができる(図 2-1-5)。

本研究と同様に、理化学研究所が提供する完全長 cDNA 配列情報を用いて行われたシロイヌナズナ全ゲノム選択的スプライシング解析(Iida et al., 2004)の結果を共通する完全長 cDNA 名で統合し、遺伝子構造を含むシロイヌナズナのゲノム情報を実装した(図 2-1-6)。Ds トランスポゾンタグ挿入変異体を用いた表現形質解析(Kuromori et al., 2006)、核コードの葉緑体関連遺伝子解析(Myouga et al., 2010)の結果を収録したデータベース RAPID (<http://rarge.psc.riken.jp/phenome/>) と葉緑体タンパク質データベース (<http://rarge.psc.riken.jp/chloroplast/>)、さらにシロイヌナズナ転写因子データベース RARTF (<http://rarge.psc.riken.jp/rartf/>) へのリンクを備え、関連する有用な情報の閲覧性を向上させた。また、実験リソースの請求に関する利便性も考慮し、本データベース RARGE に収録している生物資源の提供を行っている理化学研究所バイオリソースセンターのリソースカタログページ (<http://www.brc.riken.jp/lab/epd/Eng/species/arabidopsis.shtml>) へのリンクも備えた。

以上のように、完全長 cDNA 配列データを用いた転写開始位置、新規転写領域の同定、Ds トランスポゾンタグ挿入変異体のタグ挿入位置の推算とタグ挿入位置と転写領域との関係の獲得を行うことで、ゲノム注釈情報を改善し、シロイヌナズナのゲノム研究の推進に貢献した。また、解析結果および関連情報をデータベース RARGE として統合し、シロイヌナズナゲノムに関する円滑な情報参照環境を提供した。

【図表】

表 2-1-1 データベース構築に用いたシロイヌナズナ完全長 cDNA 配列情報

	末端読み	全長読み	全て
獲得配列数	265,107	20,683	285,790
ゲノムへマップした配列数	261,322	20,645	281,967
遺伝子に対応した配列数	260,064	20,401	280,465
遺伝子に対応しなかった配列数	1,258	244	1,502
cDNA に対応した遺伝子数	16,216	14,821	16,911

核ゲノムにコードされているシロイヌナズナ遺伝子は、TAIR10 ゲノム注釈情報によると 27,206 個である。したがって、本研究の対象とした完全長 cDNA によって、約 60%の遺伝子が単離されていることになる。

表 2-1-2 データベース構築に用いた *Ds* トランスポゾンタグ挿入変異体タグ挿入位置の分類

Ds タグ推定挿入位置	系統数	遺伝子モデル数
プロモータ領域	6,802	6,966
タンパク質コード領域	8,997	7,236
5'側非翻訳領域	1,940	2,142
3'側非翻訳領域	1,076	1,251
非翻訳性 RNA	641	465
遺伝子間領域	3,160	4,172
合計	17,671	15,876

タグ挿入位置をゲノム上の 1 か所に同定できない、または 1 つのタグ挿入位置が複数の項目を満たすため、合計は変異体系統総数を上回る。

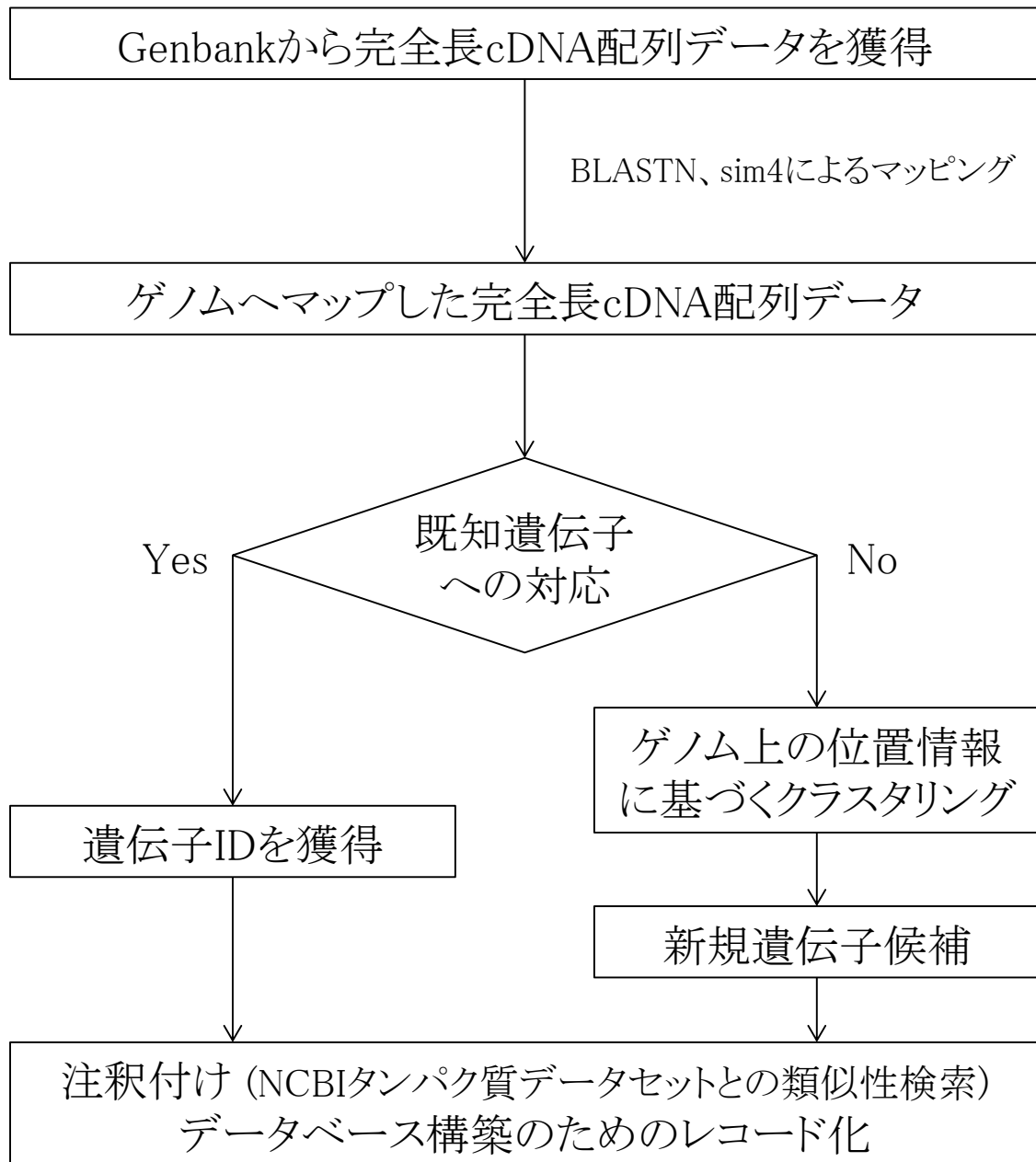


図 2-1-1 シロイヌナズナ完全長 cDNA 配列データの解析処理の流れ

獲得した完全長 cDNA の配列データは、ゲノム塩基配列へのマッピングを経て、クラスタリングされ、16,420 のシロイヌナズナ遺伝子(転写単位)に対応した。既知のシロイヌナズナ遺伝子と対応しなかった 2,195 配列をクラスタリングし、新規遺伝子候補である 987 クラスタを得た。

RAFL cDNAs Keyword Search:

RAFL clone name search (e.g., RAFL02-02-B06)
Enter only one clone name in capitals.

AGI code (MIPS) (e.g., At2g33830)
Search for RAFLcDNA by MIPS protein entry code. Enter only one code.

Gene function (e.g., MAP kinase)
Search for RAFLcDNA putative functions that are expected from blastx(nr).

Distinguish capital letter No Yes

図 2-1-2 完全長 cDNA の検索画面

cDNA 名、AGI シロイヌナズナ遺伝子名、キーワードによる検索が可能。

cDNA detailed information:

Gene Identity

RAFL Code: RAFL02-01-A03
 Sequence type: Sequencing finished(←→)
 AGI Code: At5g17530
 Putative Function: phosphoglucomutase-like protein
 Accession: AY061751
 Chromosome: 5
 Position(5'-3'): 5775610 - 5780157
 Identity: 100%

[View Genome Map](#)
[Link to GPS](#)
[Get Sequence](#)
[5' upstream seq PLACE result](#)

Blastx nr Result

1 NP_568350.1 phosphoglucomutase-related protein [Arabidopsis thaliana]
 Score: 1139
 E-value: 0

2 T51457 phosphoglucomutase-like protein - Arabidopsis thalianaemb|CAC01897.1|
 phosphoglucomutase-like protein [Arabidopsis thaliana]
 Score: 1115
 E-value: 0

3 BAC79968.1 putative phosphoglucomutase precursor, chloroplast [Oryza sativa(japonica
 cultivar-group)]
 Score: 800
 E-value: 0

図 2-1-3 完全長 cDNA の詳細画面

配列タイプ(全長読み、末端読み)、対応 AGI シロイヌナズナ遺伝子名、遺伝子機能注釈、ゲノム上の物理的位置を閲覧できる。

Transposon Mutant Keyword Search

Gene Function Search

Keyword: (e.g., disease TMV)

Score threshold value: (e.g., 1000) *option*

BLASTP Ranking threshold value: [1-10] *option*

Mutant Line Code Search

Line code: (e.g., 52-0031-1)

AGI Code Search


AGI code: (e.g., At5g45260)

図 2-1-4 *Ds* トランスポゾンタグ挿入変異体の検索画面





キーワード、変異体名、シロイヌナズナ AGI 遺伝子名による検索が可能。

Mutant detailed information:


Line code: 52-0031-1 Map


Order: From BRC 

Summary

Query	Position	Closest Gene	Distance	Function		
G-edge	Chr.5 18342770	At5g45260	0	disease resistance protein-like	 MIPS	 BLASTP
H-edge	Chr.5 18342778	At5g45260	0	disease resistance protein-like	 MIPS	 BLASTP

Genomic sequence flanking of Ds fragment:

 G-edge

 H-edge

Detail

BLASTN Result queried G-edge flanking sequence

Position

Chromosome: 5

Nucleotide No.: 18342770

Score: 559


E-value: 1e-158

Identity: 285/286 (99%)

Map

Gene near insertion site

Insertion in Coding Region

AGI code: At5g45260  MIPS

Function: disease resistance protein-like

Insertion region: coding region

図 2-1-5 Ds トランスポゾンタグ挿入変異体の詳細画面

対応 AGI シロイヌナズナ遺伝子名、その距離(塩基数)、遺伝子機能注釈、ゲノム上の物理的位置を閲覧できる。

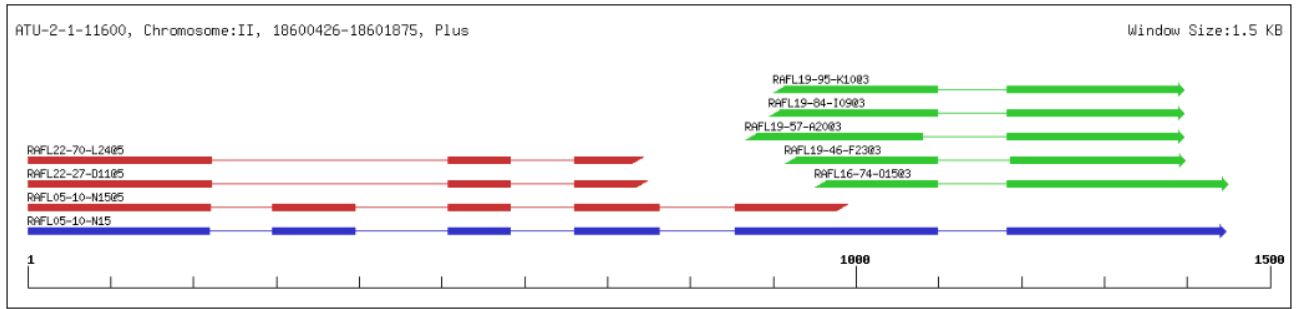


図 2-1-6 シロイヌナズナ選択的スプライシング検出結果の一例

このスクリーンショットは、第 2 エクソンをスキップするタイプの選択的スプライシングの例。それぞれ、全長読み cDNA 配列 (青)、5'-EST (赤)、3'-EST (緑) を示す。

(2) ポプラ完全長 cDNA の収集とその配列解析

【序論】

森林は地表面のおよそ 30%を覆い、生物多様性の維持に貢献している。また、大気と水の浄化、木材、繊維、燃料といった人類にとっての利益を提供し、全世界の産業における供給原料の 25%は森林資源と深く関係している(Food and Agricultural Organization of the United Nations, 2003)。体高が大きく、長い生活環を示す樹木は、一年生草本植物とは異なる選択圧の下で進化している。樹木の持つ有用な生物学的特徴を理解することは、植物の分子機構の解明に大きく貢献すると期待されている。化石資源の枯渇、および環境の維持といった面から、持続可能な炭素資源としての樹木の利用は重要性を増している。いくつかの樹木において、その生物学的特徴の理解のために研究が進められているが、他の樹木に比べゲノムサイズが小さい(およそ 4 億 8 千万塩基対)ことから、モデル樹木として、ポプラ(genus *Populus*; ハコヤナギ属)のゲノム研究資源の整備が推進されている。例えば、国際連携研究プロジェクトによって、ポプラ(*Populus trichocarpa*; コットンウッド)の全ゲノム塩基配列が決定された(Tuskan et al., 2006)。樹木であるポプラの生殖期間は長く、交雑に基づく遺伝学的な研究方法には適さないため、機能ゲノム研究に基づく逆遺伝学的アプローチを欠くことはできない。そこで、ポプラの遺伝子概観の獲得や機能ゲノム研究を推進するための重要なツールとして、cDNA の収集およびその配列決定が行われ(Sterky et al., 1998; Hertzberg et al., 2001; Wullschleger et al., 2002; Sterky et al., 2004; Ralph et al., 2006)、現在では 30 万以上の EST が利用可能になった。

完全長 cDNA は、転写産物の機能解析だけでなく、転写開始位置の把握やゲノム塩基配列の修正、遺伝子領域の同定に有効である。5'末端のキャップ構造までの全長メッセンジャーRNA を優先的に単離するための様々な方法が開発され(Kristiansen and Pandey, 2002)、これらの完全長 cDNA 単離法は、ヒト、マウス、ショウジョウバエ、シロイヌナズナ、イネ、ヒメツリガネゴケ(Konno et al., 2001; Seki et al., 2002; Stapleton et al., 2002; Suzuki et al., 2002; Kikuchi et al., 2003; Nishiyama et al., 2003)といった様々な完全長 cDNA 大規模収集プロジェクトで活用された。先に述べたように、ポプラの EST は着々と増加していったが、そのほとんどは部分長 cDNA 由来の配列データであるため、これらによる遺伝子の同定等のゲノム注釈情報は十分とは言えず、

ゲノム塩基配列からの RNA 転写領域予測に際しても、不正確な定義を含む可能性がある。

この状況を改善するため、今回、高効率に全長 RNA を収集することができるビオチン化キャップトラッパー法(Carninci et al., 2000)を用いて、ポプラ完全長 cDNA ライブラリを作製した。収集した cDNA 配列データとポプラのゲノム塩基配列由来の予測遺伝子との比較を行うことにより、新規転写単位(遺伝子座)および遺伝子モデル(転写産物)を同定した。遺伝子機能予測や遺伝子機能情報の確認によるゲノム注釈の改善だけでなく、様々なタンパク質データセットや生物種間共通遺伝子(オルソログ)、遺伝子オントロジ(Ashburner et al., 2000)を用い、収集したポプラ完全長 cDNA に注釈付けを行った。また、シロイヌナズナの遺伝子情報と比較することで、ポプラ固有の遺伝子を分類し、その遺伝子機能の傾向を把握した。

【方法】

植物材料とストレス処理

葉芽、花芽、小枝を成熟したポプラ(*Populus nigra* ver. *italica*)の雌株から採取した。採取したサンプルは、直径 90mm、高さ 130mm のバイオポット(渡辺泰株式会社、大阪)を用い、無菌環境下で外植した。外植体は、McCown 木本基礎塩類混合粉末(Sigma-Aldrich Corp., St. Louis, MO)、2%(w/v)ショ糖、2%(w/v)活性炭、0.3%(w/v)ゲランガムを含む培地で育成し、クリーンルーム温度は $25 \pm 1^\circ\text{C}$ 、光量は白色蛍光灯の使用下で $40 \sim 60 \mu\text{mol m}^{-2}\text{s}^{-1}$ 、明期 16 時間、暗期 8 時間とした。

バイオポットで 2 か月間の育成後、植物体をサンプリングし、根を蒸留水で洗浄し、一連のストレス処理を行った。乾燥、低温、高温、および塩、アブシジン酸、サリチル酸、ジャスモン酸、過酸化水素曝露の 8 種類のストレス処理を行い、各ストレス処理後、1、2、5、10、24 時間にサンプリングを行った。乾燥処理については、直径 90mm、高さ 20mm のペトリ皿に葉を乗せ、弱光、室温 25°C 、湿度 50%の条件下で行った。低温と高温処理については、湿らせたペーパータオルを敷いたペトリ皿上で、暗所、 4°C (低温処理)、および 34°C (高温処理)の条件下で行った。それ以外の処理については、葉を 50ml の溶液に浸し、各々の濃度は 400mM 塩化ナトリウム、 $100 \mu\text{M}$ アブシジン酸、 $100 \mu\text{M}$ サリチル酸、 $100 \mu\text{M}$ ジャスモン酸、200mM 過酸化水素を用い、弱光、室温 25°C の条件下で行った。すべてのサンプルは、RNA 抽出のため、液体窒素で凍結させた。

RNA 単離と完全長 cDNA ライブラリの構築

トータル RNA はフェノール-グアニジン・イソチオシアン酸溶液(TRIzol®、Invitrogen 社、カリフォルニア)を用いて抽出した。上記のすべてのサンプルに由来するトータル RNA の混合物は、 $\mu\text{MACS mRNA}$ 単離キット(Miltenyi Biotec 社、ドイツ)を用いて、さらに精製した。ポリアデニン鎖(poly A tail)が付いた RNA を選択し、cDNA ライブラリ作製に使用した。完全長 cDNA ライブラリは、トレハロース熱活性逆転写酵素を用いたビオチン化キャプトラッパー法(Carninci et al., 2000)によって作製した。

完全長 cDNA の末端読み配列決定

各 cDNA クローンの挿入 DNA に関しては、TempliPhi DNA 増幅キット(GE Healthcare 社、英国)を使用した RCA 法(Dean et al., 2001)により増幅し、384 ウェルプレートで調製した。末端読み配列決定は、AB3700 DNA 解析装置(Applied Biosystems 社、カリフォルニア)により行い、5'末端読み配列決定には、M13-21 プライマー(5'-TGTAACAACGACGGCCAGT-3')、3'末端読み配列決定には、1233 プライマー(5'-AGCGGATAACAATTTTCACACAGGA-3')を用いた。

配列データの調整と配列アセンブリ作成

シーケンサから出力された波形データは、ソフトウェア phred (Ewing et al., 1998)を用いてベースコールした。配列両端の低品質部については、スコア 20 以上の塩基が 20 塩基以上続くようになるまで除去した。併せて、配列マスクソフトウェア cross_match (Ewing et al., 1998)を使用してベクター配列部を除去した。低品質、不要部の除去処理後に残った塩基配列長が 100 未満の配列は除外した。さらに、大腸菌ゲノム塩基配列に対し、塩基配列類似性検索ソフトウェア BLASTN を実行し、 10^{-100} 未満の期待値(E value)を示す配列データを除外した。上記の調整を行った EST を日本 DNA データバンク(DNA Data Bank of Japan; DDBJ) (Kodama et al., 2012)へ登録し、以降のゲノム解析における基本データとした。

調整済みの cDNA 配列データを配列連結(アセンブル)ソフトウェア CAP3 (Huang and Madan, 1999)を用いてアセンブルし、配列アセンブリを得た。アセンブルにより、配列の一致、整列によって連結されたコンティグ配列(contiguous sequence)を得ることができる。CAP3 は、初期設定パラメータにて実行した。

スキヤフォールドの構築

非冗長な cDNA 配列セットを得るため、上記 EST のアセンブリ配列をクローン名に基づきクラスタリングした。CAP3 が出力するファイルの 1 つである ace ファイルは、投入された配列の連結結果を記す。この ace ファイルを分析することで、コンティグ配列を構成する cDNA クローンの組み合わせを解釈し、転写産物毎のスキヤフォールド(scaffold)を構築した。スキヤフォールドを構成する両末端の配列(コンティグ配列または cDNA 端読み配列)は 20 文字の 'N' で連結した。以

降の解析で配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を使用するが、BLASTN が検索時に使用する検索範囲は 11 塩基であるため、スキヤフォールド内の 20 文字の 'N' が配列検索を妨げることはない。

完全長 cDNA ライブラリの品質

本研究では、両末端読み配列データを持つ cDNA を対象に完全長 cDNA 率を算出した。3'末端読み配列にポリアデニン鎖が確認でき、既知タンパク質の配列データセット NCBI-nr(Sayers et al., 2012)との配列類似性検索において、 10^{-30} 未満の期待値(E value)、正方向の翻訳枠、かつメチオニン(M)から整列する cDNA クローンを完全長 cDNA と見なした。

代謝パスウェイマッピング

収集した完全長 cDNA の代謝パスウェイへのマッピングには、分子間ネットワークデータベース KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2012)を用いた。ポプラ cDNA 配列データを問い合わせ配列とし、KEGG 自動注釈付けサービス KAAS(Moriya et al., 2007)へ投入した。対象データセットとして Ath(シロイヌナズナ)を、対応付け方法に SBH (single-directional best hit)法を選択して実行した。

完全長 cDNA 配列のゲノムマッピングと新規遺伝子領域の探索

ポプラのゲノム塩基配列は植物ゲノム情報ウェブサイト Phytozome (Goodstein et al., 2012)より獲得し、はじめに、核酸配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を用いて、完全長 cDNA 配列が対応するおおよそのゲノム領域の塩基配列データを獲得した。続いて、cDNA 配列整列ソフトウェア SIM4 (Florea et al., 1998)を用いて、完全長 cDNA 配列が対応する正確なゲノム位置、転写方向、遺伝子構造 (エクソン/イントロン構造) を得た。

得られた完全長 cDNA のゲノム上における物理的位置情報を Phytozome より獲得した既知のポプラゲノム注釈情報(転写単位、遺伝子モデル)と対比させ、新規転写単位(transcription unit、遺伝子座)、遺伝子モデル(gene model、mRNA)を同定した。末端読み配列のどちらか一方が、既知のゲノム注釈情報の転写単位領域より外側に伸長するように対応した場合を新規遺伝子モデル

とし、末両端読み配列が既知のゲノム注釈情報と対応しなかった場合を新規転写単位とした。

全代表転写物配列データの準備

真核細胞では選択的スプライシングより、同一転写領域から複数種の転写産物（スプライスバリエーション）が生産されることがある。遺伝子機能などの比較解析に際しては、スプライスバリエーションにより統計の偏りを防ぐため、転写領域につき1つの代表転写産物を選抜したデータセットを用いた。ポプラのゲノム注釈情報由来の各転写領域から選抜された代表転写産物配列データに上記解析で検出した新規転写領域のスキュフォールドを加えた配列データセットをポプラの全代表転写物配列と定義した。以降、この配列のことをポプラ代表転写物配列と表現する。

比較ゲノム解析

ポプラ代表転写物配列データをシロイヌナズナタンパク質との類似性の有無による2つのグループに分類し、比較解析を行った。分類方法を次に述べる。ポプラ代表転写物配列データセットを問い合わせ配列とし、シロイヌナズナ情報ウェブサイト TAIR(Lamesch et al., 2012)よりダウンロードしたシロイヌナズナタンパク質配列データセット (TAIR10_pep_20101214) に対して、BLASTX 検索を行った。BLASTX は、期待値が 10^{-10} 未満(実行時オプション `-e 1e-10`)、単純配列のマスキング不使用(実行時オプション `-F F`)の条件で行った。本解析では、BLASTX 検索結果にしたがい、期待値が 10^{-10} 未満であったポプラ代表転写物配列をシロイヌナズナと類似すると判定した。逆に期待値が 10^{-10} 以上、または類似対象シロイヌナズナタンパク質が無かった(結果が No hits found であった)ポプラ代表転写物配列をシロイヌナズナと類似せずとした。

遺伝子機能解析

ポプラの遺伝子機能概観を獲得するため、生物種間共通遺伝子(オルソログ)クラスター COG (Tatusov et al., 2000)と遺伝子オントロジ (Ashburner et al., 2000)を使用した。COG のタンパク質配列データセットは、NCBI の FTP サーバ(<ftp://ftp.ncbi.nih.gov/pub/COG/>)より獲得し、真核生物のオルソログタンパク質配列データセット KOG の 2003 年 3 月公開版を使用した。ポプラ代表転写物配列データを問い合わせ配列とし、KOG に対する BLASTX 検索を行った。BLASTX

は、期待値が 10^{-5} 未満(-e 1e-5)、単純配列のマスキング不使用(-F F)の条件で実行した。その配列類似性検索の結果から、最高スコアのタンパク質 ID を獲得し、そのタンパク質 ID と KOG ID および KOG 遺伝子機能分類 ID とを対応付けた。

同様に遺伝子オントロジへの対応付けについても、はじめに、ポプラ代表転写物配列を問い合わせ配列とし、欧州バイオインフォマティクス研究所(European Bioinformatics Institute; EBI) 提供のタンパク質配列データセット UniProt/TrEMBL (Dimmer et al., 2012)に対する BLASTX 検索を行った。BLASTX は、期待値が 10^{-5} 未満(-e 1e-5)、単純配列のマスキング不使用(-F F)の条件で実行した。その配列類似性検索の結果から、最高スコアのタンパク質 ID を獲得し、遺伝子オントロジウェブサイト(<http://www.geneontology.org/>)からダウンロードしたタンパク質 ID - 遺伝子オントロジ ID 対応付けデータファイル gp_association.goa_uniprot を用い、タンパク質 ID と遺伝子オントロジ ID とを対応付けた。さらに、同じく遺伝子オントロジウェブサイト(<http://www.geneontology.org/>)からダウンロードした遺伝子オントロジ ID グループ化データファイル goslim_plant.obo を用い、対応付けた遺伝子オントロジ ID を集約した。

統計解析

ポプラとシロイヌナズナの遺伝子機能分類結果、シロイヌナズナと類似-非類似ポプラ代表転写産物グループの遺伝子オントロジ解析結果についての統計解析に際し、その 2 群の独立性を確認するため、ピアソンのカイ二乗検定(Pearson chi-square test)と併せて、調整済み標準化残差分析(adjusted standardized residual analysis) (Bewick et al., 2004)を行った。この統計解析手法は、2 群間の独立系の検定後、群を構成する各項目についての残差(観測値-期待値)に基づき、その項目に関して 2 群間における差異を検定するものである。式を次に示す。

$$e_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

$$v_{ij} = \left(1 - \frac{n_{i.}}{N}\right) \left(1 - \frac{n_{.j}}{N}\right)$$

$$d_{ij} = \frac{e_{ij}}{\sqrt{v_{ij}}}$$

O : 観測値

E : 期待値

e : 標準化残差

v : 残差分散

n_i : 行周辺度数(行周辺和)

n_j : 列周辺度数(列周辺和)

d : 調整済み標準化残差

標準化残差は、近似的に平均 0、分散 1 の標準正規分布に則る。したがって、この標準化残差は標準正規分布における Z スコアと見なせる。

【結果と考察】

cDNA クローンの両末端読み配列の決定と配列アセンブリ

ポプラ(*Populus nigra*)の花芽、葉芽、根および各種ストレス処理を施した葉を出発材料とし、ビオチン化キャプトラッパー法を用いて完全長 cDNA ライブラリを作製した。作製した完全長 cDNA ライブラリから cDNA を単離し、両末端からの端読み配列を決定した。各配列データから低品質部、ベクター部を削除、大腸菌配列を除外し、89,572 配列(5'末端読み:46,000、3'末端読み配列:43,572 配列)を得た(表 2-2-1)。この配列データセットを以降のゲノム解析の基本データとし、公共データベース DDBJ へ登録した(登録番号 BP921855~BP929692、BP929693~BP937111、DB874873~DB910976)。cDNA クローンは、理化学研究所バイオリソースセンターに寄託した。

既知のタンパク質配列に対して高い類似性を示す 5'および 3'末端読み配列を持つ cDNA クローンについて、完全長 cDNA 比(完全長 cDNA 数/対象 cDNA 数)を計算したところ、0.86 であった。これらの cDNA クローンが完全なタンパク質コード領域を含んでいることを意味し、遺伝子機能や遺伝子構造の解析への使用について十分な完全長 cDNA ライブラリ品質であることを示した。

低品質部などを除去し、得られたポプラ cDNA 配列データを、配列連結ソフトウェア CAP3 を用いてアセンブルし、14,912 のコンティグと 10,451 のシングレットを得た。さらに、独立した転写産物(mRNA)を単位とする配列セットを作成するため、この CAP3 によるアセンブル結果と cDNA 配列情報から 17,838 スキャフォールドを構築した(表 2-2-1)。このスキャフォールド構築結果から、スキャフォールドを構成する cDNA クローン数を確認したところ、スキャフォールドの 54% が 1 つの cDNA クローンで、80%以上のスキャフォールドは 3 つ以下の cDNA クローンで構成されており、cDNA 重複が少なく、効率的に多様な cDNA の収集が行われたことが確認された(図 2-2-1)。

収集した cDNA の網羅性、および新規転写領域の探索

分子間ネットワークデータベース KEGG (Kyoto Encyclopedia of Genes and Genomes) を用いて、ポプラ cDNA 配列データを代謝パスウェイにマップしたところ、123 代謝経路図上の 2,090 個の酵素に対応した(表 2-2-2)。シロイヌナズナタンパク質は、2,564 個の酵素に対応しており、本研究において、シロイヌナズナの酵素遺伝子の 82%に相当するポプラ完全長 cDNA を収集した

ことを示した(表 2-2-2)。

ポプラゲノム塩基配列上における転写領域を探索するため、本研究で収集したポプラ完全長 cDNA の配列データをポプラゲノム塩基配列へマップしたところ、96.6%に相当する 84,758 配列がマップされ、ゲノム上の物理的位置情報が得られた(表 2-2-3)。獲得できた各完全長 cDNA 配列データの位置情報に基づき、ポプラのゲノム注釈情報(Phytozome v8.0 annotation)へ対応させたところ、ポプラゲノムにコードされている遺伝子の 34.9%である 14,208 のポプラ遺伝子に対応した。ポプラ遺伝子と対応しなかった 3,009 配列を対象にゲノム上の位置情報に基づくクラスタリングを行ったところ、452 個の新規転写領域を検出した。ポプラゲノム塩基配列にマップしなかった cDNA 配列を精査したところ、492 個のスキヤフォールドが対応し、これらも新規遺伝子領域の候補と考えられる。以上の結果から本研究によって、15,000 個以上の転写単位に対応するポプラ cDNA を収集できたと考えられた。また、ポプラのゲノム注釈情報と対応した配列について精査したところ、1,087 個の新規遺伝子モデルを検出した。以上より、本研究で収集した完全長 cDNA によって、新規転写単位および遺伝子モデルが検出されたことによって、既存のゲノム注釈情報を改善することができた。既存のゲノム注釈情報由来の 40,668 配列に、この解析で検出できた 944 の新規遺伝子の配列データを加えた配列データセット(41,612 配列)をポプラゲノム中の全転写領域の代表配列と定義した。以降、ポプラ代表転写物配列と表現し、配列解析で使用した。

遺伝子機能解析

ポプラの遺伝子機能概観を把握するため、上述の解析で得たポプラ代表転写物配列データを問い合わせ配列として、生物種間共通遺伝子(オルソログ)クラスタ COG の真核生物オルソログタンパク質配列データセット KOG に対する BLASTX 配列類似性検索を行い、その検索結果にしたがって、遺伝子機能を分類した。KOG は、3 動物種(センチュウ、ショウジョウバエ、ヒト)、1 植物種(シロイヌナズナ)、2 真菌種(パン酵母、分裂酵母)、1 微胞子虫種(*Encephalitozoon cuniculi*) のタンパク質配列データで構成されている。本研究で定義したポプラ代表転写物配列 41,612 の内、22,852(54.9%)が KOG に対応し、ポプラ遺伝子の機能分類を得た(図 2-2-2)。草本双子葉モデル植物シロイヌナズナとの比較解析を行うため、シロイヌナズナについても同様の遺伝子機能

分類を行った結果、シロイヌナズナ代表転写物配列 27,206 の内、15,347 (56.4%) が KOG に対応した。ポプラとシロイヌナズナの遺伝子機能分類について、各植物での分類の割合の比較を図 2-2-3 に示す。ポプラとシロイヌナズナの遺伝子機能分類結果の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ、p 値は 6.78×10^{-29} を示し、ポプラとシロイヌナズナの間で遺伝子機能の分類割合に有意な差があることが示唆された。続いて差がある分類項目を検定するため、調整済み標準化残差分析を行ったところ、“Transcription”、“Replication, recombination and repair”、“Cell cycle control, cell division, chromosome partitioning”、“Signal transduction mechanisms”、“Cytoskeleton”、“Intracellular trafficking, secretion, and vesicular transport”、“Posttranslational modification, protein turnover, chaperones”、“Secondary metabolites biosynthesis, transport and catabolism”の 8 個の分類項目について有意差が認められた($p < 0.05$)。

ポプラが持つ遺伝子機能をより詳細に把握するため、ポプラ代表転写物配列データセットを問い合わせ配列とし、シロイヌナズナタンパク質配列データセット (TAIR10_pep_20101214) に対する BLASTX 検索を行った。その結果、期待値が 10^{-10} 未満であった 34,149 ポプラ代表転写物配列をシロイヌナズナと類似すると判定した。この配列群をシロイヌナズナ類似配列と表す。期待値が 10^{-10} 以上、または類似対象シロイヌナズナタンパク質が無かった(結果が No hits found であった) 7,463 ポプラ代表転写物配列をシロイヌナズナと類似せずと分類した。この配列群をシロイヌナズナ非類似配列とした。

図 2-2-4 は、遺伝子オントロジの生物学的プロセス(biological process)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の対応遺伝子オントロジの割合の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ p 値は 1.58×10^{-89} を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、“DNA metabolic process”、“biosynthetic process”、“nucleobase-containing compound metabolic process”、“response to stress”、“protein metabolic process”、“cell death”、“photosynthesis”、“generation of precursor metabolites and energy”、“response to biotic stimulus”の 9 個の遺伝子オントロジ項目が、シロイヌナズナ非類似配列で優位に高い割合を示すことが確認できた。

($p < 0.01$)。遺伝子オントロジ"biosynthetic process"については、配下にリグニン、リグナン、ノルリグナン等のフェニルプロパノイド生成に関連する遺伝子オントロジが存在する。先の KOG による機能解析においても、"Secondary metabolites biosynthesis, transport and catabolism"の割合がポプラ転写物配列で優位に高かった(図 2-2-3)。このことから、ポプラ特有または多様性に富む生合成関連遺伝子の存在が示唆された。遺伝子オントロジ"response to stress"や KOG 機能分類の"Signal transduction mechanisms"についても、ポプラでの割合が優位に大きく、シロイヌナズナにはないストレス応答性遺伝子またはメカニズムの多様化の可能性が示唆された。遺伝子オントロジ"cellular process"、"metabolic process"、"cellular protein modification process"などの汎用的な遺伝子機能については、シロイヌナズナ類似配列での分類割合が高かった。この結果から、これらの遺伝子機能については、シロイヌナズナとポプラとで共通していると考えられる。

図 2-2-5 は、遺伝子オントロジの細胞構成要素 (cellular component)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の遺伝子オントロジ対応結果について、ピアソンのカイ二乗検定を行ったところ p 値は 9.57×10^{-42} を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、"plastid"、"nucleus"、"thylakoid"、"membrane"の 4 個の遺伝子オントロジ項目に有意差を確認した ($p < 0.01$)。非類似配列群で色素体やチラコイドについての割合が高いことから、ポプラ特有の関連遺伝子機能またはメカニズムの多様化の可能性が示唆された。

図 2-2-6 は、遺伝子オントロジの分子機能(molecular function)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の遺伝子オントロジ対応結果について、ピアソンのカイ二乗検定を行ったところ p 値は 0.0 を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、17 個の遺伝子オントロジ項目に有意差を確認した($p < 0.01$)。特に"nucleotide binding"、"catalytic activity"、"kinase activity"、"transporter activity"の 4 個の遺伝子オントロジ項目が、シロイヌナズナ類似配列で高い割合を示した。この結果から、これらの遺伝子機能については、シロイヌナズナとポプラとで共通していると考えら

れる。また、核酸結合タンパク質に関する遺伝子オントロジ項目が、シロイヌナズナ非類似配列で高い割合を示し、転写調節、シグナル伝達などに関連するポプラ特有のメカニズムの多様化の可能性を示した。

【図表】

表 2-2-1 完全長 cDNA の収集と配列アセンブリの内訳

クローン数	47,137
端読み配列数 (5'/3'末端読み)	89,572 (46,000/43,572)
コンティグ数	14,912
シングレット数	10,451
スキヤフォールド数	17,838

表 2-2-2 各代謝パスウェイに対応するポプラ cDNA とシロイヌナズナの酵素遺伝子数の比較

	KEGG Pathway	Poplar	Arabidopsis	Pathway Coverage
03010	Ribosome	105	109	0.96
03040	Spliceosome	77	101	0.76
04141	Protein processing in endoplasmic reticulum	70	74	0.95
03013	RNA transport	66	91	0.73
00190	Oxidative phosphorylation	57	67	0.85
00230	Purine metabolism	56	81	0.69
04120	Ubiquitin mediated proteolysis	46	56	0.82
00240	Pyrimidine metabolism	45	70	0.64
04075	Plant hormone signal transduction	40	43	0.93
03015	mRNA surveillance pathway	39	47	0.83
03008	Ribosome biogenesis in eukaryotes	38	56	0.68
00520	Amino sugar and nucleotide sugar metabolism	36	37	0.97
03018	RNA degradation	36	48	0.75
04146	Peroxisome	33	33	1.00
03050	Proteasome	33	35	0.94
04144	Endocytosis	32	35	0.91
00330	Arginine and proline metabolism	31	34	0.91
00500	Starch and sucrose metabolism	30	32	0.94
00270	Cysteine and methionine metabolism	28	29	0.97
00010	Glycolysis / Gluconeogenesis	27	29	0.93

00260	Glycine, serine and threonine metabolism	27	30	0.90
00860	Porphyrin and chlorophyll metabolism	27	30	0.90
00564	Glycerophospholipid metabolism	27	32	0.84
00195	Photosynthesis	26	26	1.00
00620	Pyruvate metabolism	26	29	0.90
04626	Plant-pathogen interaction	26	37	0.70
04145	Phagosome	25	26	0.96
00710	Carbon fixation in photosynthetic organisms	24	24	1.00
00970	Aminoacyl-tRNA biosynthesis	24	25	0.96
00250	Alanine, aspartate and glutamate metabolism	24	26	0.92
03060	Protein export	23	26	0.88
00630	Glyoxylate and dicarboxylate metabolism	22	23	0.96
04712	Circadian rhythm - plant	22	27	0.81
00900	Terpenoid backbone biosynthesis	22	29	0.76
00510	N-Glycan biosynthesis	22	32	0.69
00561	Glycerolipid metabolism	21	22	0.95
03022	Basal transcription factors	20	31	0.65
00020	Citrate cycle (TCA cycle)	19	19	1.00
03420	Nucleotide excision repair	19	37	0.51
00940	Phenylpropanoid biosynthesis	18	21	0.86
00051	Fructose and mannose metabolism	17	17	1.00
00350	Tyrosine metabolism	17	17	1.00
00562	Inositol phosphate metabolism	16	18	0.89
00280	Valine, leucine and isoleucine degradation	16	19	0.84

00400	Phenylalanine, tyrosine and tryptophan biosynthesis	16	23	0.70
00030	Pentose phosphate pathway	15	15	1.00
00052	Galactose metabolism	15	15	1.00
00640	Propanoate metabolism	15	16	0.94
00100	Steroid biosynthesis	15	16	0.94
00480	Glutathione metabolism	15	16	0.94
04130	SNARE interactions in vesicular transport	15	17	0.88
00410	beta-Alanine metabolism	14	14	1.00
00360	Phenylalanine metabolism	14	15	0.93
00380	Tryptophan metabolism	14	16	0.88
04070	Phosphatidylinositol signaling system	14	16	0.88
00906	Carotenoid biosynthesis	14	17	0.82
03020	RNA polymerase	14	27	0.52
03030	DNA replication	14	32	0.44
00053	Ascorbate and aldarate metabolism	13	14	0.93
00770	Pantothenate and CoA biosynthesis	13	16	0.81
00910	Nitrogen metabolism	13	17	0.76
00563	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	13	21	0.62
00196	Photosynthesis - antenna proteins	12	12	1.00
00941	Flavonoid biosynthesis	12	12	1.00
00040	Pentose and glucuronate interconversions	12	13	0.92
00061	Fatty acid biosynthesis	12	13	0.92
00592	alpha-Linolenic acid metabolism	12	14	0.86
00130	Ubiquinone and other terpenoid-quinone biosynthesis	12	18	0.67

00920	Sulfur metabolism	11	11	1.00
00600	Sphingolipid metabolism	11	12	0.92
01040	Biosynthesis of unsaturated fatty acids	11	13	0.85
00071	Fatty acid metabolism	10	10	1.00
00340	Histidine metabolism	10	11	0.91
00290	Valine, leucine and isoleucine biosynthesis	10	12	0.83
03410	Base excision repair	10	27	0.37
00300	Lysine biosynthesis	9	9	1.00
00450	Selenocompound metabolism	9	9	1.00
00650	Butanoate metabolism	9	10	0.90
00760	Nicotinate and nicotinamide metabolism	9	11	0.82
03430	Mismatch repair	9	21	0.43
03440	Homologous recombination	9	27	0.33
00960	Tropane, piperidine and pyridine alkaloid biosynthesis	8	8	1.00
00790	Folate biosynthesis	8	10	0.80
00460	Cyanoamino acid metabolism	8	11	0.73
00310	Lysine degradation	7	7	1.00
00950	Isoquinoline alkaloid biosynthesis	7	7	1.00
00511	Other glycan degradation	7	8	0.88
00670	One carbon pool by folate	7	10	0.70
04140	Regulation of autophagy	7	10	0.70
00062	Fatty acid elongation	6	7	0.86
00565	Ether lipid metabolism	6	7	0.86
00905	Brassinosteroid biosynthesis	6	8	0.75

04122	Sulfur relay system	6	9	0.67
00073	Cutin, suberine and wax biosynthesis	6	10	0.60
00966	Glucosinolate biosynthesis	6	12	0.50
00909	Sesquiterpenoid and triterpenoid biosynthesis	6	16	0.38
00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	5	5	1.00
00590	Arachidonic acid metabolism	5	6	0.83
00904	Diterpenoid biosynthesis	5	7	0.71
00750	Vitamin B6 metabolism	4	5	0.80
00908	Zeatin biosynthesis	4	5	0.80
04710	Circadian rhythm – mammal	4	5	0.80
00740	Riboflavin metabolism	4	6	0.67
00730	Thiamine metabolism	4	7	0.57
00072	Synthesis and degradation of ketone bodies	3	3	1.00
00430	Taurine and hypotaurine metabolism	3	3	1.00
00603	Glycosphingolipid biosynthesis – globo series	3	3	1.00
04650	Natural killer cell mediated cytotoxicity	3	3	1.00
00660	C5-Branched dibasic acid metabolism	3	4	0.75
00591	Linoleic acid metabolism	3	4	0.75
00531	Glycosaminoglycan degradation	3	4	0.75
00902	Monoterpenoid biosynthesis	3	4	0.75
00903	Limonene and pinene degradation	3	4	0.75
00944	Flavone and flavonol biosynthesis	3	4	0.75
03450	Non-homologous end-joining	3	8	0.38
00514	Other types of O-glycan biosynthesis	2	2	1.00

00604	Glycosphingolipid biosynthesis – ganglio series	2	2	1.00
00232	Caffeine metabolism	2	2	1.00
00942	Anthocyanin biosynthesis	1	1	1.00
00785	Lipoic acid metabolism	1	2	0.50
00901	Indole alkaloid biosynthesis	1	2	0.50
02010	ABC transporters	1	2	0.50
00780	Biotin metabolism	1	3	0.33
Total		2090	2564	0.82

シロイヌナズナが持つ酵素遺伝子を基準とするとき、ポプラ cDNA の収集率(Pathway Coverage)は、ポプラ cDNA が対応した代謝パスウェイマップ上の遺伝子数 / シロイヌナズナタンパク質が対応した代謝パスウェイマップ上の遺伝子数 = $2,090 / 2,564 = 0.82$ となる。

表 2-2-3 完全長 cDNA の末端読み配列とゲノム塩基配列へのマッピング結果

cDNA 末端読み配列数	89,572
ゲノムへマップした配列数	87,767
遺伝子に対応した配列数	84,758
遺伝子に対応しなかった配列数	3,009
cDNA に対応した遺伝子数	14,208

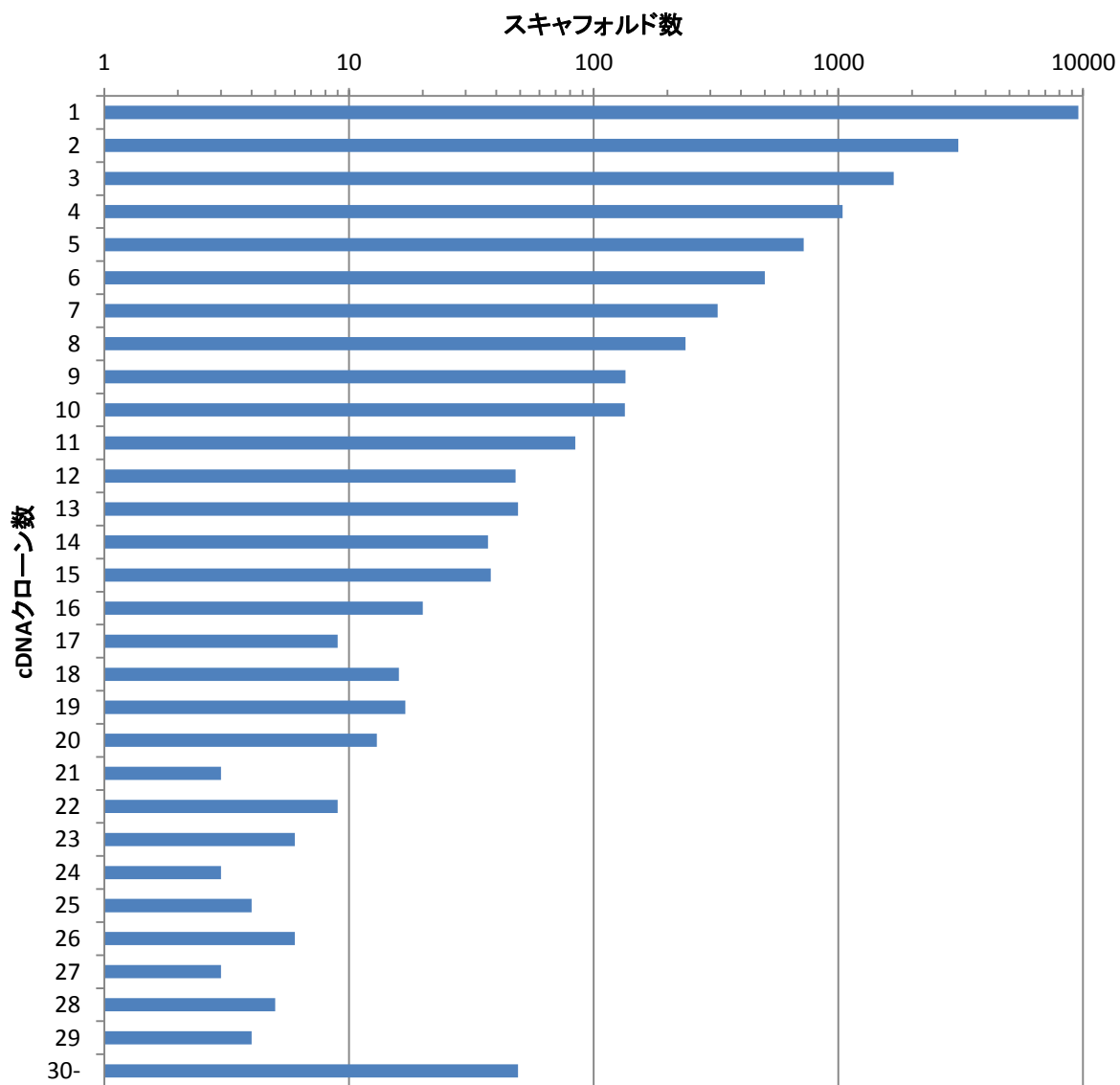


図 2-2-1 cDNA 配列アセンブリから構築したスキファールドを構成する cDNA クローン

スキファールドの 54%が 1 つの cDNA クローンで、80%は 3 つ以下の cDNA クローンで構成されており、cDNA 重複が少なく、効率的に多様な cDNA の収集が行われたことを表す。

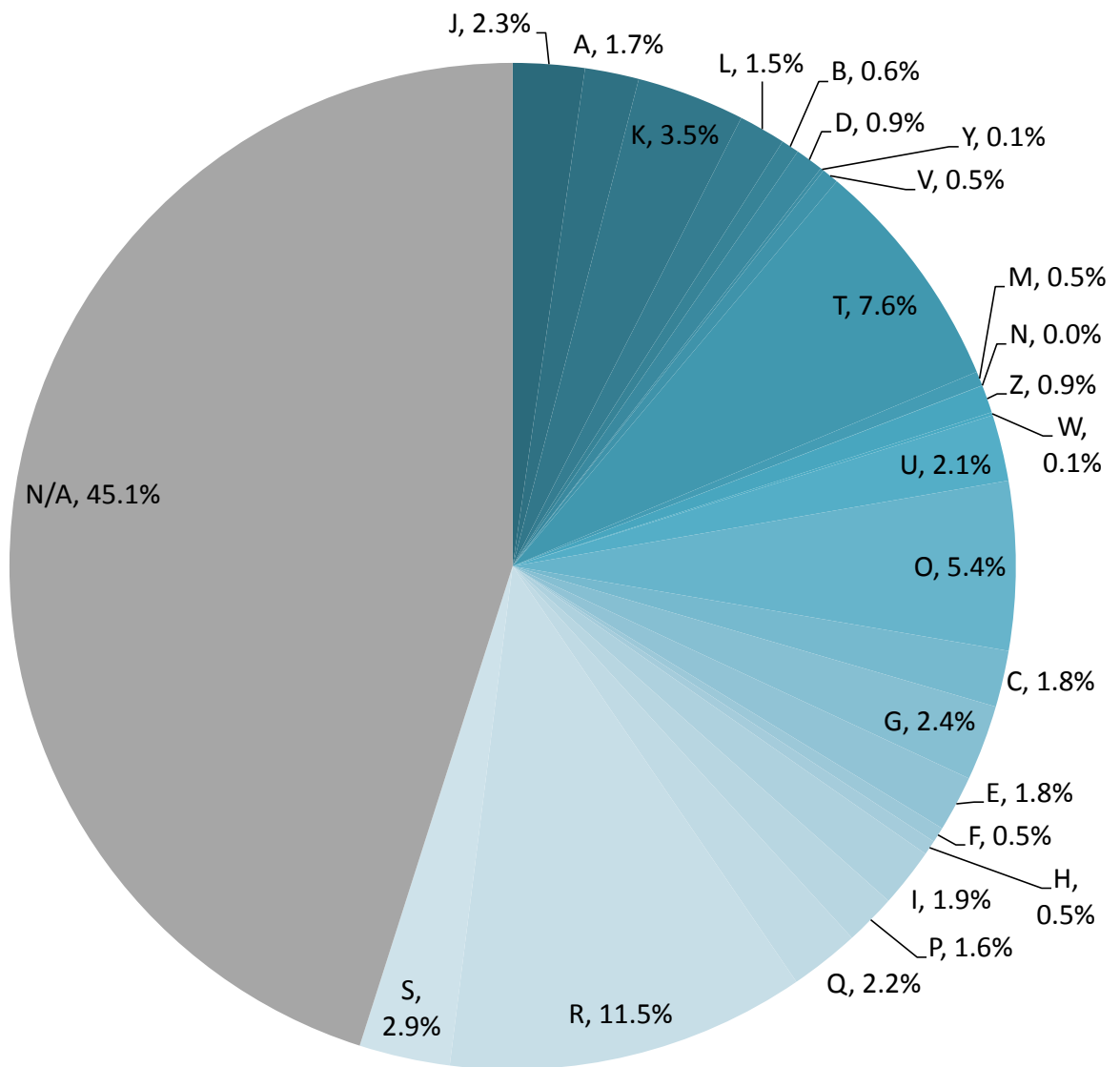


図 2-2-2 ポプラ転写物配列を用いた遺伝子機能分類

真核生物オルソログタンパク質配列データセット KOG を用いて、ポプラ転写物配列の遺伝子機能分類を行った。ポプラ代表転写物配列 41,612 の内、22,852 遺伝子(54.9%)が KOG に対応し、ポプラ遺伝子の機能分類を得た。分類項目については、以下に列挙する。A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control and mitosis; E, amino acid transport and metabolism; F, nucleotide

transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure, and biogenesis; K, transcription; L, replication and repair; M, cell wall/membrane/envelope biogenesis; O, posttranslational modification, protein turnover, and chaperone functions; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis, transport, and catabolism; T, signal transduction; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton; R, general functional prediction only; S, function unknown.

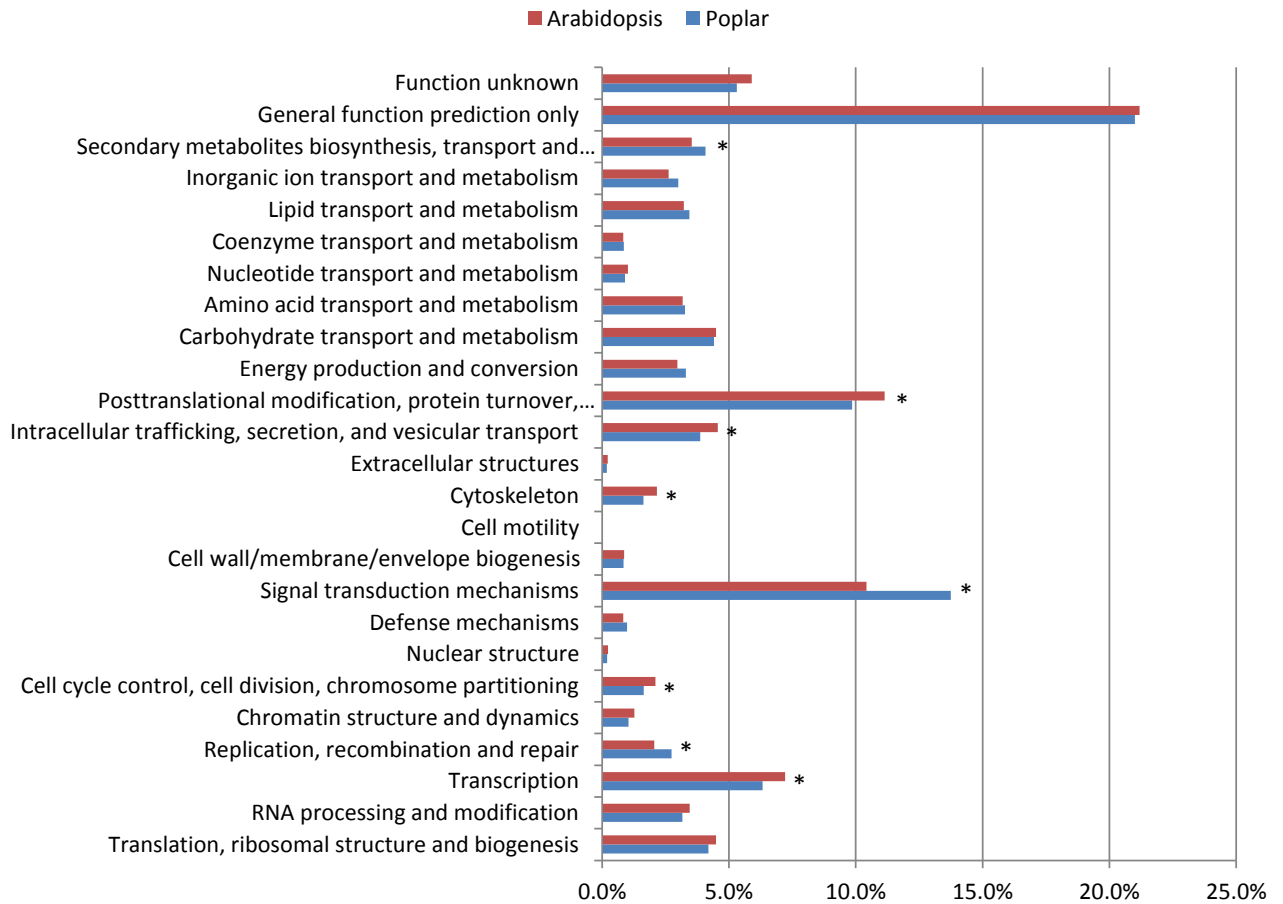


図 2-2-3 ポプラとシロイヌナズナの遺伝子機能分類の比較

ポプラとシロイヌナズナの遺伝子機能分類結果の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ、 p 値は 6.78×10^{-29} を示し、ポプラとシロイヌナズナの間で遺伝子機能の分類割合に有意な差があることが示唆された。さらに、有意差がある分類項目を検定するため、調整済み標準化残差分析を行い、8 個の分類項目に有意差が認められた(*がついている項目; $p < 0.05$)。

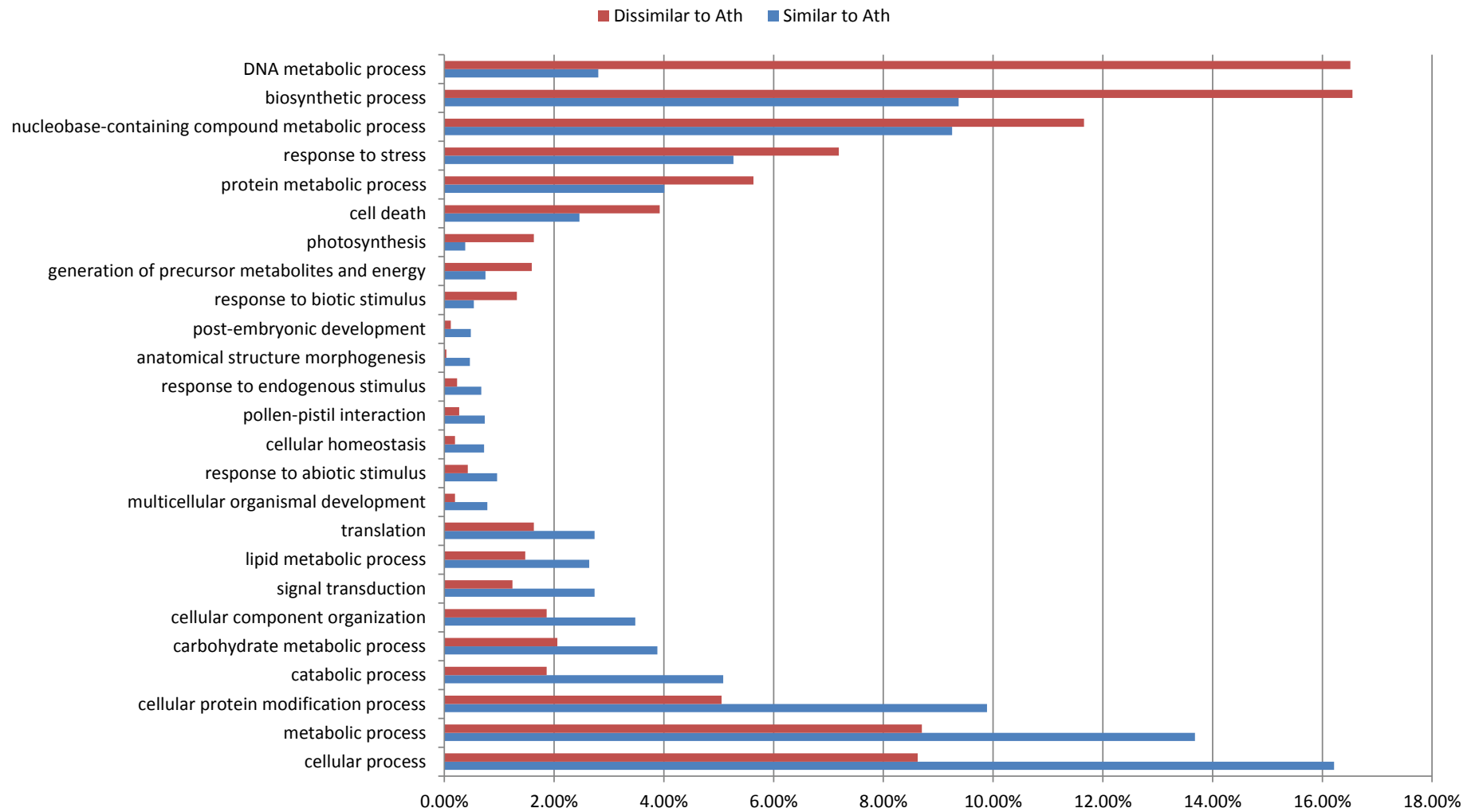


図 2-2-4 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ生物学的プロセス(biological process)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された($p=1.58 \times 10^{-39}$)。調整済み標準化残差分析の結果、有意差が認められた遺伝子オントロジ項目を示す($p < 0.01$)。"DNA metabolic process"、"biosynthetic process"、"nucleobase-containing compound metabolic process"、"response to stress"、"protein metabolic process"、"cell death"、"photosynthesis"、"generation of precursor metabolites and energy"、"response to biotic stimulus"の9個の遺伝子オントロジ項目が、シロイヌナズナ非類似配列で優位に高い割合を示した。

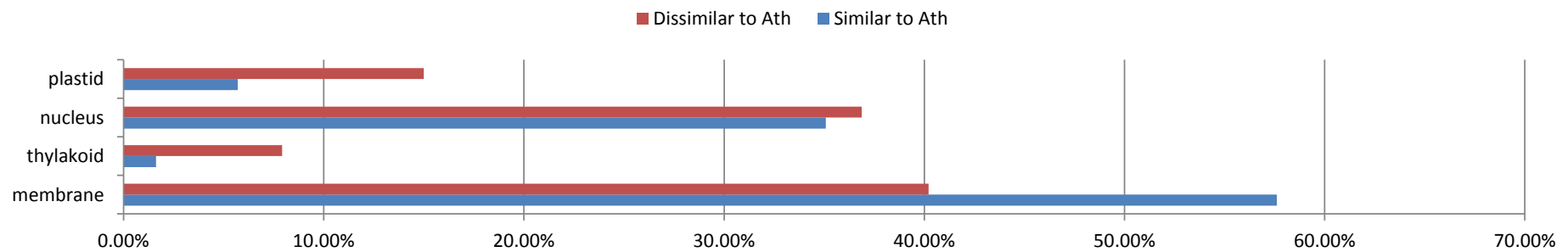


図 2-2-5 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ細胞構成要素 (cellular component)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ p 値は 9.57×10^{-42} を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、" plastid"、" nucleus"、" thylakoid "、" membrane"の 4 個の遺伝子オントロジ項目に有意差を確認した ($p < 0.01$)。

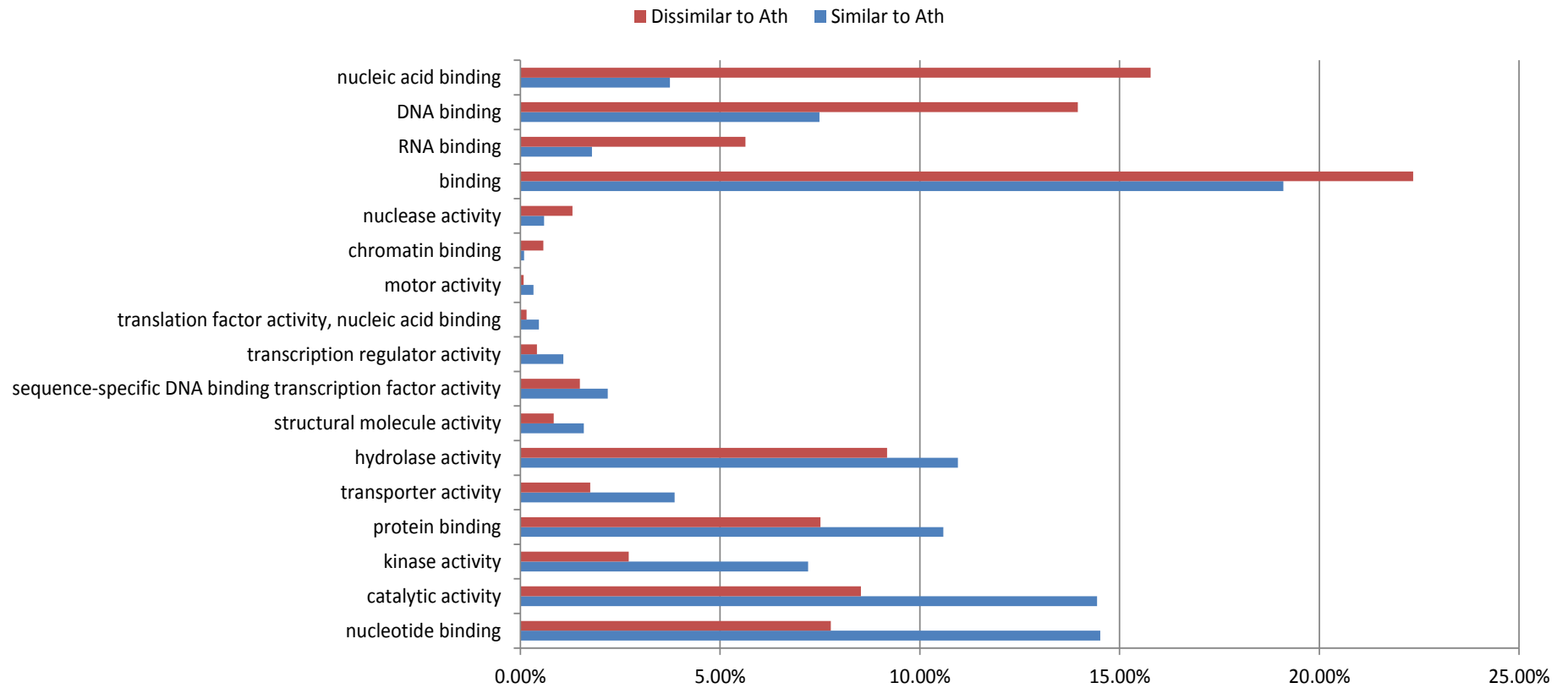


図 2-2-6 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ分子機能 (molecular function)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ p 値は 0.0 を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有

意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、17 個の遺伝子オントロジ項目に有意差を確認した($p < 0.01$)。

(3) キャッサバ完全長 cDNA の収集とその配列解析

【序論】

キャッサバ(*Manihot esculenta* Crantz)は、熱帯地域における重要作物の一つであり(Cock, 1982)、年間 2 億トン以上ものキャッサバが収穫され、5 億もの人々の主食として使われている(Food and Agricultural Organization of the United Nations, 2010)。また、キャッサバの根から抽出されるデンプンは、食品、製紙、繊維、合板などの原料として活用されて(Tonukari, 2004)、キャッサバを原料としたデンプン生産は他のデンプン作物と比較しても安価であることから注目を集めている(Amutha and Gunasekaran, 2001)。またキャッサバは、環境に対する適応性が高く、高温、多湿、貧栄養、酸性といった過酷な土壌での栽培が可能であり、度々干ばつが生じるような降水量が非常に少ない地域でもキャッサバ栽培が行われている(El-Sharkawy and Cadavid, 2002)。また、デンプンを貯蔵する根塊は土中で最低 2 年間は保持することができるため、飢饉への対応にも優れる一面も持つ(Raheem and Chukwuma, 2001)。上記のような悪条件下での生育が可能であるが、最適環境下における生育時に比べ生産性は低下し、細菌またはウイルスによる被害も問題になっている。またキャッサバは、高いデンプン含量と対照的に、食用するにあたって有毒であるシアン化合物を含む(Andersen et al., 2000)。

以前より行われてきた育種方法により、収量などに関する幾つかの問題点の改善が行なわれた。しかし、多くのキャッサバ系統が持つ遺伝的ヘテロ接合性や長い生活環のため、従来法による育種の進展は遅い(Fauquet and Tohme, 2004)。遺伝子地図の作成と量的表現形質遺伝子座の同定に関する研究報告はあるが、より詳細なマッピングや遺伝子同定を実現するためには、ゲノム塩基配列の決定を含む大きなコストが必要となる。したがって、機能ゲノム研究に基づく逆遺伝学的アプローチを欠かすことはできない。キャッサバの遺伝子概観の獲得や機能ゲノム研究の推進のための重要なツールの 1 つとして cDNA の収集およびその配列決定がある。しかし、その重要性にもかかわらず、キャッサバの cDNA 配列データの公共データベースへの登録数は、3 万程度に過ぎず、トウモロコシの 300 万、イネの 200 万に比べ、あまりにも小規模といえる。2012 年に公開されたキャッサバゲノム概要塩基配列は、想定されるゲノムサイズの 60%程度の解読であ

り(Prochnik et al., 2012)、ゲノム注釈付けに必要な EST の質、量ともに不十分であるため、RNA 転写領域予測にも不正確な領域定義が多く含まれると考えられる。

完全長 cDNA は、転写産物の機能解析だけでなく、転写開始位置の把握やゲノム塩基配列の修正、遺伝子領域の同定にも有効である。したがって本研究では、ビオチン化キャップトラッパー法(Carninci et al., 2000)を用いて、キャッサバ完全長 cDNA ライブラリを作製し、収集した cDNA 配列データとキャッサバのゲノム塩基配列由来の予測遺伝子との比較を行うことにより、新規転写単位(遺伝子座)および遺伝子モデル(転写産物)を同定した。遺伝子機能予測や遺伝子機能情報の確認によるゲノム注釈の改善だけでなく、様々なタンパク質データセットや生物種間共通遺伝子(オルソログ)、遺伝子オントロジ(Ashburner et al., 2000)を用い、収集したキャッサバ完全長 cDNA に注釈付けを行った。また、シロイヌナズナの遺伝子情報と比較することで、キャッサバ固有の遺伝子を分類し、その遺伝子機能の傾向を把握した。

【方法】

植物材料とストレス処理

通常環境下で生育または様々なストレス処理を行ったキャッサバ(*Manihot esculenta* 系統 MTAI16)から葉と根を採取した(表 2-3-1)。葉の採取に用いた植物体は、7~12 週齢まで温室内でプラスチック鉢に植えて育成した。根の採取には、圃場植えの 9 月齢の植物体を使用した。

高アルミニウム濃度-低 pH 処理については、pH4.2、200 μ M AlCl₃ と 200 μ M CaCl₂ 溶液を使った。高温処理については、植物体を 42°C で保温した。乾燥ストレス処理については、植物体をポットから取り出し、蒸留水で洗浄後、水気を除き、28°C で放置した。キャッサバの根塊部は、収穫すると品質の劣化が始まる(収穫後の生理的変敗)。この品質変化中における RNA を獲得するため、9 月齢植物体の根を収穫し、切断部をプラスチックシートで覆い、28°C で放置した。

各ストレス処理後の葉のサンプリングは、3、6、24、72 時間で行った。高アルミニウム濃度-低 pH 処理の根のサンプリングは、6、24、48 時間で、収穫後の生理的変敗処理の根は、24、48、120 時間で行った。すべてのサンプルは、RNA 抽出のため、液体窒素で凍結させた。

RNA 単離と完全長 cDNA ライブラリの構築

トータル RNA はフェノール-グアニジン・イソチオシアン酸溶液(TRIzol[®]、Invitrogen 社、カリフォルニア)を用いて抽出した。上記のすべてのサンプルに由来するトータル RNA の混合物は、 μ MACS mRNA 単離キット(Miltenyi Biotec 社、ドイツ)を用いて、さらに精製し、ポリアデニン鎖(poly A tail)が付いた RNA を cDNA ライブラリ作製に使用した。完全長 cDNA ライブラリは、トレハロース熱活性逆転写酵素を用いたビオチン化キャプトラッパー法(Carninci et al., 2000)によって作製した。

完全長 cDNA の末端読み配列決定

各 cDNA クローンの DNA は、TempliPhi DNA 増幅キット(GE Healthcare 社、英国)を使用した RCA 法(Dean et al., 2001)により増幅し、384 ウェルプレートで準備した。末端読み配列決定は、AB3700 DNA 解析装置(Applied Biosystems 社、カリフォルニア)により行い、5'末端読み配

列決定には、M13-21 プライマー(5'-TGTAACGACGGCCAGT-3')、3'末端読み配列決定には、1233 プライマー(5'-AGCGGATAACAATTTTCACACAGGA-3')を用いた。

配列データの調整と配列アセンブリ作成

シーケンサから出力された波形データは、ソフトウェア phred (Ewing et al., 1998)を用いてベースコールした。配列両端の低品質部については、スコア 20 以上の塩基が 20 塩基以上続くようになるまで除去した。併せて、配列マスクソフトウェア cross_match (Ewing et al., 1998)を使用してベクター配列部を除去した。低品質、不要部の除去処理後に残った塩基配列長が 100 未満の配列は除外した。さらに、大腸菌ゲノム塩基配列に対し、塩基配列類似性検索ソフトウェア BLASTN を実行し、 10^{-100} 未満の期待値(E value)を満たす配列データを除外した。上記の調整を行った EST を日本 DNA データバンク(DNA Data Bank of Japan; DDBJ) (Kodama et al., 2012)へ登録し、以降のゲノム解析における基本データとした。

調整済みの cDNA 配列データを配列連結(アセンブル)ソフトウェア CAP3 (Huang and Madan, 1999)を用いてアセンブルし、配列アセンブリを得た。アセンブルにより、配列の一致、整列によって連結されたコンティグ配列(contiguous sequence)を得ることができる。CAP3 は、初期設定パラメータにて実行した。

スキヤフォールドの構築

非冗長な cDNA 配列セットを得るため、上記 EST のアセンブリ配列をクローン名に基づきクラスタリングした。CAP3 が出力するファイルの 1 つである ace ファイルは、投入された配列の連結結果を記す。この ace ファイルを分析することで、コンティグ配列を構成する cDNA クローンの組み合わせを解釈し、転写産物毎のスキヤフォールド(scaffold)を構築した。スキヤフォールドを構成する両末端の配列(コンティグ配列または cDNA 端読み配列)は 20 文字の 'N' で連結した。以降の解析で塩基配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を使用するが、BLASTN が検索時に使用する検索範囲は 11 塩基であるため、スキヤフォールド内の 20 文字の 'N' が配列検索を妨げることはない。

完全長 cDNA ライブラリの品質

本研究では、両末端読み配列データを持つ cDNA を対象に完全長 cDNA 率を算出した。3'末端読み配列にポリアデニン鎖が確認でき、既知タンパク質の配列データセット NCBI-nr(Sayers et al., 2012)との配列類似性検索において、 10^{-30} 未満の期待値(E value)、正方向の翻訳枠、かつメチオニン(M)から整列する cDNA クローンを完全長 cDNA と見なした。

代謝パスウェイマッピング

収集した完全長 cDNA の代謝パスウェイへのマッピングには、分子間ネットワークデータベース KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2012)を用いた。キャッサバ cDNA 配列データを問い合わせ配列とし、KEGG 自動注釈付けサービス KAAS(Moriya et al., 2007)へ投入した。対象データセットに Ath(シロイヌナズナ)を、対応付け方法に SBH (single-directional best hit)法を選択して実行した。

完全長 cDNA 配列のゲノムマッピングと新規遺伝子領域の探索

キャッサバのゲノム塩基配列は植物ゲノム情報ウェブサイト Phytozome (Goodstein et al., 2012)より獲得し、はじめに核酸配列類似性検索ソフトウェア BLASTN (Altschul et al., 1997)を用いて、完全長 cDNA 配列が対応するおおよそのゲノム領域の塩基配列データを獲得した。続いて、cDNA 配列整列ソフトウェア SIM4 (Florea et al., 1998)を用いて、完全長 cDNA 配列が対応する正確なゲノム位置、転写方向、遺伝子構造 (エクソン/イントロン構造)を得た。

得られた完全長 cDNA のゲノム上における物理的位置情報を Phytozome より獲得した既知のキャッサバゲノム注釈情報(転写単位、遺伝子モデル)と対比させ、新規転写単位(transcription unit、遺伝子座)、遺伝子モデル(gene model、mRNA)を探索した。末端読み配列のどちらか一方が、既知のゲノム注釈情報の転写単位領域より外側に伸長するように対応した場合を新規遺伝子モデルとし、両末端読み配列が既知のゲノム注釈情報と対応しなかった場合を新規転写単位とした。

全代表転写物配列データの準備

キャッサバのゲノム注釈情報由来の代表タンパク質コード配列データ(Coding sequence; CDS)に上記解析で検出した新規転写領域のスキャフォールドを加えた配列データセットをキャッサバの全代表転写物配列と定義した。以降、この配列のことをキャッサバ代表転写物配列と表現する。

比較ゲノム解析

キャッサバ代表転写物配列データをシロイヌナズナタンパク質との類似性の有無による2つのグループに分類し、比較解析を行った。分類方法を次に述べる。キャッサバ代表転写物配列データセットを問い合わせ配列とし、シロイヌナズナ情報ウェブサイト TAIR(Lamesch et al., 2012)よりダウンロードしたシロイヌナズナタンパク質配列データセット (TAIR10_pep_20101214) に対して、BLASTX 検索を行った。BLASTX は、期待値が 10^{-10} 未満(-e 1e-10)、単純配列のマスキング不使用(-F F)の条件で行った。本解析では、BLASTX 検索結果にしたがい、期待値が 10^{-10} 未満であったキャッサバ代表転写物配列をシロイヌナズナと類似すると判定した。逆に期待値が 10^{-10} 以上、または類似対象シロイヌナズナタンパク質が無かった(結果が No hits found であった)キャッサバ代表転写物配列をシロイヌナズナと類似せずとした。

遺伝子機能解析

キャッサバの遺伝子機能概観を獲得するため、生物種間共通遺伝子(オルソログ)クラスタ COG (Tatusov et al., 2000)と遺伝子オントロジ (Ashburner et al., 2000)を使用した。COG のタンパク質配列データセットは、NCBI の FTP サーバ(<ftp://ftp.ncbi.nih.gov/pub/COG/>)より獲得し、真核生物のオルソログタンパク質配列データセット KOG の 2003 年 3 月公開版を使用した。キャッサバ代表転写物配列データを問い合わせ配列とし、KOG に対する BLASTX 検索を行った。BLASTX は、期待値が 10^{-5} 未満(-e 1e-5)、単純配列のマスキング不使用(-F F)の条件で実行した。その配列類似性検索の結果から、最高スコアのタンパク質 ID を獲得し、そのタンパク質 ID と KOG ID および KOG 遺伝子機能分類 ID とを対応付けた。

同様に遺伝子オントロジへの対応付けについても、はじめにキャッサバ代表転写物配列データを問い合わせ配列とし、欧州バイオインフォマティクス研究所(European Bioinformatics Institute; EBI)提供のタンパク質配列データセット UniProt/TrEMBL (Dimmer et al., 2012)に対する BLASTX 検索を行った。BLASTX は、期待値が 10^{-5} 未満(-e 1e-5)、単純配列のマスキング不使用(-F F)の条件で実行した。その配列類似性検索の結果から、最高スコアのタンパク質 ID を獲得し、遺伝子オントロジウェブサイト(<http://www.geneontology.org/>)からダウンロードしたタンパク質 ID-遺伝子オントロジ ID 対応付けデータファイル gp_association.goa_uniprot を用い、タンパク質 ID と遺伝子オントロジ ID とを対応付けた。さらに、同じく遺伝子オントロジウェブサイト(<http://www.geneontology.org/>)からダウンロードした遺伝子オントロジ ID グループ化データファイル goslim_plant.obo を用い、対応付けた遺伝子オントロジ ID を集約した。

統計解析

キャッサバとシロイヌナズナの遺伝子機能分類結果、シロイヌナズナと類似-非類似キャッサバ代表転写産物グループの遺伝子オントロジ解析結果についての統計解析に際し、その 2 群の独立性を確認するため、ピアソンのカイ二乗検定(Pearson chi-square test)と併せて、調整済み標準化残差分析(adjusted standardized residual analysis) (Bewick et al., 2004)を行った。この統計解析手法は、2 群間の独立系の検定後、群を構成する各項目についての残差(観測値-期待値)に基づき、その項目に関して 2 群間における差異を検定するものである。式を次に示す。

$$e_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

$$v_{ij} = \left(1 - \frac{n_{i.}}{N}\right) \left(1 - \frac{n_{.j}}{N}\right)$$

$$d_{ij} = \frac{e_{ij}}{\sqrt{v_{ij}}}$$

O : 観測値

E : 期待値

e : 標準化残差

v : 残差分散

n_i : 行周辺度数(行周辺和)

n_j : 列周辺度数(列周辺和)

d : 調整済み標準化残差

標準化残差は、近似的に平均 0、分散 1 の標準正規分布に則る。したがって、この標準化残差は標準正規分布における Z スコアと見なせる。

【結果と考察】

cDNA クローンの両末端読み配列の決定と配列アセンブリ

キャッサバ(*Manihot esculenta*)の各種ストレス処理を施した植物体の葉と根を出発材料とし、ビオチン化キャップトラッパー法を用いて完全長 cDNA ライブラリを作製した。作製した cDNA ライブラリから cDNA を単離し、両末端からの端読み配列を決定した。各配列データから低品質部、ベクター部を削除、大腸菌の混入配列を除外し、35,400 配列(5'末端読み:18,790、3'末端読み配列:16,610 配列)を得た(表 2-3-2)。この配列データセットを以降のゲノム解析の基本データとし、公共データベース DDBJ へ登録した(登録番号 DB920056~DB955455)。cDNA クローンは、理化学研究所バイオリソースセンターに寄託した。

既知のタンパク質配列に対して高い類似性を示す 5'および 3'末端読み配列を持つ cDNA クローンについて、完全長 cDNA 比(完全長 cDNA 数/対象 cDNA 数)を計算したところ、0.84 であった。これらのクローンが完全なタンパク質コード領域を含んでいることを意味し、遺伝子機能や遺伝子構造の解析への使用について十分な完全長 cDNA ライブラリ品質であることを示した。

低品質部などを削除し得られたキャッサバ cDNA 配列データを、配列連結ソフトウェア CAP3 を用いてアセンブルし、6,355 のコンティグと 9,026 のシングレットを得た。さらに、独立した転写産物(mRNA)を単位とする配列セットを作成するため、この CAP3 によるアセンブル結果と cDNA 配列情報から 10,363 スキャフォールドを構築した(表 2-3-2)。このスキャフォールド構築結果から、スキャフォールドを構成する cDNA クローン数を確認したところ、スキャフォールドの 67% が 1 つの cDNA クローンで、90%以上のスキャフォールドは 3 つ以下の cDNA クローンで構成されており、cDNA 重複が少なく、効率的に多様な cDNA の収集が行われたことが確認できた(図 2-3-1)。

収集した cDNA の網羅性、および新規転写領域の探索

分子間ネットワークデータベース KEGG (Kyoto Encyclopedia of Genes and Genomes) を用いて、キャッサバ cDNA 配列データを代謝パスウェイにマップしたところ、123 代謝経路図上の 1,711 個の酵素に対応した(表 2-3-3)。シロイヌナズナタンパク質は、2,564 個の酵素に対応して

おり、本研究において、シロイヌナズナの酵素遺伝子の 67%に相当するキャッサバ完全長 cDNA を収集したことを示した(表 2-3-3)。

キャッサバゲノム塩基配列上における新規転写領域の探索のため、今回収集したキャッサバ完全長 cDNA の配列データをキャッサバゲノム塩基配列へマップしたところ、97.2%に相当する 34,432 配列がマップされ、ゲノム上の位置情報を獲得できた(表 2-3-4)。獲得できた各完全長 cDNA 配列データの位置情報に基づき、キャッサバのゲノム注釈情報(Phytozome v8.0 annotation)へ対応させたところ、キャッサバゲノムにコードされている遺伝子の 25.4%である 7,785 のキャッサバ遺伝子に対応した。キャッサバ遺伝子と対応しなかった 2,992 配列を対象にゲノム上の位置情報に基づくクラスタリングを行ったところ、672 個の新規遺伝子領域を検出した。キャッサバゲノム塩基配列にマップしなかった cDNA 配列を精査したところ、302 個のスキヤフォールドが対応し、これらも新規遺伝子領域の候補と考えられる。以上の結果から本研究によって、約 9,000 個の転写単位に対応するキャッサバ cDNA が収集されたことが示された。また、キャッサバのゲノム注釈情報と対応した配列について精査したところ、751 個の新規遺伝子モデルを検出した。本研究で収集した完全長 cDNA を解析することで新規転写単位および新規遺伝子モデルが検出された。これによって、既存のゲノム注釈情報を改善することができた。既存のゲノム注釈情報由来の 30,666 配列に、この解析で検出できた 974 の新規遺伝子の配列データを加えた配列データセット(31,640 配列)をキャッサバゲノム中の全転写領域の代表配列と定義した。以降、キャッサバ代表転写物配列と表現し、配列解析で使用した。

遺伝子機能解析

キャッサバの遺伝子機能概観を把握するため、上述の解析で得たキャッサバ代表転写物配列データを問い合わせ配列として、生物種間共通遺伝子(オルソログ)クラスタ COG の真核生物オルソログタンパク質配列データセット KOG に対する BLASTX 配列類似性検索を行い、その検索結果にしたがって、遺伝子機能を分類した。KOG は、3 動物種(センチュウ、ショウジョウバエ、ヒト)、1 植物種(シロイヌナズナ)、2 真菌種(パン酵母、分裂酵母)、1 微孢子虫種(*Encephalitozoon cuniculi*)のタンパク質配列データで構成されている。本研究で定義したキャッサバ代表転写物配

列 31,640 の内、18,187 遺伝子(57.5%)が KOG に対応し、キャッサバ遺伝子の機能分類を得た(図 2-3-2)。草本双子葉モデル植物シロイヌナズナとの比較解析を行うため、シロイヌナズナについても同様の遺伝子機能分類を行った。キャッサバとシロイヌナズナの遺伝子機能分類について、各植物での分類の割合の比較を図 2-3-3 に示す。キャッサバとシロイヌナズナの遺伝子機能分類結果の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ、p 値は 2.45×10^{-9} を示し、キャッサバとシロイヌナズナの間で遺伝子機能の分類割合に有意な差があることが示唆された。続いて差がある分類項目を検定するため、調整済み標準化残差分析を行ったところ、"Chromatin structure and dynamics"、"Cell cycle control, cell division, chromosome partitioning"、"Signal transduction mechanisms"、"Cytoskeleton"、"Energy production and conversion"の 5 個の分類項目について有意差が認められた($p < 0.05$)。

キャッサバが持つ遺伝子機能をより詳細に把握するため、キャッサバ代表転写物配列データセットを問い合わせ配列とし、シロイヌナズナタンパク質配列データセット(TAIR10_pep_20101214)に対する BLASTX 検索を行った。その結果、期待値が 10^{-10} 未満であった 27,915 キャッサバ代表転写物配列をシロイヌナズナと類似すると判定した。この配列群をシロイヌナズナ類似配列と表す。期待値が 10^{-10} 以上、または類似対象シロイヌナズナタンパク質が無かった(結果が No hits found であった) 3,725 キャッサバ代表転写物配列をシロイヌナズナと類似せずと分類した。この配列群をシロイヌナズナ非類似配列とした。

図 2-3-4 は、遺伝子オントロジの生物学的プロセス(biological process)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の対応遺伝子オントロジの割合の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ p 値は 0.0 を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、"DNA metabolic process"、"biosynthetic process"、"transport"、"protein metabolic process"、"nucleobase-containing compound metabolic process"、"cell death"、"response to stress"、"photosynthesis"の 8 個の遺伝子オントロジ項目が、シロイヌナズナ非類似配列で優位に高い割合を示し($p < 0.01$)、キャッサバ特有または多様化の可能性が示唆された。遺伝子オントロ

”cellular process”、”metabolic process”、” cellular protein modification process”などの汎用的な遺伝子機能については、シロイヌナズナ類似配列での分類割合が高かった。この結果から、これらの遺伝子機能については、シロイヌナズナとキャッサバとで共通していると考えられる。

図 2-3-5 は、遺伝子オントロジの細胞構成要素 (cellular component)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の遺伝子オントロジ対応結果について、ピアソンのカイ二乗検定を行ったところ p 値は 6.01×10^{-61} を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、" plastid"、"intracellular"、" thylakoid "、"endoplasmic reticulum"、" plasma membrane"、" membrane"の 6 個の遺伝子オントロジ項目に有意差を確認した ($p < 0.01$)。非類似配列群で色素体やチラコイドについての割合が高いことから、光合成、デンプン貯蔵などのキャッサバ特有の関連遺伝子機能またはメカニズムの多様化の可能性が示唆された。

図 2-3-6 は、遺伝子オントロジの分子機能(molecular function)へ対応した配列の割合を示す。シロイヌナズナ類似/非類似配列群の遺伝子オントロジ対応結果について、ピアソンのカイ二乗検定を行ったところ p 値は 0.0 を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、21 個の遺伝子オントロジ項目に有意差を確認した($p < 0.01$)。特に” nucleotide binding”、"catalytic activity"、"kinase activity"、"transporter activity"の 4 個の遺伝子オントロジ項目が、シロイヌナズナ類似配列で高い割合を示した。この結果から、これらの遺伝子機能については、シロイヌナズナとキャッサバとで共通していると考えられる。また、核酸結合タンパク質に関する遺伝子オントロジ項目が、シロイヌナズナ非類似配列で高い割合を示し、転写調節、シグナル伝達などに関連するキャッサバ特有のメカニズムの多様化の可能性が示された。

【図表】

表 2-3-1 RNA 抽出に使用した組織、処理条件

処理	生育期間	部位	処理後サンプリング時間
無処理	9、11、12 週間	葉	-
無処理	9 か月間	根	-
乾燥	7 週間	葉	3、6、24、72 時間
高温	9 週間	葉	3、6、24、72 時間
収穫後生理的変敗	9 か月間	根	24、48、120 時間
高アルミニウム濃度-低 pH	9 週間	葉	3、6、24、72 時間
高アルミニウム濃度-低 pH	9 か月間	根	6、24、48 時間

表 2-3-2 完全長 cDNA の収集と配列アセンブリの内訳

クローン数	19,450
端読み配列数 (5'/3'末端読み)	35,400 (18,790/16,610)
コンティグ数	6,355
シングレット数	9,026
スキヤフォールド数	10,363

表 2-3-3 各代謝パスウェイに対応するキャッサバ cDNA とシロイヌナズナの酵素遺伝子数の比較

	KEGG Pathway	Cassava cDNA	Arabidopsis genes	Pathway Coverage
03010	Ribosome	104	109	0.95
03040	Spliceosome	65	101	0.64
00190	Oxidative phosphorylation	58	67	0.87
04141	Protein processing in endoplasmic reticulum	57	74	0.77
03013	RNA transport	46	91	0.51
00230	Purine metabolism	40	81	0.49
04120	Ubiquitin mediated proteolysis	39	56	0.70
00240	Pyrimidine metabolism	34	70	0.49
04075	Plant hormone signal transduction	32	43	0.74
03015	mRNA surveillance pathway	30	47	0.64
00010	Glycolysis / Gluconeogenesis	29	29	1.00
03050	Proteasome	28	35	0.80
00520	Amino sugar and nucleotide sugar metabolism	26	37	0.70
00270	Cysteine and methionine metabolism	25	29	0.86
00500	Starch and sucrose metabolism	25	32	0.78
03018	RNA degradation	25	48	0.52
00710	Carbon fixation in photosynthetic organisms	24	24	1.00
00195	Photosynthesis	24	26	0.92
00620	Pyruvate metabolism	24	29	0.83
00260	Glycine, serine and threonine metabolism	24	30	0.80

04144	Endocytosis	24	35	0.69
04626	Plant-pathogen interaction	24	37	0.65
04145	Phagosome	23	26	0.88
00510	N-Glycan biosynthesis	23	32	0.72
04146	Peroxisome	23	33	0.70
00860	Porphyrin and chlorophyll metabolism	22	30	0.73
00330	Arginine and proline metabolism	22	34	0.65
00250	Alanine, aspartate and glutamate metabolism	21	26	0.81
00564	Glycerophospholipid metabolism	21	32	0.66
00630	Glyoxylate and dicarboxylate metabolism	20	23	0.87
00970	Aminoacyl-tRNA biosynthesis	20	25	0.80
03008	Ribosome biogenesis in eukaryotes	20	56	0.36
00020	Citrate cycle (TCA cycle)	19	19	1.00
03060	Protein export	19	26	0.73
00561	Glycerolipid metabolism	18	22	0.82
00900	Terpenoid backbone biosynthesis	17	29	0.59
03420	Nucleotide excision repair	17	37	0.46
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	16	23	0.70
04712	Circadian rhythm - plant	16	27	0.59
00940	Phenylpropanoid biosynthesis	15	21	0.71
00030	Pentose phosphate pathway	14	15	0.93
00350	Tyrosine metabolism	14	17	0.82
03022	Basal transcription factors	14	31	0.45
00053	Ascorbate and aldarate metabolism	13	14	0.93

00640	Propanoate metabolism	13	16	0.81
00051	Fructose and mannose metabolism	13	17	0.76
04130	SNARE interactions in vesicular transport	13	17	0.76
03020	RNA polymerase	13	27	0.48
00196	Photosynthesis – antenna proteins	12	12	1.00
00052	Galactose metabolism	12	15	0.80
00360	Phenylalanine metabolism	12	15	0.80
00480	Glutathione metabolism	12	16	0.75
00562	Inositol phosphate metabolism	12	18	0.67
00280	Valine, leucine and isoleucine degradation	12	19	0.63
00941	Flavonoid biosynthesis	11	12	0.92
00061	Fatty acid biosynthesis	11	13	0.85
00592	alpha-Linolenic acid metabolism	11	14	0.79
00040	Pentose and glucuronate interconversions	10	13	0.77
00410	beta-Alanine metabolism	10	14	0.71
04070	Phosphatidylinositol signaling system	10	16	0.63
00910	Nitrogen metabolism	10	17	0.59
00071	Fatty acid metabolism	9	10	0.90
00670	One carbon pool by folate	9	10	0.90
00920	Sulfur metabolism	9	11	0.82
01040	Biosynthesis of unsaturated fatty acids	9	13	0.69
00906	Carotenoid biosynthesis	9	17	0.53
03410	Base excision repair	9	27	0.33
00290	Valine, leucine and isoleucine biosynthesis	8	12	0.67

00100	Steroid biosynthesis	8	16	0.50
00380	Tryptophan metabolism	8	16	0.50
00130	Ubiquinone and other terpenoid–quinone biosynthesis	8	18	0.44
03030	DNA replication	8	32	0.25
00565	Ether lipid metabolism	7	7	1.00
00960	Tropane, piperidine and pyridine alkaloid biosynthesis	7	8	0.88
00300	Lysine biosynthesis	7	9	0.78
04122	Sulfur relay system	7	9	0.78
00340	Histidine metabolism	7	11	0.64
00600	Sphingolipid metabolism	7	12	0.58
00770	Pantothenate and CoA biosynthesis	7	16	0.44
03440	Homologous recombination	7	27	0.26
00950	Isoquinoline alkaloid biosynthesis	6	7	0.86
00450	Selenocompound metabolism	6	9	0.67
00650	Butanoate metabolism	6	10	0.60
00460	Cyanoamino acid metabolism	6	11	0.55
00760	Nicotinate and nicotinamide metabolism	6	11	0.55
00909	Sesquiterpenoid and triterpenoid biosynthesis	6	16	0.38
00563	Glycosylphosphatidylinositol(GPI)–anchor biosynthesis	6	21	0.29
03430	Mismatch repair	6	21	0.29
00945	Stilbenoid, diarylheptanoid and gingerol biosynthesis	5	5	1.00
00590	Arachidonic acid metabolism	5	6	0.83
00740	Riboflavin metabolism	5	6	0.83
00310	Lysine degradation	5	7	0.71

00511	Other glycan degradation	5	8	0.63
00790	Folate biosynthesis	5	10	0.50
04140	Regulation of autophagy	5	10	0.50
00591	Linoleic acid metabolism	4	4	1.00
00750	Vitamin B6 metabolism	4	5	0.80
04710	Circadian rhythm – mammal	4	5	0.80
00062	Fatty acid elongation	4	7	0.57
00730	Thiamine metabolism	4	7	0.57
00073	Cutin, suberine and wax biosynthesis	4	10	0.40
04650	Natural killer cell mediated cytotoxicity	3	3	1.00
00660	C5-Branched dibasic acid metabolism	3	4	0.75
00531	Glycosaminoglycan degradation	3	4	0.75
00902	Monoterpenoid biosynthesis	3	4	0.75
00944	Flavone and flavonol biosynthesis	3	4	0.75
00908	Zeatin biosynthesis	3	5	0.60
00904	Diterpenoid biosynthesis	3	7	0.43
00905	Brassinosteroid biosynthesis	3	8	0.38
03450	Non-homologous end-joining	3	8	0.38
00901	Indole alkaloid biosynthesis	2	2	1.00
00430	Taurine and hypotaurine metabolism	2	3	0.67
00903	Limonene and pinene degradation	2	4	0.50
00966	Glucosinolate biosynthesis	2	12	0.17
00942	Anthocyanin biosynthesis	1	1	1.00
00514	Other types of O-glycan biosynthesis	1	2	0.50

00604	Glycosphingolipid biosynthesis – ganglio series	1	2	0.50
00232	Caffeine metabolism	1	2	0.50
02010	ABC transporters	1	2	0.50
00072	Synthesis and degradation of ketone bodies	1	3	0.33
00603	Glycosphingolipid biosynthesis – globo series	1	3	0.33
00780	Biotin metabolism	1	3	0.33
00785	Lipoic acid metabolism	0	2	0.00
Total		1711	2564	0.67

シロイヌナズナが持つ酵素遺伝子を基準とするとき、キャッサバ cDNA の収集率(Pathway Coverage)は、キャッサバ cDNA が対応した代謝パスウェイマップ上の遺伝子数 / シロイヌナズナタンパク質が対応した代謝パスウェイマップ上の遺伝子数 = $1,711 / 2,564 = 0.67$ となる。

表 2-3-4 完全長 cDNA の末端読み配列とゲノム塩基配列へのマッピング結果

cDNA 末端読み配列数	35,400
ゲノムへマップした配列数	34,432
遺伝子に対応した配列数	31,440
遺伝子に対応しなかった配列数	2,992
cDNA に対応した遺伝子数	7,785

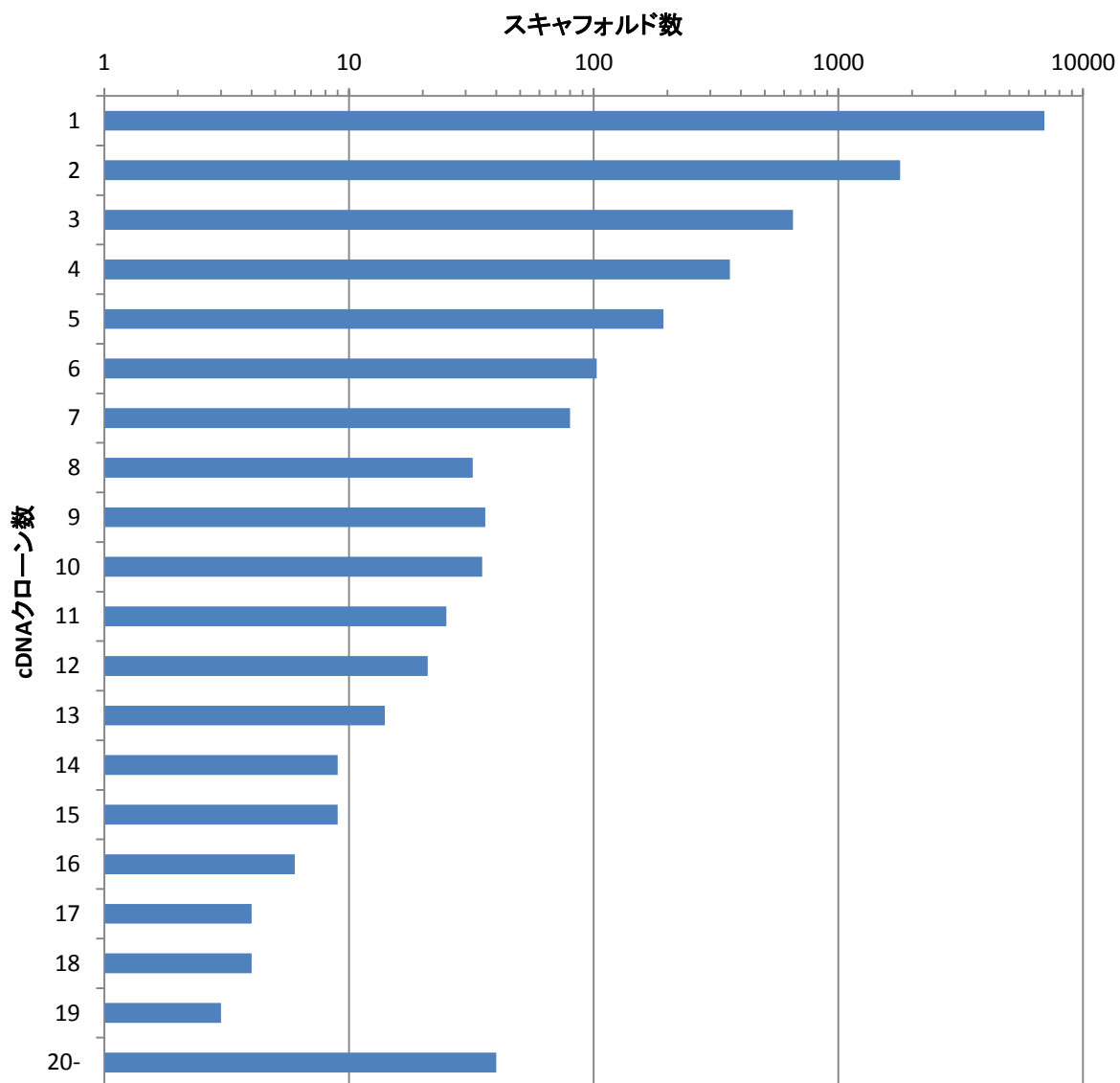


図 2-3-1 cDNA 配列アセンブリから構築したスキュフォールドを構成する cDNA クローン

スキュフォールドの 67%が 1 つの cDNA クローンで、90%は 3 つ以下の cDNA クローンで構成されており、cDNA 重複が少なく、効率的に多様な cDNA の収集が行われたことを表す。

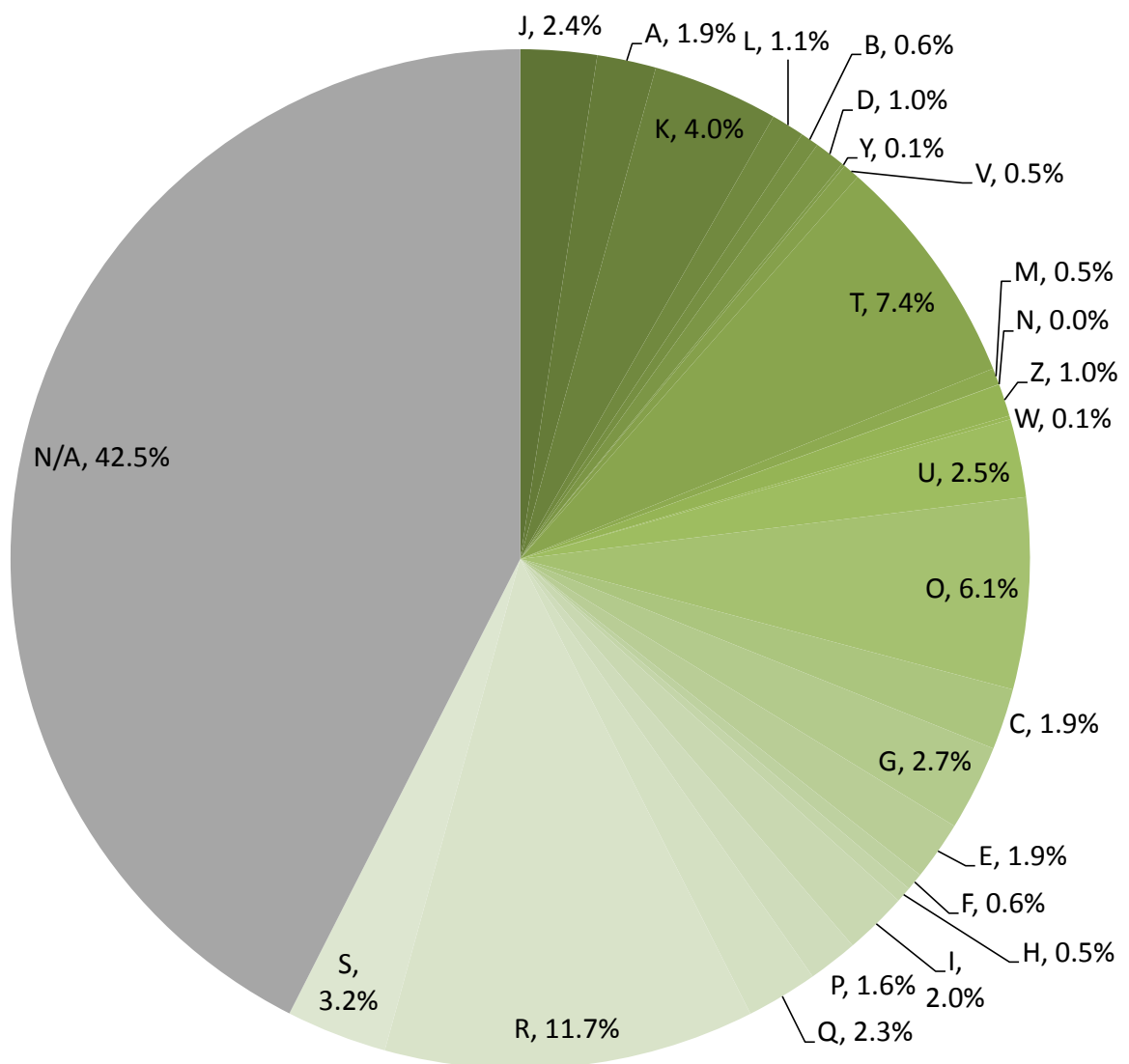


図 2-3-2 キャッサバ転写物配列を用いた遺伝子機能分類

真核生物オルソログタンパク質配列データセット KOG を用いて、キャッサバ転写物配列の遺伝子機能分類を行った。キャッサバ代表転写物配列 31,640 の内、18,187 遺伝子(57.5%)が KOG に対応し、キャッサバ遺伝子の機能分類を得た。分類項目については、以下に列挙する。A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control and mitosis; E, amino acid transport and metabolism; F,

nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure, and biogenesis; K, transcription; L, replication and repair; M, cell wall/membrane/envelope biogenesis; O, posttranslational modification, protein turnover, and chaperone functions; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis, transport, and catabolism; T, signal transduction; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton; R, general functional prediction only; S, function unknown.

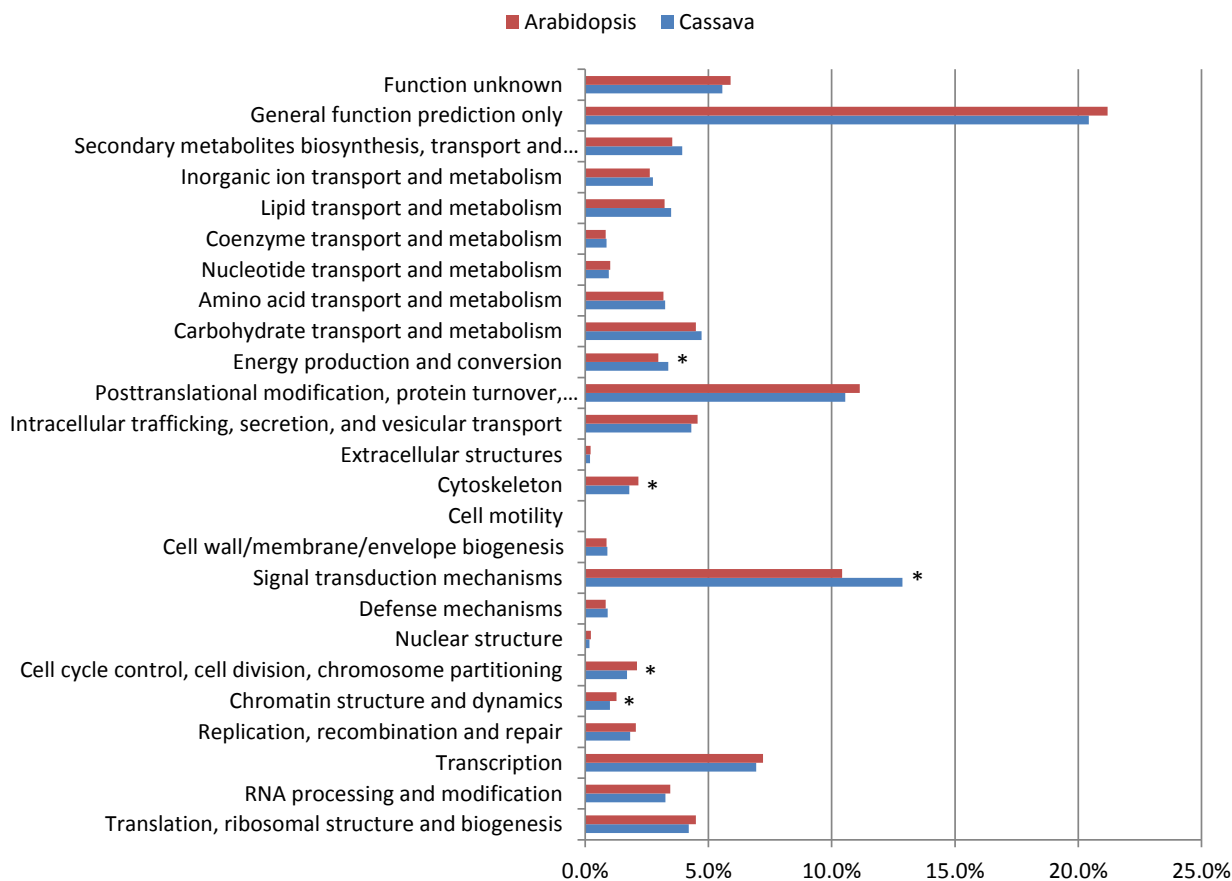


図 2-3-3 キャッサバとシロイヌナズナの遺伝子機能分類の比較

キャッサバとシロイヌナズナの遺伝子機能分類結果の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ、 p 値は 2.45×10^{-9} を示し、キャッサバとシロイヌナズナの間で遺伝子機能の分類割合に有意な差があることが示唆された。さらに、有意差がある分類項目を検定するため、調整済み標準化残差分析を行い、5 個の分類項目に有意差が認められた(*がついている項目; $p < 0.05$)。

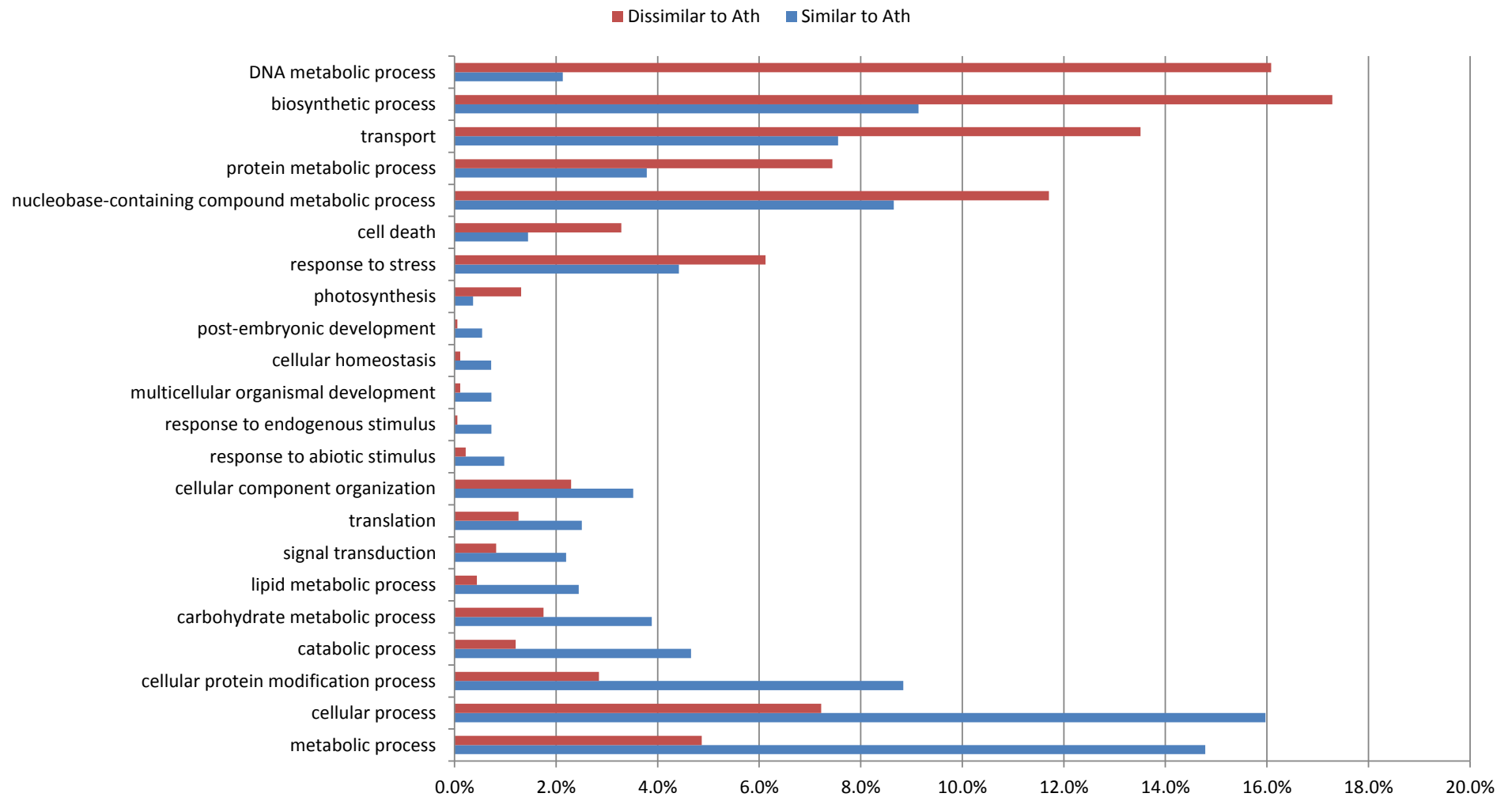


図 2-3-4 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ生物学的プロセス(biological process)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された($p=0.0$)。調整済み標準化残差分析の結果、有意差が認められた遺伝子オントロジ項目を示す($p<0.01$)。"DNA metabolic process"、"biosynthetic process"、"transport"、"protein metabolic process"、"nucleobase-containing compound metabolic process"、"cell death"、"response to stress"、"photosynthesis"の 8 個の遺伝子オントロジ項目が、シロイヌナズナ非類似配列で優位に高い割合を示した。

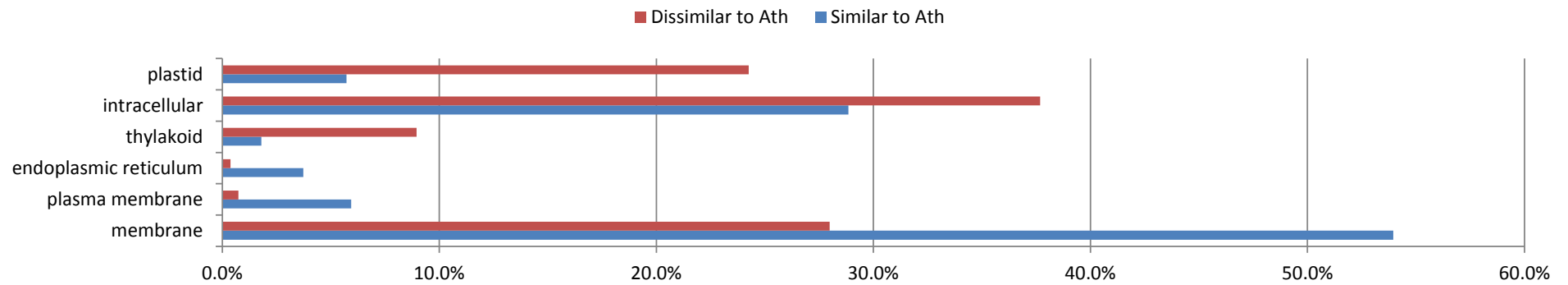


図 2-3-5 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ細胞構成要素 (cellular component)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ p 値は 6.01×10^{-61} を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、" plastid"、"intracellular"、" thylakoid "、"endoplasmic reticulum"、" plasma membrane"、" membrane"の 6 個の遺伝子オントロジ項目に有意差を確認した ($p < 0.01$)。

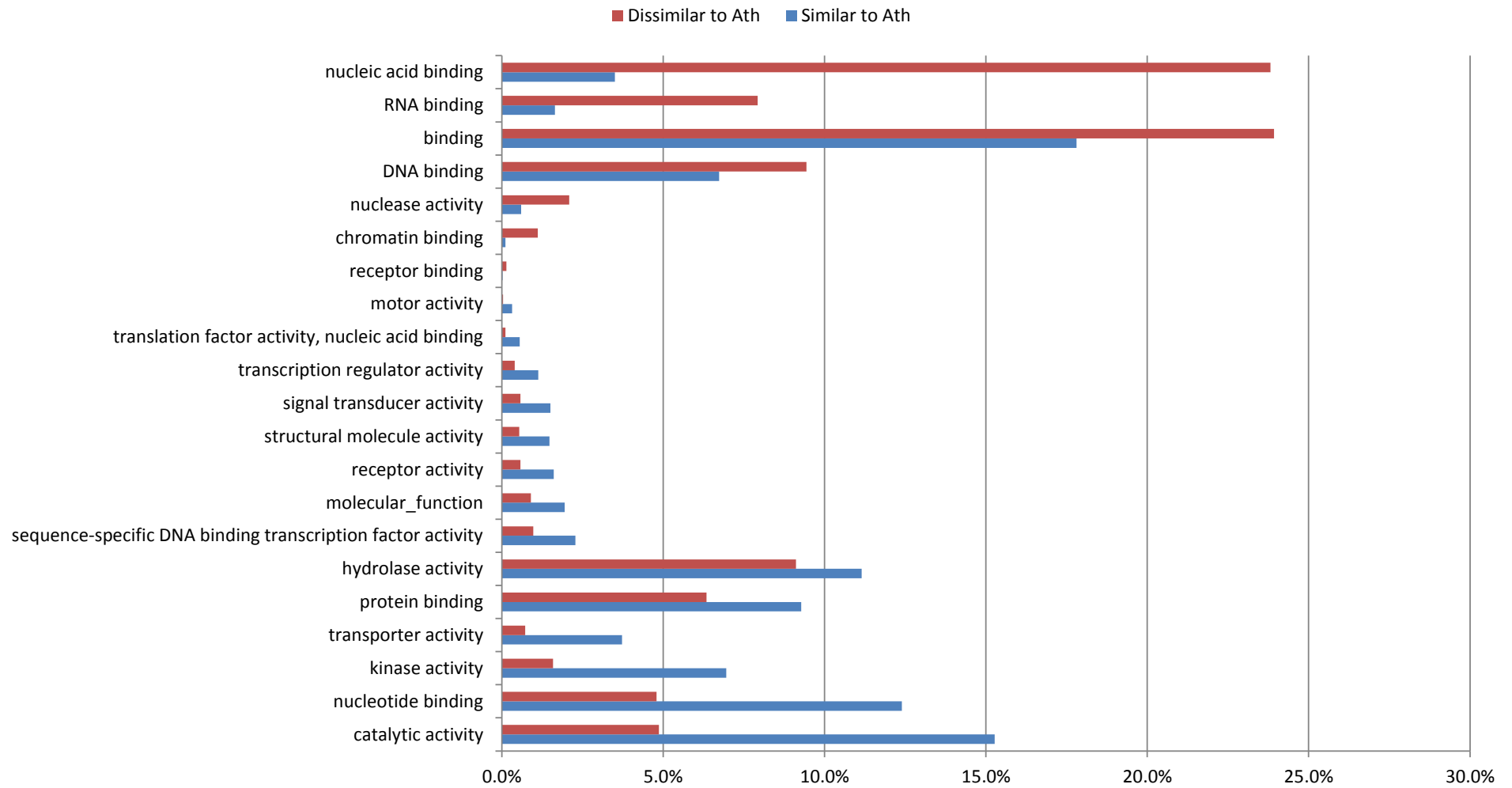


図 2-3-6 シロイヌナズナ類似配列と非類似配列の遺伝子オントロジ分子機能 (molecular function)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ p 値は 0.0 を示し、シロイヌナズナ類似配列と非類似配列との間での遺伝子オントロジ対応の割合に有意な差があることが示された。遺伝子オントロジ項目毎の有意差を検定するため、調整済み標準化残差分析を行ったところ、21 個の遺伝子オントロジ項目に有意差を確認した($p < 0.01$)。

第3章 ゲノム情報を活用した有用作物キャッサバの DNA 多型探索および DNA 多型と遺伝子機能の関連性の解析

【序論】

前章でも取り上げたように、キャッサバ(*Manihot esculenta* Crantz)は、熱帯地域における重要作物の一つであり(Cock, 1982)、年間 2 億トン以上ものキャッサバが収穫され、5 億もの人々の主食として使われている(Food and Agricultural Organization of the United Nations, 2010)。また、キャッサバの根から抽出されるデンプンは、食品、製紙、繊維、合板などの原料として活用されている(Tonukari, 2004)。キャッサバは、環境適応性が高く、高温、多湿、貧栄養、酸性といった耕作不適地での栽培が可能であることから、気候の変化、耕地の酷使、塩害などによる耕地減少に対応する作物としても注目されつつある。

近年、遺伝子探索の効果的な研究手法として、核酸配列データが盛んに収集されている(Mochida and Shinozaki, 2010)。実際、幾つかの植物種に対して行われた cDNA 収集とその配列決定は、機能ゲノム研究の推進に貢献した(Kikuchi et al., 2003; Nanjo et al., 2007; Taji et al., 2008; Umezawa et al., 2008; Soderlund et al., 2009)。キャッサバにおいても、幾つかの研究グループによって cDNA 収集が行われ(Anderson et al., 2004; Lopez et al., 2004; Lokko et al., 2007; Sakurai et al., 2007)、その成果は網羅的な発現遺伝子を実装したマイクロアレイ設計に活用され、トランスクリプトーム研究を推進した(Sojikul et al., 2010; An et al., 2012; Utsumi et al., 2012)。現在、キャッサバのゲノム概要塩基配列が公開され、推定ゲノムサイズ 770Mbp の 54%に相当する 419.5Mbp のゲノム塩基配列データが利用可能であり、このキャッサバゲノム概要塩基配列から 30,666 のタンパク質コード遺伝子が予測されている(Prochnik et al., 2012)。

分子マーカーは、ゲノム研究だけでなく育種研究にも重要であり、マーカー支援選抜(marker-assisted selection; MAS)を通じ、個体選抜の効率化に活用されている。植物遺伝学研究において、分子マーカーは集団構造解析、進化関連研究で活用され、モデル植物シロイヌナズナでは全ゲノムレベルでの遺伝構造解析が行われている(Cao et al., 2011)。さらに近年、一塩基多型(single nucleotide polymorphism; SNP)マーカーが注目され、動物やヒトのゲノム解析におい

て何千何万もの SNP マーカーを用いたハプロタイプ解析が行われており (International HapMap Consortium et al., 2007)、植物育種コミュニティにおいても関心が高まっている。

キャッサバにおいても遺伝子マッピングや分子マーカーの整備が進められ (Akano et al., 2002; Okogbenin and Fregene, 2002; Rabbi et al., 2012)、単純反復配列 (simple sequence repeat; SSR) の検出による分子マーカー整備やそれを用いた量的形質遺伝子座マッピング (quantitative trait loci mapping; QTL mapping) について報告されている (Raji et al., 2009; Sraphet et al., 2011)。さらに遺伝学研究と育種を推進するためには、SNP マーカーの様なより高密度なマーカー整備が求められる。SNP やゲノム挿入/欠損 (insertion and deletion; InDel) は、集団内で生じる突然変異であり、キャッサバでもこれらに関する研究が報告されている (Lopez et al., 2005; Ferguson et al., 2012)。これら DNA 多型の検出は、分子育種の推進だけでなく、遺伝子機能の理解にも重要である (Yamaguchi-Kabata et al., 2008)。

主要植物のゲノム情報は統合され、インターネットを介して円滑な閲覧が可能になり、研究推進に活用されている。遺伝マーカーと対応するゲノムまたは転写産物の配列データの統合化は、ゲノムワイドな遺伝学的研究を支援している。The Arabidopsis Information Resource (TAIR) (Lamesch et al., 2012) は、シロイヌナズナ研究者を始め多くの植物研究者に利用されている。同様に Gramene は、単子葉植物の比較ゲノム研究に関する情報を提供している (Liang et al., 2008)。ムギ類やトウモロコシなどのその他の植物についても情報を統合した有用なデータベースが整備されつつある。したがって、キャッサバ研究に関する情報を統合し、更なる研究促進を図るため、キャッサバ DNA 多型と関連遺伝子情報を編纂した新たなデータベースを構築した。

本章では、キャッサバの様々な系統由来の EST を活用することで同定した 10,000 ヶ所を超える DNA 多型とその DNA 多型と遺伝子機能、遺伝子重複との関係性について述べる。また、それら解析結果や関連する遺伝子注釈などを統合し、インターネット上に公開したデータベース Cassava Online Archive (<http://cassava.psc.riken.jp/>) についても述べる。

【方法】

キャッサバ転写産物配列データの獲得と系統毎の分類

キャッサバの DNA 多型を検出するための材料として、cDNA の部分配列データ(EST)を使用した。EST は、公共配列データベース GenBank(Benson et al., 2012)より獲得した。獲得した配列データは、配列マスクソフトウェア SeqClean(<http://compbio.dfci.harvard.edu/tgi/software/>)を用いて、低品質配列や不適切な単純繰り返し配列を除去した。残存するベクター部配列は、NCBI の UniVec データベース (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>)の 2011 年 11 月 22 日公開版を比較対象配列として、配列マスクソフトウェア cross_match (Ewing et al., 1998)を用いて検出し除去した。大腸菌の混入配列は、大腸菌ゲノム塩基配列(U00096)を比較対象配列として、配列類似性検索ソフトウェア BLASTN を用いて検出した。BLASTN 検索結果から期待値が 10^{-100} 未満の配列データを除外した。

続いて、上記の処理を経たキャッサバ EST を各 EST の GenBank 形式ファイルに記述されている系統情報により分類した。一部のキャッサバ系統については、EST 登録者へ連絡して系統名を確認した。また、キャッサバゲノム概要塩基配列(系統 AM560-2)から予測されたタンパク質および転写産物配列データ(JGI annotation v4.1)を植物ゲノム情報ウェブサイト Phytozome(Goodstein et al., 2012)より獲得した。

DNA 多型の同定と PCR プライマーの設計

上記で調整したキャッサバ EST とゲノム塩基配列から予測した転写産物配列データを配列整列化ソフトウェア CAP3(Huang and Madan, 1999)を用いてアセンブルした。CAP3 の実行は、初期設定条件で行った。本解析では、次の 4 つの条件を満たすものを DNA 多型と見なした。(1) コンティグ配列がキャッサバゲノム概要塩基配列にマップされる。(2)DNA 多型の塩基が'N'(不明)ではない。(3)SNP は検出された系統間での対立塩基数は 2 である。(4)低品質配列による誤検出を防ぐため、周辺 5 塩基以内に他の SNP が存在しない。

各 DNA 多型について、キャッサバゲノムスキファールド上における物理的位置を独自のプログラムを用い、ゲノム注釈情報から推算した。DNA 多型周辺のゲノム DNA 増幅用の PCR 用プ

ライマーペア配列をプライマー設計ソフトウェア Primer3(Rozen and Skaletsky, 2000)を用いて設計した。プライマー配列の設計に際しては、プライマー配列長が 18~25 塩基、増幅 DNA 長が 150~200 塩基、GC 含量が 45~65%、融解温度が 58~72°Cを満たすプライマーペア配列を選別した。

検出した SNP の検証

ファーガソンらの以前の報告において、キャッサバの SNP の検出とその検証が行われている (Ferguson et al., 2012)。この論文の検証結果と本研究で得た SNP 情報を対照させることで、本研究で同定した SNP およびその同定方法の検証を行った。比較対象のファーガソンらの論文より、SNP 情報をダウンロードし、彼らの SNP 同定時に使用された SNP 周辺配列データをキャッサバゲノム配列データに整列させることで、SNP の物理的な位置と SNP アリルを獲得した。その検証済み SNP データと本研究で同定した SNP の物理的な位置と SNP アリルの一致を確認した。

DNA 多型と遺伝子機能解析

検出した SNP について、トランジション置換(プリン塩基から別のプリン塩基、またはピリミジン塩基が別のピリミジン塩基へ変わる置換)ートランスバージョン置換(プリン塩基からピリミジン塩基、またはピリミジン塩基からプリン塩基へ変わる置換)、非同義置換(SNP によって翻訳されるアミノ酸が変わる)ー同義置換(SNP によって翻訳されるアミノ酸が変わらない)、ナンセンス置換 (premature stop substitution; アミノ酸が終止コドンに置換)ー読み過ごし置換 (read-through substitution; 終止コドンがアミノ酸に置換)を上記配列アセンブリとゲノム注釈情報を用いて解析した。ゲノム注釈情報に基づくキャッサバ遺伝子モデル(転写産物)の範囲内の SNP に関しては、独自のプログラムによって、5'側、3'側非翻訳領域(un-translated region; UTR)、タンパク質コード領域(protein coding region; CDS)の 3 つに分類した。SNP が CDS 中にある場合は、さらに翻訳されるアミノ酸が非同義もしくは同義置換かを確認した。同様に、ナンセンス置換と読み過ごし置換についても確認した。キャッサバ遺伝子モデルのタンパク質ドメイン情報

については、Phytozome 提供のタンパク質ドメインデータベース Pfam(Punta et al., 2012)との対応情報 (Mesculenta_147_annotation_info.txt) を使用した。

統計解析

キャッサバの重複遺伝子-非重複遺伝子、SNP によるアミノ酸変異-非変異といった 2 群間における比較解析に際し、その 2 群の独立性を確認するため、ピアソンのカイ二乗検定(Pearson chi-square test)と併せて、調整済み標準化残差分析(adjusted standardized residual analysis) (Bewick et al., 2004)を行った。この統計解析手法は、2 群間の独立系の検定後、群を構成する各項目についての残差(観測値-期待値)に基づき、その項目に関して 2 群間における差異を検定するものである。式を次に示す。

$$e_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

$$v_{ij} = \left(1 - \frac{n_{i.}}{N}\right) \left(1 - \frac{n_{.j}}{N}\right)$$

$$d_{ij} = \frac{e_{ij}}{\sqrt{v_{ij}}}$$

O : 観測値

E : 期待値

e : 標準化残差

v : 残差分散

$n_{i.}$: 行周辺度数(行周辺和)

$n_{.j}$: 列周辺度数(列周辺和)

d: 調整済み標準化残差

標準化残差は、近似的に平均 0、分散 1 の標準正規分布に則る。したがって、この標準化残差は標準正規分布における Z スコアと見なせる。

【結果と考察】

キャッサバ転写産物配列データの系統毎分類と DNA 多型探索

キャッサバの DNA 多型を検出するための材料として、cDNA の部分配列データ(expressed sequence tag; EST)を使用した。キャッサバの EST80,631 配列を公共配列データベース GenBank(Benson et al., 2012)より獲得し、混入する大腸菌やベクター部の配列を除去した結果、有効な配列データとして 80,523 配列を得た。獲得した EST の配列長の分布を図 3-1 に示す。GenBank 形式記述中の系統情報(cultivar tag)および配列情報登録者への確認により、16 系統または cDNA ライブラリ(CAS36.01、CAS36.04、CM21772、CM523-7、MCol22、IAC 12.829、KU50/MTAI16、MBra685、MCol1522、MNga2、MPer183、Mirassol、SG107-35、‘Sauti, Gomani, Mbundumali, TME 1, and Mkondezi’、‘TMS30572 and CM2177-2’、不明)に由来することが明らかになった。この EST 配列データにキャッサバゲノム概要塩基配列(系統 AM560-2)由来の予測転写産物配列データを加えた 17 系統、114,674 配列を使用して以降の解析を行った(表 3-1)。

上記配列データを配列整列化ソフトウェア CAP3(Huang and Madan, 1999)を使用してアセンブルした結果、16,363 コンティグと 17,789 シングレットを得た。このアセンブリには、96,885 配列が使用され(表 3-1)、コンティグを構成する配列数の分布は 2 から 707 であり、コンティグを構成する平均配列数は 5.9 であった(図 3-2)。この配列アセンブリを 1 塩基ずつ精査し、方法で示した条件を満たす、10,546 ヶ所の SNP と 674 ヶ所の InDel を同定した。配列数が少なかった系統 MNga2、‘TMS30572 and CM2177-2’からは DNA 多型が検出されなかった。コンティグを構成する配列数毎の DNA 多型検出数とコンティグあたりの平均 DNA 多型検出数を図 3-3 に示す。コンティグあたりの平均 DNA 多型検出数は 3.8 で、2.6 から 6.2 の範囲にあった。この結果は、コンティグを構成する配列数と同定される DNA 多型数に特に偏りがなかったことを示し、解析に用いた配列のシーケンスエラーによる誤った DNA 多型の同定ではないことを示す。検出した 10,546 ヶ所の SNP の内、8,794 ヶ所について、SNP 周辺ゲノム領域を増幅させるための PCR プライマーペア配列を設計することができた。また、SNP が検出されたキャッサバ遺伝子モデルは、3,252 個であり、SNP 頻度は、1 遺伝子モデルあたり 3.2 ヶ所、1,072.5 塩基に 1 ヶ所であった(表 3-2)。同様に、検出した 674 ヶ所の InDel の内 522 ヶ所について、InDel 周辺ゲノ

ム領域を増幅させるための PCR プライマーペア配列を設計することができた。また、InDel が検出されたキャッサバ遺伝子モデルは、583 個であり、InDel 頻度は、1 遺伝子モデルあたり 1.2 ヶ所、3,291.4 塩基に 1 ヶ所であった(表 3-2)。キャッサバは 3 万個以上の遺伝子を持つと考えられていることから(Prochnik et al., 2012)、この結果は、10 万以上の SNP と 4 万以上の InDel が見つかる可能性を示唆した。過去にキャッサバの DNA 多型検出について 2 件の報告があり、その検出数はそれぞれ、186 と 2,954 であった(Lopez et al., 2005; Ferguson et al., 2012)。本研究では、その同定数を大きく上回り、キャッサバにおいて最大規模であった。

検出した SNP の検証

ファーガソンらの以前の報告において、SNP の検出とその検証が行われている(Ferguson et al., 2012)。この論文の検証結果と本研究で得た SNP 情報を対照させることで、本研究で同定した SNP およびその同定方法の検証を行った。比較対象のファーガソンらの論文のオンラインサプリメントファイルより、彼らが同定した 1,190 ヶ所の SNP の周辺ゲノム塩基配列と各 SNP 周辺ゲノム塩基配列における SNP、アليل情報を獲得した。この SNP 周辺ゲノム塩基配列データをキャッサバのゲノム概要塩基配列に整列させ、ゲノム概要塩基配列における SNP 位置情報を獲得した。この内の 103 個は、本研究で同定した SNP と位置情報が一致し、そのすべての SNP のアليلに齟齬は生じなかった(表 3-3)。したがって、この結果は本研究の同定方法が有効であることを示した。

キャッサバ系統間で多様性を示す遺伝子と SNP の特徴

SNP について、置換が起こり得る 6 種類の塩基組み合わせについての分類を行った結果を表 3-4 に示す。C/T、G/A、A/C、A/T、C/G、T/G の組み合わせになる SNP は、それぞれ 2,893、2,952、1,108、1,285、996、1,312 ヶ所であった。トランジション置換(C/T または G/A)とトランスバージョン置換(C/G、A/T、C/A または T/G) は、それぞれ 5,845 と 4,701 ヶ所であった。トランジション置換ートランスバージョン置換比は、1.24 であり、以前に報告された 1.27(Lopez et al., 2005; Ferguson et al., 2012)と非常に近い値であった。他の植物のトランジション置換ー

トランスバージョン置換比は、シロイヌナズナとイネで報告があり、それぞれ 1.5、2.1 であった。

500 配列以上の EST を獲得することができた 8 系統(AM560-2、CM523-7、MCol22、KU50、MBra685、MCol1522、MPer183、SG107-35)について、SNP が検出された遺伝子モデルに対して、各々 2 系統間において SNP アリルが検出された遺伝子モデルとの比(該当 2 系統間で SNP アリルが存在した遺伝子モデル数/SNP 検出に該当 2 系統の EST が関与した遺伝子モデル数)を確認した(表 3-5)。この値は、キャッサバの系統間における遺伝子の違いを示しており、遺伝学的研究または育種に有意義である。例えば、遺伝地図作成や QTL 解析におけるマッピング集団作製に際して効果的な系統選択を補助するものである。

SNP による非同義—同義塩基置換とタンパク質機能

上述のように、10,546 ヶ所の SNP と 674 ヶ所の InDel を検出した(表 3-2)。検出した SNP の 6,613 ヶ所(62.7%)は、タンパク質コード領域(protein coding region; CDS)に位置していた(表 3-6)。シロイヌナズナは 78.0% (Cao et al., 2011)、イネは 73.4% (McNally et al., 2009)の SNP が CDS 中に位置し、キャッサバよりも大きな割合であった。これは、本解析で使用した配列データの多くが部分長 cDNA 由来の EST で構成されており、5'末端側よりの配列データの割合が小さく、その結果、CDS 中から検出される SNP の割合が減少したものと考えられる。CDS 中に位置した InDel の割合は 20.0%と SNP の割合よりも非常に小さかった。これは、CDS 中の InDel がタンパク質翻訳時のフレームシフトを生じさせ、InDel により変異が生じた翻訳産物は SNP よりも、細胞内活動に大きな影響を与えやすく、変異が後代に保存されにくいためと考えられる。

CDS 内で検出された SNP のアミノ酸翻訳フレーム内の位置を確認したところ、コドンの第 1、第 2、第 3 塩基目の SNP 数は、それぞれ 1,638、1,240、3,735 であった。基本的に CDS における塩基置換は、非同義置換になる頻度が同義置換よりも小さく、さらに、コドンの同義アミノ酸の冗長性は主に第 3 塩基目に見られる。したがって、第 3 塩基目の SNP が多く検出された結果を説明できる。

どの程度 of 同義または非同義置換が、タンパク質機能に影響を及ぼすのか明らかにするため、タンパク質コード領域における SNP に注目し、さらなる解析を進めた。まず、CDS 内で SNP

が検出されたタンパク質配列を、非同義置換が生じたキャッサバタンパク質配列と同義置換であったタンパク質配列の2つに分類した。その結果、非同義置換を生じさせた SNP は 46.8%であった(表 3-6)。この割合は、シロイヌナズナ(45.3%) (Clark et al., 2007)、トマト(46.3%) (Jimenez-Gomez and Maloof, 2009)と似ており、イネ(56.2%) (Xu et al., 2012)よりも小さかった。SNP による非同義-同義置換とタンパク質ドメインとの関係性を確かめるため、タンパク質ドメインデータベース Pfam (Punta et al., 2012)に基づき、SNP による非同義-同義置換で分類したタンパク質の対応付けを行った。ユビキチンファミリーや ATP 合成酵素を含むタンパク質では、アミノ酸が変化しない同義置換が生じる割合が多く、対照的に、NB-ARC ドメインやロイシンリッチリピートを含むタンパク質では、SNP によって非同義置換が生じたタンパク質で大きかった(図 3-4)。NB-ARC ドメインやロイシンリッチリピートといったドメインは、植物の病害応答性遺伝子に存在し、この高い非同義-同義置換比を示した結果は、病原菌感染応答性遺伝子の多様化と一致する(Bakker et al., 2006; Tameling et al., 2006; Clark et al., 2007)。

SNP によるナンセンス置換-読み過ごし置換とタンパク質機能および遺伝子重複との関係

CDS 内の SNP から 38 ヶ所のナンセンス置換(premature stop substitution; アミノ酸が終止コドンに置換)と 24 か所の読み過ごし置換(read-through substitution; 終止コドンがアミノ酸に置換)を同定し、これらの置換を検出したタンパク質配列を2群に分類した。ナンセンス置換、読み過ごし置換を検出したタンパク質を以降、それぞれナンセンス置換タンパク質、読み過ごし置換タンパク質と表現する。この2群のタンパク質配列を遺伝子オントロジーの生物学的プロセスに対応させた(図 3-5)。ナンセンス置換と読み過ごし置換タンパク質の遺伝子機能分類結果の独立性を確認するため、ピアソンのカイ二乗検定を行ったところ、p 値は 1.85×10^{-3} を示し、ナンセンス置換-読み過ごし置換タンパク質間で遺伝子機能の分類割合に有意な差があることが示唆された。続いて差がある分類項目を検定するため、調整済み標準化残差分析を行ったところ、“response to abiotic or biotic stimulus”、“response to stress”、“cell organization”の3項目について、2群間における有意差を確認した ($p < 0.05$)。ストレス応答に関して、読み過ごし置換タンパク質群で優位に割合が高かった。この結果から、ストレスへの適応などの有用な形質の獲得と

タンパク質の伸長との関連性が示された。

30,666 の遺伝子が、キャッサバゲノム概要塩基配列から予測され、配列類似性に基づき、遺伝子重複を確認したところ、その 43.3%が重複遺伝子であることが示され、ナンセンス置換タンパク質の 58.3%が重複遺伝子であることが明らかになった。フィッシャーの正確確率検定により $p < 0.05$ を示し、有意にナンセンス置換タンパク質で遺伝子重複が見られることが示された。これは、遺伝子が重複していることにより、ナンセンス置換が許容される確率が高まったためであると考えられる。一方、読み過ごし置換タンパク質の 27.8%で遺伝子重複があらわれ、全遺伝子よりも低い割合を示した。読み過ごし置換は、タンパク質の伸長を生じさせ、新規遺伝子の獲得に関与したと考えられる。

データベース **Cassava Online Archive** の構築

本研究によって 1 万を超えるキャッサバの DNA 多型が検出され、大量な関連情報が生産された。これらの解析結果をデータベース **Cassava Online Archive**(<http://cassava.psc.riken.jp/>)として構築、インターネット上に公開した。このデータベース構築によって、円滑なキャッサバゲノム情報の閲覧が可能になった。データベース **Cassava Online Archive** では、SNP や InDel の情報を閲覧するために、キーワード、DNA 多型 ID、遺伝子モデル ID、キャッサバ系統名による検索が可能であり、ゲノムブラウザも実装している(図 3-6)。各 DNA 多型の詳細ページには、キャッサバゲノムスキャフォールド上の物理的位置情報、関連する遺伝子モデル、系統毎の対立塩基を含む。同定された DNA 多型周辺のゲノム DNA 増幅のためのプライマーペア配列や想定増幅ゲノム塩基配列も詳細ページ上で提供する(図 3-7)。キャッサバゲノム、遺伝子の注釈情報、独自の遺伝子機能注釈など本研究で同定された DNA 多型情報と多岐にわたる有用情報を提供するデータベースはこれまでには存在せず、キャッサバ研究における世界標準のデータベースを構築したと考えられる。

【図表】

表 3-1 使用したキャッサバ転写産物配列

キャッサバ系統またはcDNA ライブラリ	公共データベースか ら獲得した配列数	不適切なものを除去 した配列数	コンティグ作成に使用 された配列数
AM560-2 ^a	34,151	34,151	22,589
CAS36.01	254	254	249
CAS36.04	488	488	488
CM21772	95	95	89
CM523-7	3,608	3,581	3,495
MCol22	4,764	4,764	4,604
IAC 12.829	63	63	63
KU50 (MTA116)	35,572	35,500	32,984
MBra685	2,506	2,506	2,355
MCol1522	1,979	1,975	1,854
MNga2	40	40	33
MPer183	3,391	3,388	3,206
Mirassol	210	210	208
SG107-35	720	720	651
Sauti, Gomani, Mbundumali, TME 1 and Mkondezi	5,046	5,046	4,607
TMS30572 and CM2177-2	7	7	5
Unknown	21,888	21,886	19,405
Total	114,782	114,674	96,885

* AM560-2 は、ゲノム概要塩基配列由来の予測転写産物配列。

表 3-2 検出した DNA 多型の概観

DNA 多型タイプ	SNP	InDel	合計
検出できた DNA 多型数	10,546	674	11,220
設計できたプライマーペア配列数	8,794	522	9,316
DNA 多型が検出された遺伝子モデル数	3,252	583	3,402
遺伝子モデルあたりの DNA 多型数	3.2	1.2	3.3
DNA 多型の検出頻度(エクソン領域のみ)	337.7	1,012.0	378.2
DNA 多型の検出頻度(含イントロン領域)	1,072.5	3,291.4	1,205.8

表 3-3 検出した SNP の検証

SNP ID (本研究)	SNP ID (Ferguson et al., 2012)	遺伝子モデル ID	ゲノムスキヤフォールド ID	ゲノムスキヤフォールド上の位置	アليل
R_MesP_000279m.01	Me.MEF.c.2744	cassava4.1_000279m	scaffold10222	23384	A/T
R_MesP_001567m.03	Me.MEF.c.3206	cassava4.1_001567m	scaffold01701	524265	A/T
R_MesP_001640m.02	Me.MEF.c.2663	cassava4.1_001640m	scaffold05219	124972	C/T
R_MesP_002375m.08	Me.MEF.c.2473	cassava4.1_002375m	scaffold00271	73027	C/T
R_MesP_002648m.00	Me.MEF.c.3245	cassava4.1_002648m	scaffold08265	12105	A/T
R_MesP_002747m.00	Me.MEF.c.3176	cassava4.1_002747m	scaffold11243	96145	A/T
R_MesP_003090m.05	Me.MEF.c.3339	cassava4.1_003090m	scaffold05421	111112	A/G
R_MesP_003144m.00	Me.MEF.c.2755	cassava4.1_003144m	scaffold04274	213187	A/G
R_MesP_003321m.01	Me.MEF.c.1794	cassava4.1_003321m	scaffold02431	174176	A/C
R_MesP_003427m.01	Me.MEF.c.2363	cassava4.1_003427m	scaffold09294	205067	A/G
R_MesP_004227m.01	Me.MEF.c.3056	cassava4.1_004227m	scaffold12585	11918	A/G
R_MesP_004227m.02	Me.MEF.c.3057	cassava4.1_004227m	scaffold12585	12038	A/G
R_MesP_004630m.01	Me.MEF.c.3345	cassava4.1_004630m	scaffold03264	203508	A/G
R_MesP_004975m.00	Me.MEF.c.3066	cassava4.1_004975m	scaffold08265	4313944	A/G
R_MesP_005040m.12	Me.MEF.c.3336	cassava4.1_005040m	scaffold04285	281079	A/G
R_MesP_006632m.00	Me.MEF.c.3217	cassava4.1_006632m	scaffold04043	804524	A/G
R_MesP_007404m.01	Me.MEF.c.2570	cassava4.1_007404m	scaffold11495	796742	A/G
R_MesP_007552m.00	Me.MEF.c.3199	cassava4.1_007552m	scaffold03049	867704	A/G
R_MesP_007560m.00	Me.MEF.c.1958	cassava4.1_007560m	scaffold06582	904165	C/G
R_MesP_008478m.01	Me.MEF.c.3361	cassava4.1_008478m	scaffold04457	1003339	C/T
R_MesP_008791m.00	Me.MEF.c.2962	cassava4.1_008791m	scaffold10305	629767	A/G

R_MesP_008822m.00	Me.MEF.c.1127	cassava4.1_008822m	scaffold02717	237099	G/T
R_MesP_008965m.01	Me.MEF.c.2384	cassava4.1_008965m	scaffold12262	132277	A/T
R_MesP_009150m.04	Me.MEF.c.1459	cassava4.1_009150m	scaffold08265	2078665	A/C
R_MesP_009532m.02	Me.MEF.c.1982	cassava4.1_009532m	scaffold08380	34042	G/T
R_MesP_009764m.00	Me.MEF.c.0774	cassava4.1_009764m	scaffold07238	149170	C/G
R_MesP_010236m.02	Me.MEF.c.1049	cassava4.1_010236m	scaffold04587	95636	A/C
R_MesP_010383m.02	Me.MEF.c.1268	cassava4.1_010383m	scaffold00847	2075370	C/T
R_MesP_010832m.02	Me.MEF.c.1071	cassava4.1_010832m	scaffold02870	5127	C/T
R_MesP_010884m.03	Me.MEF.c.2344	cassava4.1_010884m	scaffold12088	25866	G/T
R_MesP_010884m.06	Me.MEF.c.2346	cassava4.1_010884m	scaffold12088	30223	C/G
R_MesP_011879m.05	Me.MEF.c.2979	cassava4.1_011879m	scaffold06656	794006	C/G
R_MesP_011947m.03	Me.MEF.c.2248	cassava4.1_011947m	scaffold01551	2841586	C/T
R_MesP_012068m.01	Me.MEF.c.2001	cassava4.1_012068m	scaffold08265	61457	C/G
R_MesP_012375m.00	Me.MEF.c.2402	cassava4.1_012375m	scaffold08265	1858135	A/C
R_MesP_012582m.01	Me.MEF.c.2454	cassava4.1_012582m	scaffold08085	36479	A/G
R_MesP_012703m.01	Me.MEF.c.2119	cassava4.1_012703m	scaffold12498	309545	A/T
R_MesP_012776m.01	Me.MEF.c.3152	cassava4.1_012776m	scaffold00321	179660	A/G
R_MesP_012882m.01	Me.MEF.c.0670	cassava4.1_012882m	scaffold11110	92729	A/C
R_MesP_012882m.02	Me.MEF.c.0671	cassava4.1_012882m	scaffold11110	92855	A/C
R_MesP_013010m.00	Me.MEF.c.0996	cassava4.1_013010m	scaffold07233	29151	A/T
R_MesP_013452m.00	Me.MEF.c.2177	cassava4.1_013452m	scaffold02431	443087	A/C
R_MesP_013631m.00	Me.MEF.c.0744	cassava4.1_013631m	scaffold07478	34782	C/T
R_MesP_013795m.02	Me.MEF.c.2927	cassava4.1_013795m	scaffold09294	87456	A/T
R_MesP_013999m.00	Me.MEF.c.1284	cassava4.1_013999m	scaffold03237	157721	C/G
R_MesP_014216m.00	Me.MEF.c.0963	cassava4.1_014216m	scaffold11960	73682	A/G
R_MesP_014405m.06	Me.MEF.c.0240	cassava4.1_014405m	scaffold06598	425721	A/G

R_MesP_014693m.00	Me.MEF.c.2428	cassava4.1_014693m	scaffold06407	370274	A/G
R_MesP_015054m.01	Me.MEF.c.1671	cassava4.1_015054m	scaffold12794	25269	A/G
R_MesP_015068m.03	Me.MEF.c.2698	cassava4.1_015068m	scaffold05363	196220	A/T
R_MesP_015093m.01	Me.MEF.c.3044	cassava4.1_015093m	scaffold12455	220233	C/T
R_MesP_015119m.02	Me.MEF.c.0458	cassava4.1_015119m	scaffold10504	17877	C/T
R_MesP_015358m.01	Me.MEF.c.0262	cassava4.1_015358m	scaffold06407	113040	C/T
R_MesP_015358m.02	Me.MEF.c.0263	cassava4.1_015358m	scaffold06407	113448	A/T
R_MesP_015786m.00	Me.MEF.c.0114	cassava4.1_015786m	scaffold11661	293210	C/T
R_MesP_015850m.01	Me.MEF.c.1874	cassava4.1_015850m	scaffold06278	142397	G/T
R_MesP_015958m.00	Me.MEF.c.3229	cassava4.1_015958m	scaffold10045	438228	A/G
R_MesP_015972m.01	Me.MEF.c.3209	cassava4.1_015972m	scaffold05859	584798	C/T
R_MesP_016045m.00	Me.MEF.c.2928	cassava4.1_016045m	scaffold09631	95044	A/T
R_MesP_016262m.01	Me.MEF.c.2655	cassava4.1_016262m	scaffold05875	845674	G/T
R_MesP_016312m.00	Me.MEF.c.2236	cassava4.1_016312m	scaffold06916	1353707	A/T
R_MesP_016339m.06	Me.MEF.c.1052	cassava4.1_016339m	scaffold05875	1512655	A/T
R_MesP_016344m.05	Me.MEF.c.2554	cassava4.1_016344m	scaffold01514	54209	C/G
R_MesP_016364m.01	Me.MEF.c.0795	cassava4.1_016364m	scaffold02421	475471	A/G
R_MesP_016411m.00	Me.MEF.c.3310	cassava4.1_016411m	scaffold01551	1608759	C/T
R_MesP_016484m.01	Me.MEF.c.3379	cassava4.1_016484m	scaffold08799	183769	G/T
R_MesP_016917m.11	Me.MEF.c.1658	cassava4.1_016917m	scaffold07520	1722908	A/C
R_MesP_016917m.16	Me.MEF.c.1664	cassava4.1_016917m	scaffold07520	1724876	A/T
R_MesP_017270m.00	Me.MEF.c.0570	cassava4.1_017270m	scaffold00847	1732669	A/G
R_MesP_017713m.00	Me.MEF.c.0859	cassava4.1_017713m	scaffold02264	132995	C/T
R_MesP_017738m.00	Me.MEF.c.2873	cassava4.1_017738m	scaffold11110	232152	C/G
R_MesP_017755m.04	Me.MEF.c.3002	cassava4.1_017755m	scaffold12946	396504	C/T
R_MesP_018048m.00	Me.MEF.c.1986	cassava4.1_018048m	scaffold07027	61508	C/T

R_MesP_018091m.00	Me.MEF.c.2562	cassava4.1_018091m	scaffold03735	83703	A/G
R_MesP_018095m.00	Me.MEF.c.0381	cassava4.1_018095m	scaffold06005	225219	C/T
R_MesP_018102m.01	Me.MEF.c.1279	cassava4.1_018102m	scaffold03168	55619	A/T
R_MesP_018355m.00	Me.MEF.c.2297	cassava4.1_018355m	scaffold05859	590922	A/G
R_MesP_018401m.03	Me.MEF.c.1597	cassava4.1_018401m	scaffold03150	159657	C/T
R_MesP_018439m.00	Me.MEF.c.2128	cassava4.1_018439m	scaffold03614	493353	C/G
R_MesP_018524m.00	Me.MEF.c.2747	cassava4.1_018524m	scaffold12794	504065	A/G
R_MesP_018548m.04	Me.MEF.c.0767	cassava4.1_018548m	scaffold09458	13818	A/C
R_MesP_018555m.02	Me.MEF.c.1617	cassava4.1_018555m	scaffold11279	197763	C/T
R_MesP_018724m.00	Me.MEF.c.1393	cassava4.1_018724m	scaffold01551	2804872	G/T
R_MesP_018755m.00	Me.MEF.c.2448	cassava4.1_018755m	scaffold12742	63374	G/T
R_MesP_018881m.00	Me.MEF.c.0768	cassava4.1_018881m	scaffold02477	376181	C/T
R_MesP_018897m.02	Me.MEF.c.0951	cassava4.1_018897m	scaffold05703	60839	G/T
R_MesP_018900m.00	Me.MEF.c.1382	cassava4.1_018900m	scaffold10222	26834	C/T
R_MesP_018952m.00	Me.MEF.c.0829	cassava4.1_018952m	scaffold10524	7426	C/T
R_MesP_018955m.03	Me.MEF.c.2759	cassava4.1_018955m	scaffold02421	61392	A/C
R_MesP_019054m.00	Me.MEF.c.1454	cassava4.1_019054m	scaffold00847	305160	A/G
R_MesP_019054m.01	Me.MEF.c.1455	cassava4.1_019054m	scaffold00847	305643	A/T
R_MesP_019293m.04	Me.MEF.c.2124	cassava4.1_019293m	scaffold07238	193443	A/G
R_MesP_019641m.02	Me.MEF.c.2953	cassava4.1_019641m	scaffold09151	975576	A/T
R_MesP_020084m.00	Me.MEF.c.3324	cassava4.1_020084m	scaffold01796	17779	A/G
R_MesP_020289m.00	Me.MEF.c.1729	cassava4.1_020289m	scaffold04538	100541	C/T
R_MesP_020722m.01	Me.MEF.c.2714	cassava4.1_020722m	scaffold10173	959438	C/G
R_MesP_020733m.00	Me.MEF.c.3139	cassava4.1_020733m	scaffold04209	542273	C/T
R_MesP_024632m.00	Me.MEF.c.0677	cassava4.1_024632m	scaffold03237	297465	C/G
R_MesP_026456m.00	Me.MEF.c.2401	cassava4.1_026456m	scaffold06158	303583	A/G

R_MesP_027198m.00	Me.MEF.c.1902	cassava4.1_027198m	scaffold06512	1446290	A/T
R_MesP_027526m.14	Me.MEF.c.0724	cassava4.1_027526m	scaffold11837	111409	C/T
R_MesP_032608m.00	Me.MEF.c.1382	cassava4.1_032608m	scaffold01645	30975	C/T
R_MesP_034057m.00	Me.MEF.c.1235	cassava4.1_034057m	scaffold08265	4464846	A/G

本研究で同定した SNP の内 103 個は、Ferguson ら(2012)が検証を行った SNP と位置情報が一致した。そのすべての SNP のアレルに齟齬は生じなかった。したがって、この結果は本研究の同定方法が有効であることを示した。

表 3-4 トランジション置換とトランスバージョン変異の内訳

トランジション置換	
C/T	2,893
G/A	2,952
合計	5,845
トランスバージョン置換	
A/C	1,108
A/T	1,285
C/G	996
G/T	1,312
合計	4,701

トランジション置換(C/T または G/A)とトランスバージョン置換(C/G、A/T、C/A または T/G) は、それぞれ 5,845 と 4,701 ヶ所であった。トランジション置換ートランスバージョン置換比は、1.24 であった。

表 3-5 2 系統間における対立 SNP 遺伝子比

	AM560-2	CM523-7	MCol22	KU50	MBra685	MCol1522	MPer183
CM523-7	0.67						
MCol22	0.85	0.74					
KU50	0.82	0.54	0.66				
MBra685	0.40	0.52	0.67	0.58			
MCol1522	0.40	0.50	0.65	0.59	0.41		
MPer183	0.73	0.61	0.64	0.65	0.61	0.39	
SG107-35	0.72	0.50	0.40	0.50	0.50	0.45	0.71

500 以上の EST が獲得できた系統のみを対象とした。SNP が検出された遺伝子に対して、各々 2 系統間において対立 SNP が検出された遺伝子との比(該当 2 系統間で対立 SNP が存在した遺伝子数/SNP 検出に該当 2 系統の EST が関与した遺伝子数)を確認した。

表 3-6 ゲノム注釈情報に基づく SNP 位置の分類

	SNP		InDel
CDS			
(非同義/同義 置換)	6,613	(3,095/3,518)	123
5' UTR		1,466	188
3' UTR		2,467	363
合計		10,546	674

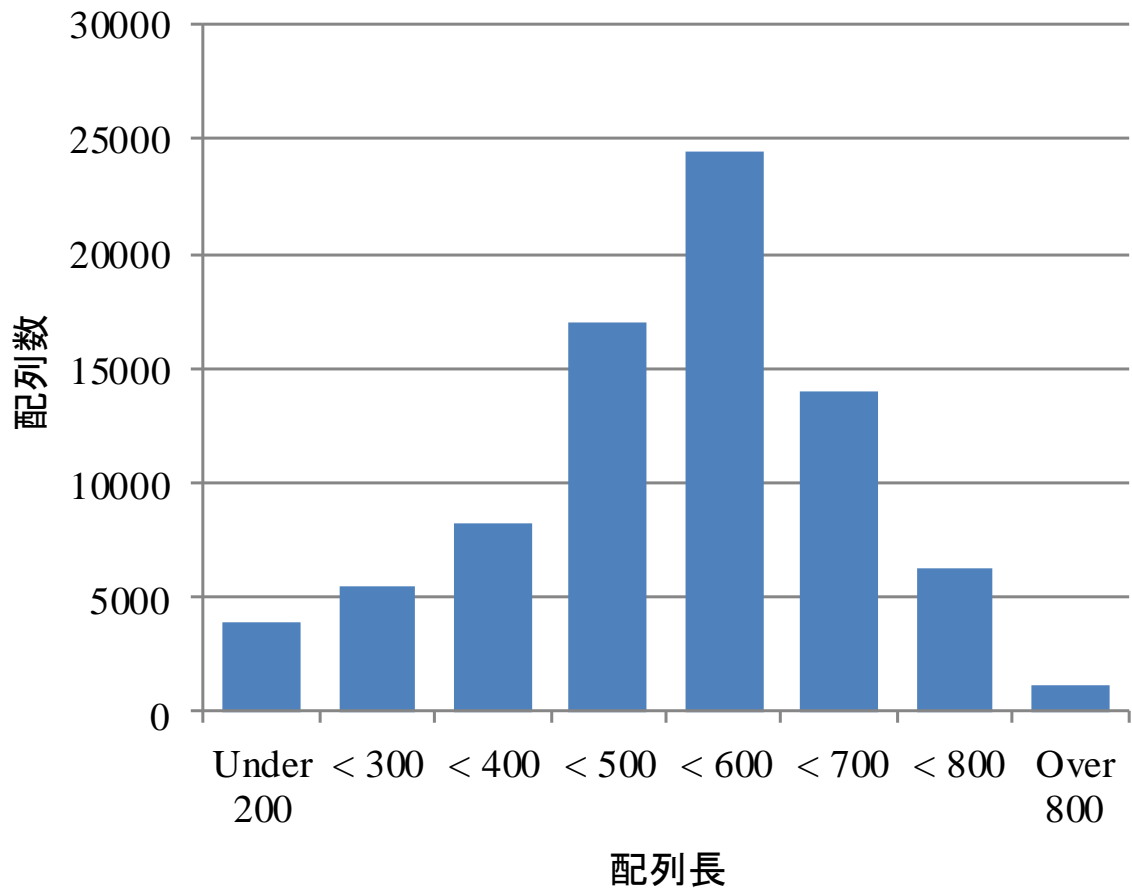


図 3-1 獲得した EST の配列長の分布

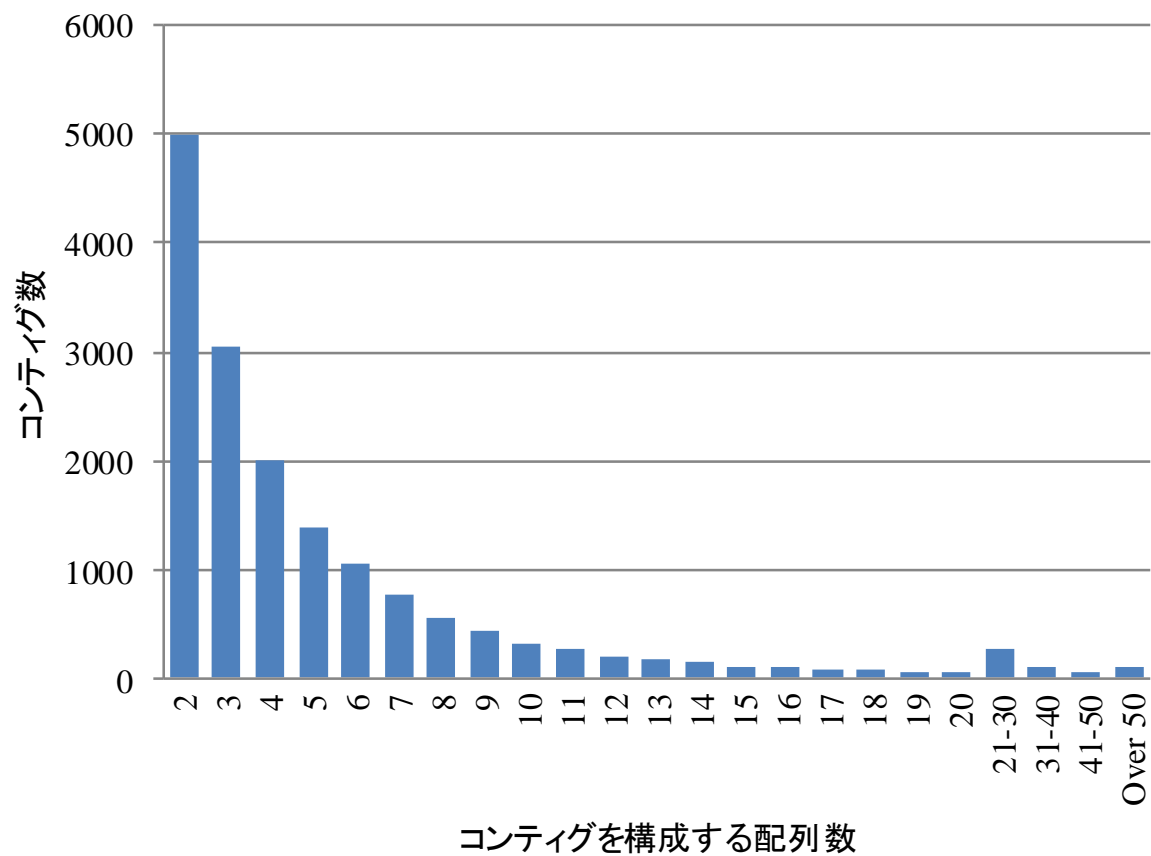


図 3-2 コンティグを構成する EST の配列数の分布

コンティグを構成する配列数の分布は 2 から 707 であり、コンティグを構成する平均配列数は 5.9 であった。

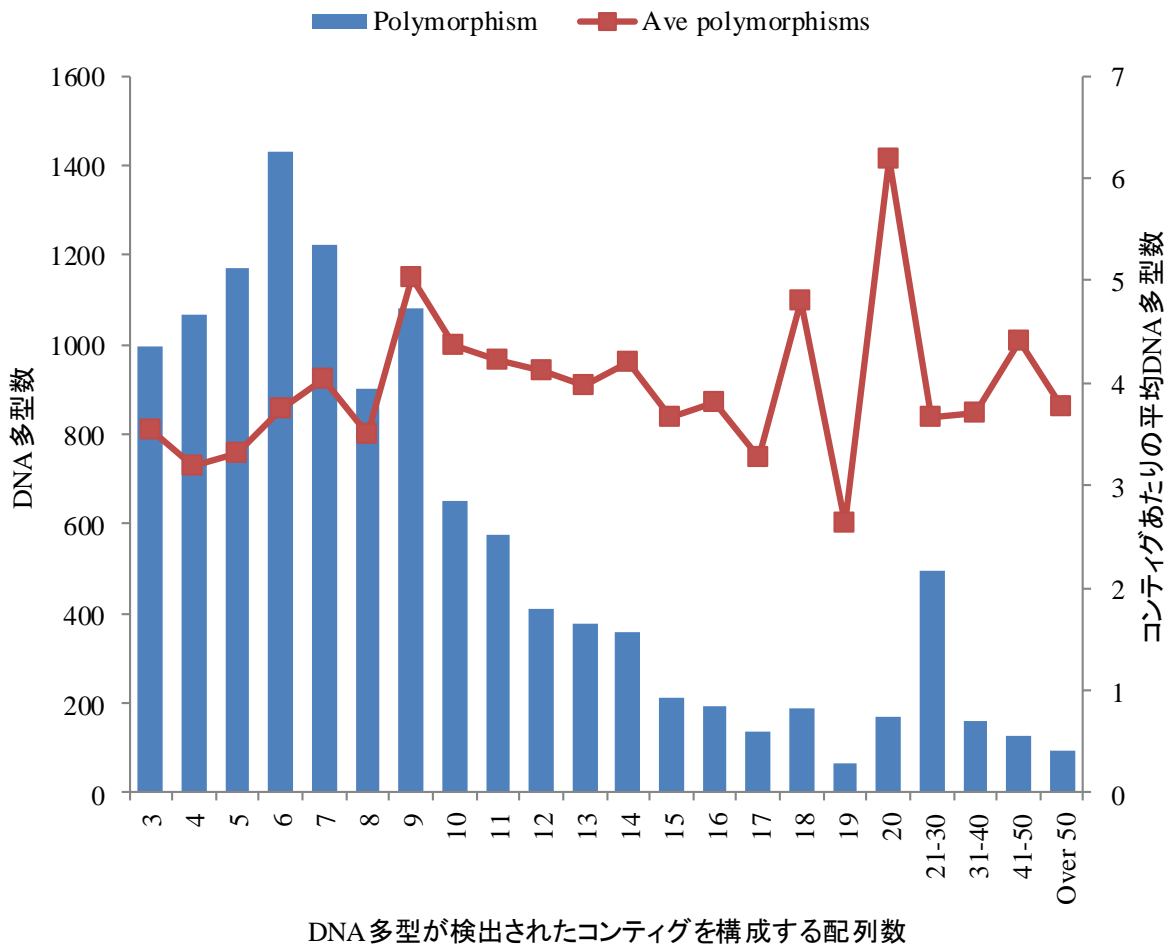


図 3-3 コンティグを構成する EST の配列数と検出された DNA 多型数の分布

コンティグあたりの平均 DNA 多型検出数は 3.8 で、2.6 から 6.2 の範囲にあった。この結果は、コンティグを構成する配列数と同定される DNA 多型数に特に偏りがなかったことを示し、解析に用いた配列のシーケンスエラーによる誤った DNA 多型の同定ではないことを示す。

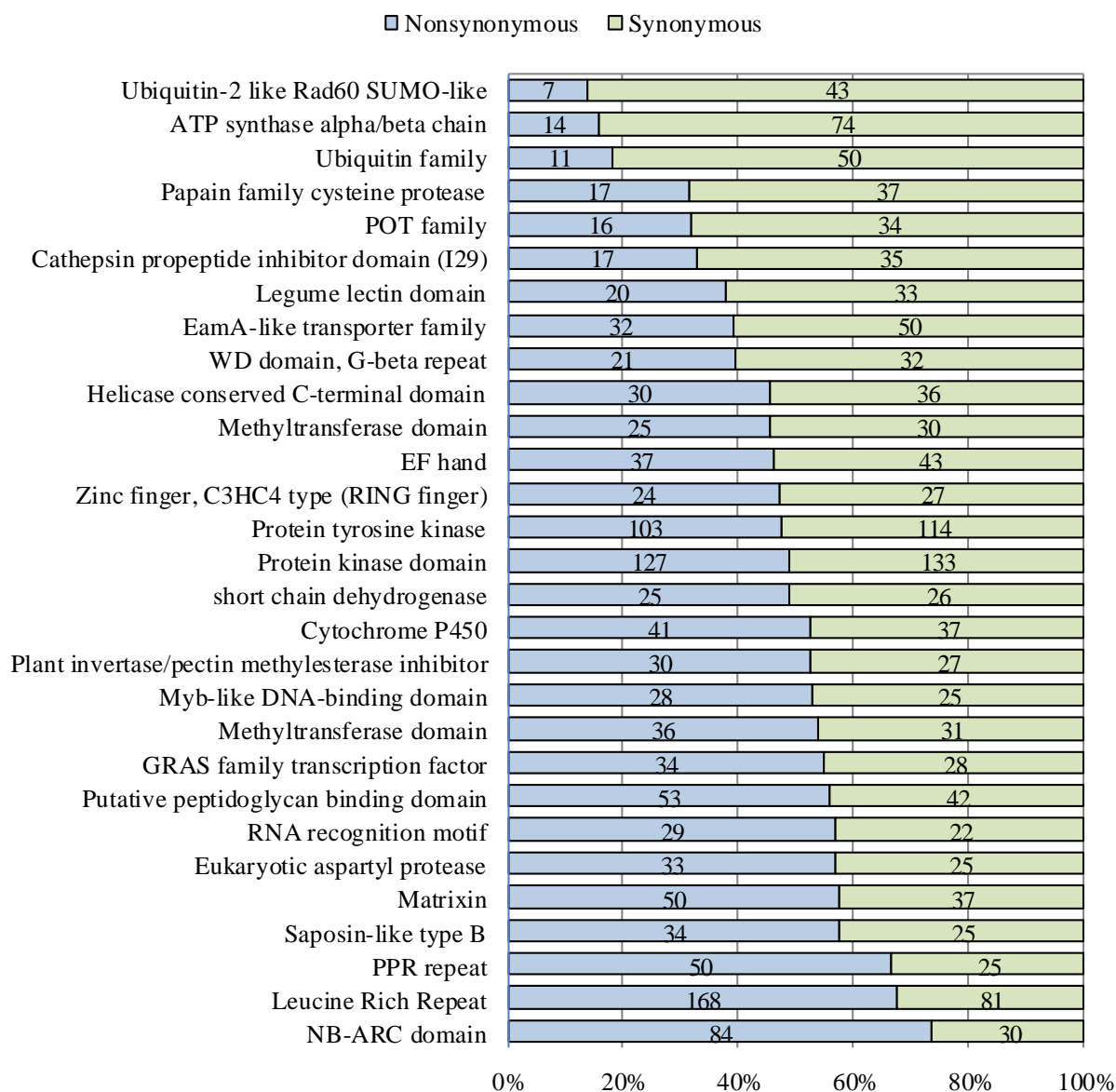


図 3-4 非同義および同義塩基置換 SNP の概観

1,196 の非同義塩基置換タンパク質と 1,232 同義塩基置換タンパク質から、30 以上の SNP が対応した Pfam ドメインについて表した。ユビキチンファミリーや ATP 合成酵素などは、低い非同義-同義塩基置換比を示した。一方、NB-ARC ドメインやロイシンリッチリピートなどは、高い非同義-同義塩基置換比を示した。

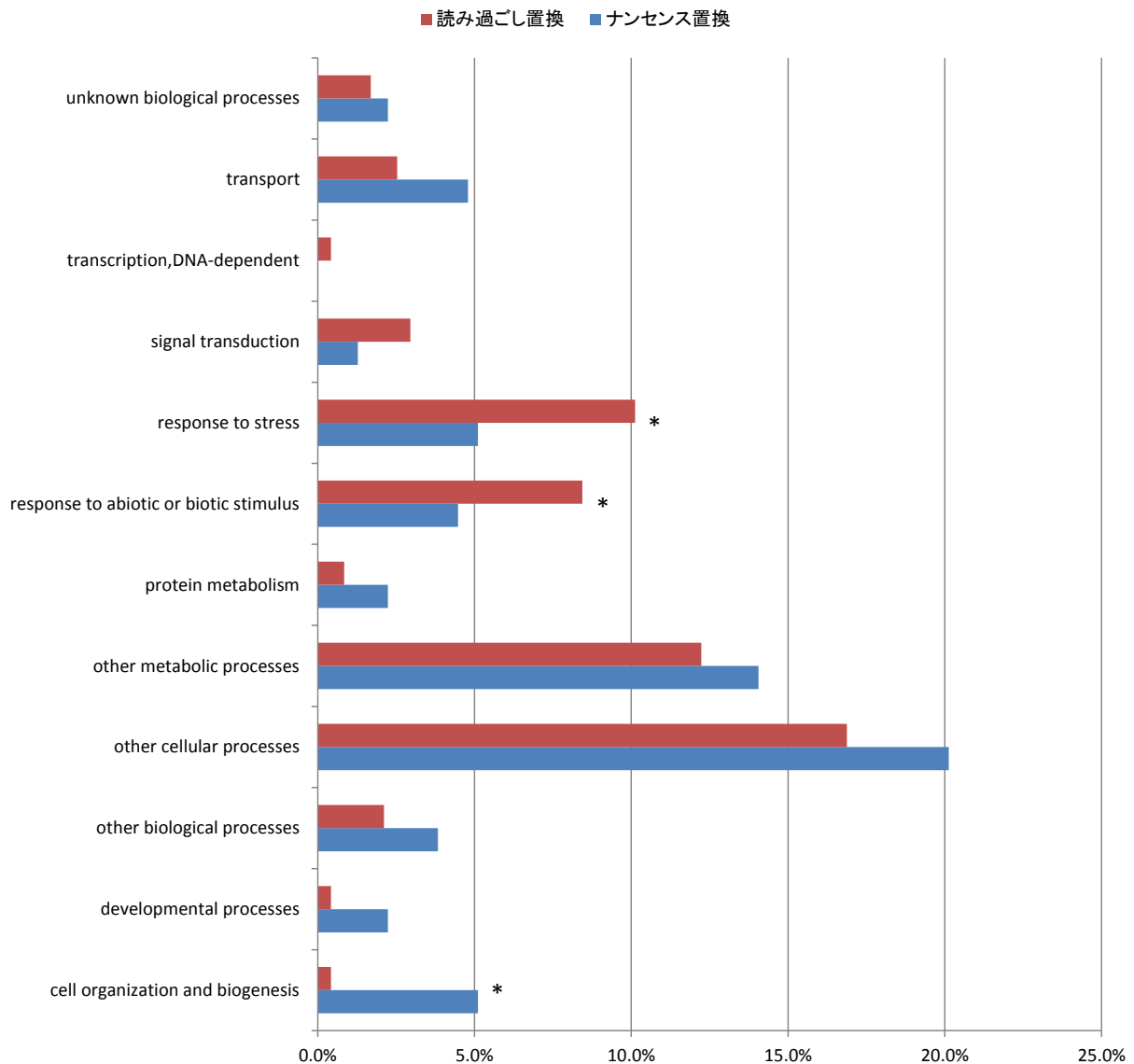


図 3-5 ナンセンス置換と読み過ごし置換タンパク質の遺伝子オントロジ生物学的プロセス (biological process)への対応割合の比較

ピアソンのカイ二乗検定を行ったところ、 p 値は 1.85×10^{-3} を示し、ナンセンス置換—読み過ごし置換タンパク質間で遺伝子機能の分類割合に有意な差があることが示唆された。続いて差がある分類項目を検定するため、調整済み標準化残差分析を行ったところ、“response to abiotic or biotic stimulus”、“response to stress”、“cell organization”の3項目について、2群間における有意差を確認した (* $p < 0.05$)。

Home > Polymorphism

Search About

Keyword:

Variety:

- AM560-2
- CAS36.01
- CAS36.04
- CM21772
- CM523-7
- MCo22
- IAC 12.829
- KU50
- MBra685
- MCo11522
- MNg2
- MPer183
- Mirasol
- SG107-35
- Sauti, Gomani, Mbundumali, TME 1 and Mtondezi
- TMS30572 and CM2177-2
- Unknown

Search Clear

e.g., "map kinase cassava4_1_000716m"

Search result 14 hits

Page: 1 - 14 displayed

Polymorphism ID	Scaffold	Position	Related gene model	Type	Description	Primer info.	Polymorphism																				
R_MesP_001712m.01	scaffold03614	1736445	cassava4_1_001712m	SNP	best arabidopsis TAR10 hit name: AT5G17620.2 Cobalamin-independent synthase family protein	Yes	<table border="1"> <thead> <tr><th colspan="2">Allele</th></tr> <tr><th>Allele</th><th>G/A</th></tr> </thead> <tbody> <tr><td>AM560-2</td><td>G</td></tr> <tr><td>KU50</td><td>G</td></tr> <tr><td>MBra685</td><td>A</td></tr> <tr><td>MCo11522</td><td>G</td></tr> <tr><td>MPer183</td><td>A</td></tr> <tr><td>Unknown</td><td>A</td></tr> </tbody> </table>	Allele		Allele	G/A	AM560-2	G	KU50	G	MBra685	A	MCo11522	G	MPer183	A	Unknown	A				
Allele																											
Allele	G/A																										
AM560-2	G																										
KU50	G																										
MBra685	A																										
MCo11522	G																										
MPer183	A																										
Unknown	A																										
R_MesP_001760m.00	scaffold01551	1168404	cassava4_1_001760m	SNP	best arabidopsis TAR10 hit name: AT3G22890.1 ATP sulfurylase 1	Yes	<table border="1"> <thead> <tr><th colspan="2">Allele</th></tr> <tr><th>Allele</th><th>T/C</th></tr> </thead> <tbody> <tr><td>AM560-2</td><td>T</td></tr> <tr><td>KU50</td><td>T</td></tr> <tr><td>MBra685</td><td>T</td></tr> <tr><td>MCo11522</td><td>T</td></tr> <tr><td>MCo22</td><td>C</td></tr> <tr><td>MPer183</td><td>T</td></tr> <tr><td>Sauti, Gomani, Mbundumali, TME 1 and Mtondezi</td><td>T</td></tr> <tr><td>Unknown</td><td>T</td></tr> </tbody> </table>	Allele		Allele	T/C	AM560-2	T	KU50	T	MBra685	T	MCo11522	T	MCo22	C	MPer183	T	Sauti, Gomani, Mbundumali, TME 1 and Mtondezi	T	Unknown	T
Allele																											
Allele	T/C																										
AM560-2	T																										
KU50	T																										
MBra685	T																										
MCo11522	T																										
MCo22	C																										
MPer183	T																										
Sauti, Gomani, Mbundumali, TME 1 and Mtondezi	T																										
Unknown	T																										

File Help

Manihot esculenta: 20 kbp from scaffold08265:3,871,493..3,891,492

Browser Select Tracks Snapshots Custom Tracks Preferences

Search

Landmark or Region:

Examples: cassava4_1_017479m.g, scaffold00224.275,001..325,000, scaffold04796.35..19000, AT3G65750.1.

Data Source: Manihot esculenta

Scroll/Zoom: Show 20 kbp Flip

Overview

Region

Details

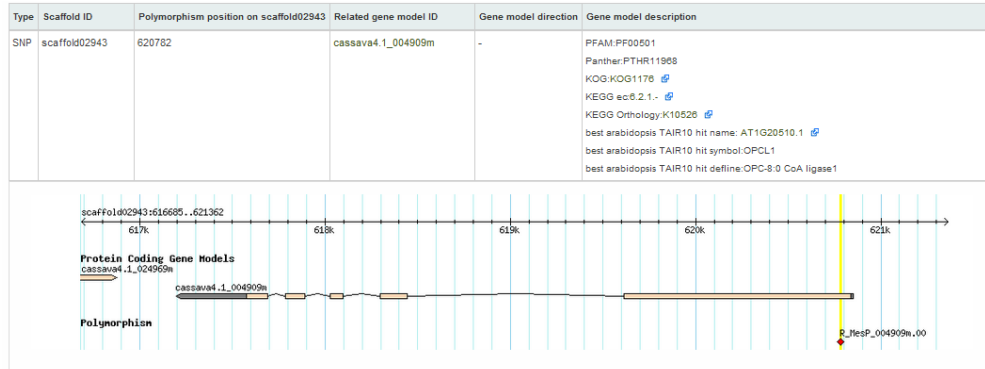
Polymorphism

Select Tracks Clear highlighting

図 3-6 データベース Cassava Online Archive の検索機能

(A)利用者は、DNA 多型を多型 ID だけでなく、キーワード、遺伝子モデル、キャッサバ系統から検索可能である。(B)検索結果画面。(C)ゲノムブラウザにより、視覚的に DNA 多型を確認することが可能である。

A



B

Left Primer sequence	CAGCTTCGAGATCGACCCA
Left Primer start position	620821
Left Primer sequence length	19
Left Primer TM (DegC)	62.503
Left Primer %GC	57.895
Right Primer sequence	CGTCGATGAAGCCAGTCTTG
Right Primer start position	620667
Right Primer sequence length	20
Right Primer TM (DegC)	61.935
Right Primer %GC	55.000
Amplicon	CAGCTTCGAGATCGACCCAAGAGCCGCTTTGCAGATCAAAATCTATTTTCTACAGCAAAACCCAAATTCCTCTCCACCCAACTCACTCAATTGATATCACCACCTTTTATTCTTCCCAAGCTCACCGTGGCAAGACTGCCTTCATCGAGC
Amplicon length	155

C

Allele	A/C
AM560-2 (Predicted CDS from genome sequence)	A
CAS36.01	-
CAS36.04	-
CM21772	-
CM523-7	-
MCol22	-
IAC 12.829	-
KU50	C
MBra685	A
MCol1522	-
MNga2	-
MPer183	-
Mirassol	-
SG107-35	-
Sauti, Goman, Mbundumali, TME 1 and Mkondezi	-
TMS30572 and CM2177-2	-
Unknown	-

D

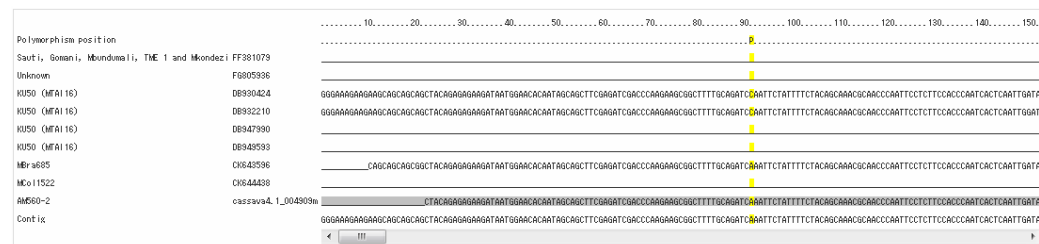


図 3-7 DNA 多型情報の詳細画面

(A)物理的位置情報。(B)プライマーペア配列情報。(C)系統毎の対立塩基。(D)DNA 多型検出における整列化配列。

第4章 結論

(1) シロイヌナズナのゲノム解析とゲノム統合データベース RARGE の構築

シロイヌナズナのゲノム注釈情報は、大規模 cDNA 収集プロジェクトによる豊かな配列データから転写領域、遺伝子構造の解析が行われ、他の植物よりも高品質なゲノム注釈情報が提供されてきた。本研究では、完全長 cDNA の配列情報を活用して、新規の転写領域および遺伝子モデルを同定し、さらなるゲノム注釈情報の改善を図った。また、本解析に使用した完全長 cDNA の配列データに遺伝子機能などの注釈情報を付加した。遺伝子機能同定のための研究基盤として開発された *Ds* トランスポゾン挿入遺伝子破壊系統の情報についても、各々のタグ挿入変異体のゲノム上のタグ挿入位置を同定し、タグ挿入位置周辺の遺伝子や遺伝子構造上のタグ挿入位置の分類を行った。この解析により、タグ挿入変異体が関与する遺伝子の情報が整備された。シロイヌナズナの機能ゲノム研究のさらなる推進を目的として、上記の解析により生産した情報をゲノム統合データベース RARGE(<http://rarge.psc.riken.jp/>)として編纂し、インターネット上に公開した。

シロイヌナズナのゲノム注釈情報の改善、新規転写領域の同定

シロイヌナズナのゲノム研究成果は、主にシロイヌナズナ情報資源データベース(The Arabidopsis Information Resource; TAIR <http://arabidopsis.org/>) (Lamesch et al., 2012)に統合され、インターネットを通じて閲覧できるオンラインデータベースとして編纂されている。2000年12月に解読されたシロイヌナズナの全ゲノム塩基配列データは、ゲノム配列決定が一般的になった今日においても高い品質といわれている。このゲノム塩基配列データに加え、シロイヌナズナのcDNA配列データも、2012年9月現在でも200万配列に達しようとしており、これらcDNA配列データを使用したゲノム注釈情報は、植物ゲノムの内で最高水準といえる。しかし、転写領域、遺伝子機能および研究リソースについては改善の余地があるため、本研究を行った。

使用する配列データの品質を考慮し、理化学研究所が作製したシロイヌナズナ完全長 cDNA 由来のデータのみを扱うこととし、公共データバンク GenBank から該当のシロイヌナズナ完全長

cDNA の末端読み(EST)および全長読み配列データを獲得した。この配列データを使用した解析の結果、完全長 cDNA 配列データの 98.7%がマップされた。既知のシロイヌナズナゲノム注釈情報と対応しなかった 1,502 配列を対象にゲノム上の位置情報に基づくクラスタリングを行ったところ、503 クラスタを得た。この 503 クラスタは、新規シロイヌナズナ遺伝子の候補とみなせる。また、ゲノム注釈情報と対応した配列についても精査したところ、4,119 のシロイヌナズナ遺伝子が既存のゲノム注釈による転写領域を 500 塩基以上延長できる新規な遺伝子モデルを持つ可能性を示した。以上のように、シロイヌナズナのゲノム注釈情報を改善した。既知遺伝子モデルを含むゲノム注釈情報へ cDNA リソースを対応させることで、バイオリソースを必要とする関連研究の推進にも貢献すると思われる。

シロイヌナズナ *Ds* トランスポゾンタグ挿入変異体のタグ近傍ゲノム DNA 配列データの獲得と物理的位置の把握

Ds トランスポゾンや T-DNA の挿入による遺伝子破壊系統の開発は、遺伝子機能同定のための効果的なアプローチの 1 つであり、シロイヌナズナの機能ゲノム研究において、cDNA 収集と同様に大規模な変異体系統の開発に基づく遺伝子機能解析プロジェクトが推進され、多くのタグ挿入変異体が作出された。しかし、各変異体におけるゲノム上のタグ挿入位置など注釈情報の整備は限定的であるため、これらタグ挿入変異体に関する情報整備を行った。理化学研究所の研究チームが作製したシロイヌナズナ *Ds* トランスポゾンタグ挿入変異体(Ito et al., 2002; Kuromori et al., 2004; Kuromori et al., 2006)の 17,671 系統について、挿入された *Ds* トランスポゾンタグの近傍ゲノム塩基配列を獲得し、この配列データをゲノム塩基配列へマップすることで、ゲノム上における *Ds* トランスポゾンタグ挿入位置を計算した。さらに、推算できた各々の *Ds* トランスポゾン挿入位置情報に基づき、既知の遺伝子注釈情報(AGI gene)へ対応させ、プロモータ領域、タンパク質コード領域、5'側非翻訳領域、3'側非翻訳領域、遺伝子間領域に分類した。大量な *Ds* トランスポゾンタグ挿入変異体について、タグ挿入位置の推算、関連遺伝子の同定などの体系的な注釈付けが行われたことはこれまでにはほとんどなかったが、本研究によりタグ挿入位置による変異体探索の効率化が図られ、変異体についての円滑な情報参照が可能になった。

シロイヌナズナゲノム情報データベース RARGE の構築

上述の解析で生産した情報は、データベース RARGE として統合し、インターネット上に公開した(<http://rarge.psc.riken.jp/>)。各完全長 cDNA と Ds トランスポゾンタグ挿入変異体の詳細情報ページを作成するだけでなく、膨大な収録データの円滑な閲覧を可能にするため、各種検索機能を実装した。完全長 cDNA に特化した配列データとその注釈情報が整備された情報資源は他には見られないため、転写開始点に基づく数々の RNA 転写関連の研究で活用された(Yamamoto et al., 2007; Schindler et al., 2008; Ramirez and Basu, 2009; Iida et al., 2011; Shahriari et al., 2011; Gruber et al., 2012; Herberth et al., 2012; Maruyama et al., 2012)。また、特定の遺伝子ファミリーの同定などの解析では高品質かつ均一な配列情報が有効であり、これらの解析においても本データベースが活用された(Iida et al., 2005)。Ds トランスポゾンタグ挿入変異体については、キーワード、変異体名、シロイヌナズナ遺伝子名(AGI 遺伝子 ID)による検索機能を実装し、検索結果から該当する Ds トランスポゾンタグ挿入変異体の詳細情報ページに遷移することができる。シロイヌナズナ情報資源データベース TAIR や SALK 研究所のシロイヌナズナデータベース T-DNA Express (<http://signal.salk.edu/cgi-bin/tdnaexpress>) でもタグ挿入変異体の情報を参照できるが、上述のようなタグ挿入関連遺伝子による検索機能を有しておらず、本データベースが利便性に優れる面を持つ。

本研究と同じように理化学研究所が提供する完全長 cDNA 配列情報を用いて行われたシロイヌナズナ全ゲノム選択的スプライシング解析(Iida et al., 2004)の結果を共通する完全長 cDNA 名で統合し、遺伝子構造を含むシロイヌナズナのゲノム情報を実装した。Ds トランスポゾンタグ挿入変異体を用いた表現形質解析(Kuromori et al., 2006)、核コードの葉緑体関連遺伝子解析(Myouga et al., 2010)の結果を収録したデータベース RAPID (<http://rarge.psc.riken.jp/phenome/>) と葉緑体タンパク質データベース (<http://rarge.psc.riken.jp/chloroplast/>)、さらにシロイヌナズナ転写因子データベース RARTF (<http://rarge.psc.riken.jp/rartf/>) へのリンクを備え、関連する有用な情報についての閲覧性を向上させた。また、実験リソースの請求に関する利便性も考慮し、本データベース RARGE に収録している生物資源の提供を行っている理化学研究所バイオリソース

センターのリソースカタログページ

(<http://www.brc.riken.jp/lab/epd/Eng/species/arabidopsis.shtml>) へのリンクも備えた。この点

に関して、数々の論文が本データベースを引用しており、本データベースによる研究効率の促進が評価された結果と思われる。

以上のように、完全長 cDNA 配列データを用いた転写開始位置、新規転写領域の同定、*Ds* トランスポゾンタグ挿入変異体のタグ挿入位置の推算とタグ挿入位置と転写領域との関係の獲得を行うことで、ゲノム注釈情報を改善し、シロイヌナズナのゲノム研究の推進に貢献できたと考えられる。今後は、タグ挿入変異体や形質転換体の表現形質情報の整備、編纂を行うことで、シロイヌナズナ遺伝子機能の解析のさらなる推進に貢献したい。

(2) 有用植物の完全長 cDNA 収集と草本植物シロイヌナズナとの比較解析

モデル生物は、分子生物学とその周辺の研究分野において、普遍的な生命現象の研究に用いられる生物のことである。シロイヌナズナは、草丈が約 20cm 程度と小さく、生活環も 2 か月程度である上に簡便な形質転換技術が確立していることなどから、植物研究において最も幅広く使用されている生物種の 1 つであり、ゲノム研究においても顕花植物として初めて全ゲノム塩基配列が解読され、モデル植物として盛んに研究が行われてきた。シロイヌナズナの研究成果を他の植物研究に活用し、遺伝子領域、機能の比較などの対象植物種のゲノム研究の推進がより加速されることが期待できるようになった。そこで本研究では、樹木ポプラと熱帯作物キャッサバの完全長 cDNA を収集、解析することで各々の植物種の遺伝子概観を獲得するとともに、シロイヌナズナとの比較解析を行うことで、各植物種の分子生物学的特性の解明を試みた。

完全長 cDNA クローンの収集と両末端読み配列の決定および配列アセンブリ

cDNA 配列は、ゲノム研究推進に重要な情報資源であるが、ポプラの EST は 30 万、キャッサバにいたっては 3 万に過ぎず、他の有用植物トウモロコシの 300 万配列データやイネの 200 万配列データなどに比べ、あまりにも小規模といえる。また、2012 年に公開されたキャッサバゲノム概要塩基配列は、想定されるゲノムサイズの 60%程度の解読であり(Prochnik et al., 2012)、EST の質、量ともに不十分であるため、不正確な転写領域定義を多く含むことが考えられる。ポプラのゲノム注釈情報についてもそれが当てはまる。そこで本研究では、高効率に全長 RNA を収集することができるビオチン化キャップトラッパー法(Carninci et al., 2000)を用いて、ポプラおよびキャッサバの完全長 cDNA ライブラリを作製し、収集した完全長 cDNA クローンの両末端からの塩基配列決定を行った。本研究で決定されたポプラとキャッサバの EST はそれぞれ、89,572 と 35,400 であった。これは、ポプラ EST 数を 1.3 倍増、キャッサバ EST 数を 2.2 倍増させるものであり、関連する包括的解析などで必要となる情報基盤の整備に関する大きな貢献といえる。本研究で収集した cDNA の完全長率は、いずれも 85%前後と推算され、転写開始点の把握、タンパク質合成等の研究に活用し得ると思われる。また、生物遺伝資源の整備という面においても有

意義と考えられる。

新規転写領域の探索と遺伝子機能解析

ポプラとキャッサバの完全長 cDNA 配列を各々のゲノム塩基配列に整列させたところ、それぞれ、ポプラで 944 ヶ所、キャッサバで 974 ヶ所の新規転写領域を検出した。また、新規遺伝子モデルについても、ポプラで 1,087 個、キャッサバで 751 個を検出するなど、本研究の完全長 cDNA 配列データの有用性が示された。

ゲノム注釈情報の各々のゲノム配列から予測した複合的な遺伝子機能予測や適切な遺伝子機能情報の確認による遺伝子機能注釈は、注釈情報生産の達成だけでなく、実際の分子機能の理解に不可欠であるため、様々なタンパク質データセットや生物種間共通遺伝子(オルソログ)、遺伝子オントロジ(Ashburner et al., 2000)を用い、収集したポプラおよびキャッサバの完全長 cDNA に注釈付けを行った。また、シロイヌナズナの遺伝子情報と比較することで、各生物種固有の遺伝子を推定し、その遺伝子機能の傾向の把握を試みた。その結果、ポプラとキャッサバにおいて、“Signal transduction mechanisms” と “Secondary metabolites biosynthesis, transport and catabolism” に割り当てられた遺伝子の割合が、シロイヌナズナよりも顕著に多かった(調整済み標準化残差分析 $p < 0.05$)。これは、ポプラ、キャッサバといった木本植物の形質と分子機構の関係性を示すものと考えられる。また、遺伝子オントロジを使用することで、より詳細な遺伝子機能分類結果を比較解析することで、“response to stress”、“response to biotic stimulus”、“biosynthetic process”などでシロイヌナズナよりも顕著に高い割合を示した。これらの遺伝子機能比較解析により、木本植物と草本植物の間での差異を示すことができた。

(3) 有用植物の DNA 多型解析と分子育種基盤整備の推進

ゲノム情報の集積により、遺伝子構造と機能の解析が推進され、生物種間の比較解析はますます加速してゲノム情報はより充実すると思われる。ゲノム情報が豊富になることで実現することの1つとして、生物種内における比較解析がある。例えば、個体間、系統間におけるゲノムの違い、すなわち DNA 多型を検出することが可能になり、個体、品種を識別する分子マーカーの整備が推進される。第2章で述べたように、本研究によって大規模にキャッサバ完全長 cDNA 配列データが収集された。その配列情報資源を活用し、キャッサバの DNA 多型の探索を行った。さらに、検出した各 DNA 多型周辺ゲノム領域を増幅するための PCR プライマーペア配列を設計することで、分子育種基盤の整備に貢献した。また、DNA 多型が検出された遺伝子機能やキャッサバゲノムにおける遺伝子重複を精査することで、キャッサバの DNA 多型の意義と傾向を把握した。

キャッサバ転写産物配列データの系統毎分類と DNA 多型探索

キャッサバの EST を公共配列データベース GenBank(Benson et al., 2012)より獲得し、混入する大腸菌やベクター部の配列を除去した結果、有効な配列データとして 80,523 配列を得た。GenBank 形式記述中の系統情報(cultivar tag)および配列情報登録者への確認により、16 系統または cDNA ライブラリに分類した。この EST 配列データにキャッサバゲノム概要塩基配列(系統 AM560-2)由来の予測転写産物配列データを加えた 17 系統、114,674 配列を使用して DNA 多型を探索したところ、10,546 ヶ所の SNP と 674 ヶ所の InDel を検出できた。過去にキャッサバの DNA 多型検出について 2 件の報告があり、その検出数はそれぞれ、186 と 2,954 であった(Lopez et al., 2005; Ferguson et al., 2012)。本研究では、その検出数を大きく上回り、キャッサバにおいて最大規模である。

SNP による非同義一同義塩基置換とタンパク質機能

検出した SNP の内、CDS に位置するものは、6,613 ヶ所(62.7%)であった。CDS 内に SNP が

生じることにより、非同義塩基置換が生じるタンパク質と同義塩基置換であるタンパク質との機能について精査し、比較解析した結果、ユビキチンファミリーや ATP 合成酵素をコードするタンパク質は、顕著に同義塩基置換が生じたタンパク質で大きな割合を示した。対照的に、NB-ARC ドメインやロイシンリッチリピートをコードするタンパク質の割合が、SNP によって非同義塩基置換が生じたタンパク質で顕著に大きかった。この結果は、DNA 多型とキャッサバ系統間での形質の違いに関係すると考えられる。また、タンパク質機能による SNP 傾向と SNP を許容できるタンパク質機能を示唆した。

SNP によるナンセンス置換—読み過ごし置換とタンパク質機能および遺伝子重複との関係性

CDS 内の SNP から 38 ヶ所のナンセンス置換(premature stop substitution; アミノ酸が終止コドンに置換)と 24 か所の読み過ごし置換(read-through substitution; 終止コドンがアミノ酸に置換)を同定し、これらの置換を検出したタンパク質配列を 2 群に分類した。ナンセンス置換、読み過ごし置換を検出したタンパク質を以降、それぞれナンセンス置換タンパク質、読み過ごし置換タンパク質と表現する。この 2 群のタンパク質の機能分類結果を比較解析した結果、“response to abiotic or biotic stimulus”、“response to stress”といったストレス応答に関して、読み過ごし置換タンパク質群で優位に割合が多く、細胞組織化(cell organization)に関しては、ナンセンス置換タンパク質群で割合が多かった。ストレスへの適応などの有用な形質の獲得とタンパク質の伸長が関連し、逆に一般的な細胞組織化に関連するタンパク質のナンセンス置換による縮退は許容されることが示された。

配列類似性に基づき、キャッサバ遺伝子の重複を確認したところ、その 43.3%が重複遺伝子であることが示された。また、ナンセンス置換タンパク質の 58.3%が重複遺伝子であることが明らかになった。ナンセンス置換タンパク質で遺伝子重複が顕著に多く見られた(フィッシャーの正確確率検定 $p < 0.05$)一方、読み過ごし置換タンパク質での重複遺伝子は 27.8%であり、全遺伝子よりも低い頻度を示した。この解析結果は、ナンセンス置換が生じても遺伝子が重複しているため、残った遺伝子が機能することで機能上不利になるタンパク質の縮退を許容できるためと考えられる。読み過ごし置換は、タンパク質を伸長させ、新規機能の獲得に関与したと考えられる。

データベース Cassava Online Archive の構築

本研究によって 1 万を超えるキャッサバの DNA 多型が検出され、大量な関連情報が生産された。これらの解析結果をデータベース Cassava Online Archive(<http://cassava.psc.riken.jp/>)として構築し、インターネット上に公開した。このデータベース構築によって、円滑なキャッサバゲノム情報の閲覧が可能になった。データベース Cassava Online Archive では、SNP や InDel の情報を閲覧するために、キーワード、DNA 多型 ID、遺伝子モデル ID、キャッサバ系統による多様な検索機能を有し、ゲノムブラウザも実装している。各 DNA 多型の詳細ページには、キャッサバゲノムスキャフォールド上の物理的位置情報、関連する遺伝子モデル、系統毎の対立塩基を含み、同定された DNA 多型周辺のゲノム DNA 増幅のためのプライマーペア配列や想定増幅ゲノム塩基配列も提供するなど、分子マーカー整備の推進に貢献するものである。

本データベースは、キャッサバゲノム、遺伝子の注釈情報、独自の遺伝子機能注釈、本研究で同定された DNA 多型情報など、キャッサバに関する多岐にわたる有用情報を提供している。これまでにはこのような総合的データベースは存在していないことから、キャッサバ研究における世界標準のデータベースが構築されたと考えられる。

(4) まとめ

このように本研究は、モデル植物シロイヌナズナの完全長 cDNA 配列データを使用したゲノム解析とタグ挿入変異体を含むゲノム情報の編纂を行い、ポプラとキャッサバの完全長 cDNA 収集とその配列によるゲノム解析およびモデル植物シロイヌナズナとの比較ゲノム解析を行った。さらに、その研究過程で生産した配列データを利用し、キャッサバの DNA 多型解析を行った。すなわち、ゲノム研究におけるモデル生物の研究、それを利用した有用生物種との比較ゲノム研究、さらに有用生物種内の比較ゲノム研究を網羅したものである。

また、本研究で得られた植物のゲノムおよび作物の分子育種研究に関する情報をデータベースとして編纂し、インターネット上に公開した。高速シーケンサの普及と共に、大量データに基づく研究が展開されることが予想され、本研究で得られた知見と構築されたデータベースは、今後のゲノム研究、分子育種の推進に役立つと期待される。

参考文献

- Akano O, Dixon O, Mba C, Barrera E, Fregene M (2002) Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor Appl Genet* **105**: 521-525
- Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D, Carter DE, Marchand T, Risseuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**: 653-657
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389-3402
- Amutha R, Gunasekaran P (2001) Production of ethanol from liquefied cassava starch using co-immobilized cells of *Zymomonas mobilis* and *Saccharomyces diastaticus*. *J Biosci Bioeng* **92**: 560-564
- An D, Yang J, Zhang P (2012) Transcriptome profiling of low temperature-treated cassava apical shoots showed dynamic responses of tropical plant to cold stress. *BMC Genomics* **13**: 64
- Andersen MD, Busk PK, Svendsen I, Moller BL (2000) Cytochromes P-450 from cassava (*Manihot esculenta* Crantz) catalyzing the first steps in the biosynthesis of the cyanogenic glucosides linamarin and lotaustralin. Cloning, functional expression in *Pichia pastoris*, and substrate specificity of the isolated recombinant enzymes. *J Biol Chem* **275**: 1966-1975
- Anderson JV, Delseny M, Fregene MA, Jorge V, Mba C, Lopez C, Restrepo S, Soto M, Piegu B, Verdier V, Cooke R, Tohme J, Horvath DP (2004) An EST resource for cassava and other species of Euphorbiaceae. *Plant Mol Biol* **56**: 527-539
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815
- Asamizu E, Nakamura Y, Sato S, Tabata S (2000) A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res* **7**: 175-180
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29
- Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**: 1803-1818
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res* **41**: D36-42

- Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW** (2012) GenBank. *Nucleic Acids Res* **40**: D48-53
- Bewick V, Cheek L, Ball J** (2004) Statistics review 8: Qualitative data - tests of association. *Crit Care* **8**: 46-53
- Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D** (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**: 956-963
- Carninci P, Shibata Y, Hayatsu N, Sugahara Y, Shibata K, Itoh M, Konno H, Okazaki Y, Muramatsu M, Hayashizaki Y** (2000) Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res* **10**: 1617-1630
- Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Hu TT, Fu G, Hinds DA, Chen H, Frazer KA, Huson DH, Scholkopf B, Nordborg M, Ratsch G, Ecker JR, Weigel D** (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**: 338-342
- Cochrane G, Alako B, Amid C, Bower L, Cerdeno-Tarraga A, Cleland I, Gibson R, Goodgame N, Jang M, Kay S, Leinonen R, Lin X, Lopez R, McWilliam H, Oisel A, Pakseresht N, Pallreddy S, Park Y, Plaister S, Radhakrishnan R, Riviere S, Rossello M, Senf A, Silvester N, Smirnov D, Ten Hoopen P, Toribio A, Vaughan D, Zalunin V** (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res* **41**: D30-35
- Cock JH** (1982) Cassava: a basic energy source in the tropics. *Science* **218**: 755-762
- Cooke R, Raynal M, Laudie M, Grellet F, Delseny M, Morris PC, Guerrier D, Giraudat J, Quigley F, Clabault G, Li YF, Mache R, Krivitzky M, Gy IJ, Kreis M, Lecharny A, Parmentier Y, Marbach J, Fleck J, Clement B, Philipps G, Herve C, Bardet C, Tremousaygue D, Hofte H, et al.** (1996) Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. *Plant J* **9**: 101-124
- Dean FB, Nelson JR, Giesler TL, Lasken RS** (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**: 1095-1099
- Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, Martin MJ, Bely B, Browne P, Mun Chan W, Eberhardt R, Gardner M, Laiho K, Legge D, Magrane M, Pichler K, Poggioli D, Sehra H, Auchincloss A, Axelsen K, Blatter MC, Boutet E, Braconi-Quintaje S, Breuza L, Bridge A, Coudert E, Estreicher A, Famiglietti L, Ferro-Rojas S, Feuermann M, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, James J, Jimenez S, Jungo F, Keller G, Lemercier P, Lieberherr D, Masson P, Moinat M, Pedruzzi I, Poux S, Rivoire C, Roechert B, Schneider M, Stutz A, Sundaram S, Tognolli M, Bougueleret L, Argoud-Puy G, Cusin I, Duek-Roggli P, Xenarios I, Apweiler R** (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res* **40**: D565-570
- El-Sharkawy MA, Cadavid LE** (2002) Response of cassava to prolonged water stress imposed at different stages of growth. *Experimental Agriculture* **38**: 333-350

- Ewing B, Hillier L, Wendl MC, Green P** (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185
- Fauquet CM, Tohme J** (2004) The global cassava partnership for genetic improvement. *Plant Mol Biol* **56**: v-x
- Ferguson ME, Hearne SJ, Close TJ, Wanamaker S, Moskal WA, Town CD, de Young J, Marri PR, Rabbi IY, de Villiers EP** (2012) Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. *Theor Appl Genet* **124**: 685-695
- Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W** (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**: 967-974
- Food and Agricultural Organization of the United Nations** (2003) State of the World's Forests 2003.
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS** (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186
- Gruber M, Wu LM, Links M, Gjetvaj B, Durkin J, Lewis C, Sharpe A, Lydiate D, Hegedus D** (2012) Analysis of expressed sequence tags in *Brassica napus* cotyledons damaged by crucifer flea beetle feeding. *Genome* **55**: 118-133
- Herberth S, Shahriari M, Bruderek M, Hessner F, Muller B, Hulskamp M, Schellmann S** (2012) Artificial ubiquitylation is sufficient for sorting of a plasma membrane ATPase to the vacuolar lumen of *Arabidopsis* cells. *Planta* **236**: 63-77
- Hertzberg M, Aspeborg H, Schrader J, Andersson A, Erlandsson R, Blomqvist K, Bhalerao R, Uhlen M, Teeri TT, Lundberg J, Sundberg B, Nilsson P, Sandberg G** (2001) A transcriptional roadmap to wood formation. *Proc Natl Acad Sci U S A* **98**: 14732-14737
- Higo K, Ugawa Y, Iwamoto M, Higo H** (1998) PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res* **26**: 358-359
- Huang X, Madan A** (1999) CAP3: A DNA sequence assembly program. *Genome Res* **9**: 868-877
- Iida K, Kawaguchi S, Kobayashi N, Yoshida Y, Ishii M, Harada E, Hanada K, Matsui A, Okamoto M, Ishida J, Tanaka M, Morosawa T, Seki M, Toyoda T** (2011) ARTADE2DB: Improved Statistical Inferences for *Arabidopsis* Gene Functions and Structure Predictions by Dynamic Structure-Based Dynamic Expression (DSDE) Analyses. *Plant and Cell Physiology* **52**: 254-264
- Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K** (2004) Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. *Nucleic Acids Res* **32**: 5096-5103
- Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K** (2005) RARTF: database and tools for complete sets of *Arabidopsis* transcription factors. *DNA Res* **12**: 247-256
- International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L,**

- Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Wayne MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861
- Ito T, Motohashi R, Kuromori T, Mizukado S, Sakurai T, Kanahara H, Seki M, Shinozaki K (2002) A new resource of locally transposed Dissociation elements for screening gene-knockout lines in silico on the Arabidopsis genome. *Plant Physiol* **129**: 1695-1699
- Jimenez-Gomez JM, Maloof JN (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. *BMC Plant Biol* **9**: 85
- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**: D109-114
- Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Foundation of Advancement of International Science Genome S, Analysis G, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, Riken, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H,

- Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* **301**: 376-379
- Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, Okubo K, Nakamura Y (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res* **40**: D38-42
- Konno H, Fukunishi Y, Shibata K, Itoh M, Carninci P, Sugahara Y, Hayashizaki Y (2001) Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library. *Genome Res* **11**: 281-289
- Kristiansen TZ, Pandey A (2002) Resources for full-length cDNAs. *Trends Biochem Sci* **27**: 266-267
- Kuromori T, Hirayama T, Kiyosue Y, Takabe H, Mizukado S, Sakurai T, Akiyama K, Kamiya A, Ito T, Shinozaki K (2004) A collection of 11 800 single-copy Ds transposon insertion lines in *Arabidopsis*. *Plant J* **37**: 897-905
- Kuromori T, Wada T, Kamiya A, Yuguchi M, Yokouchi T, Imura Y, Takabe H, Sakurai T, Akiyama K, Hirayama T, Okada K, Shinozaki K (2006) A trial of phenome analysis using 4000 Ds-insertional mutants in gene-coding regions of *Arabidopsis*. *Plant J* **47**: 640-651
- 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 (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* **40**: D1202-1210
- Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Teclé I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res* **36**: D947-953
- Lokko Y, Anderson JV, Rudd S, Raji A, Horvath D, Mikel MA, Kim R, Liu L, Hernandez A, Dixon AG, Ingelbrecht IL (2007) Characterization of an 18,166 EST dataset for cassava (*Manihot esculenta* Crantz) enriched for drought-responsive genes. *Plant Cell Rep* **26**: 1605-1618
- Lopez C, Jorge V, Piegu B, Mba C, Cortes D, Restrepo S, Soto M, Laudie M, Berger C, Cooke R, Delseny M, Tohme J, Verdier V (2004) A unigene catalogue of 5700 expressed genes in cassava. *Plant Mol Biol* **56**: 541-554
- Lopez C, Piegu B, Cooke R, Delseny M, Tohme J, Verdier V (2005) Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). *Theor Appl Genet* **110**: 425-431
- Martienssen RA (1998) Functional genomics: probing plant gene function and expression with transposons. *Proc Natl Acad Sci U S A* **95**: 2021-2026
- Maruyama K, Todaka D, Mizoi J, Yoshida T, Kidokoro S, Matsukura S, Takasaki H, Sakurai T, Yamamoto YY, Yoshiwara K, Kojima M, Sakakibara H, Shinozaki K, Yamaguchi-Shinozaki K (2012) Identification of Cis-Acting Promoter Elements in Cold- and Dehydration-Induced Transcriptional Pathways in *Arabidopsis*, Rice, and Soybean. *DNA Research* **19**: 37-49
- McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR,

- Bureau TE, Stokowski R, Ballinger DG, Frazer KA, Cox DR, Padhukasahasram B, Bustamante CD, Weigel D, Mackill DJ, Bruskiewich RM, Ratsch G, Buell CR, Leung H, Leach JE** (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A* **106**: 12273-12278
- Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M** (1998) *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282**: 662, 679-682
- Mochida K, Shinozaki K** (2010) Genomics and bioinformatics resources for crop improvement. *Plant Cell Physiol* **51**: 497-523
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M** (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**: W182-185
- Myouga F, Akiyama K, Motohashi R, Kuromori T, Ito T, Iizumi H, Ryusui R, Sakurai T, Shinozaki K** (2010) The Chloroplast Function Database: a large-scale collection of *Arabidopsis* Ds/Spm- or T-DNA-tagged homozygous lines for nuclear-encoded chloroplast proteins, and their systematic phenotype analysis. *Plant J* **61**: 529-542
- Nanjo T, Sakurai T, Totoki Y, Toyoda A, Nishiguchi M, Kado T, Igasaki T, Futamura N, Seki M, Sakaki Y, Shinozaki K, Shinohara K** (2007) Functional annotation of 19,841 *Populus nigra* full-length enriched cDNA clones. *BMC Genomics* **8**: 448
- Nishiyama T, Fujita T, Shin IT, Seki M, Nishide H, Uchiyama I, Kamiya A, Carninci P, Hayashizaki Y, Shinozaki K, Kohara Y, Hasebe M** (2003) Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. *Proc Natl Acad Sci U S A* **100**: 8007-8012
- Ogasawara O, Mashima J, Kodama Y, Kaminuma E, Nakamura Y, Okubo K, Takagi T** (2013) DDBJ new system and service refactoring. *Nucleic Acids Res* **41**: D25-29
- Okogbenin E, Fregene M** (2002) Genetic analysis and QTL mapping of early root bulking in an F1 population of non-inbred parents in cassava (*Manihot esculenta* Crantz). *Theor Appl Genet* **106**: 58-66
- Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S** (2012) The Cassava Genome: Current Progress, Future Directions. *Trop Plant Biol* **5**: 88-94
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD** (2012) The Pfam protein families database. *Nucleic Acids Res* **40**: D290-301
- Rabbi IY, Kulembeka HP, Masumba E, Marri PR, Ferguson M** (2012) An EST-derived SNP and SSR genetic linkage map of cassava (*Manihot esculenta* Crantz). *Theor Appl Genet* **125**: 329-342
- Raheem D, Chukwuma C** (2001) Foods from cassava and their relevance to Nigeria and other African countries. *Agriculture and Human Values* **18**: 383-390
- Raji AA, Anderson JV, Kolade OA, Ugwu CD, Dixon AG, Ingelbrecht IL** (2009) Gene-based microsatellites for cassava (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. *BMC Plant Biol* **9**: 118
- Ralph S, Oddy C, Cooper D, Yueh H, Jancsik S, Kolosova N, Philippe RN, Aeschliman D, White R,**

- Huber D, Ritland CE, Benoit F, Rigby T, Nantel A, Butterfield YS, Kirkpatrick R, Chun E, Liu J, Palmquist D, Wynhoven B, Stott J, Yang G, Barber S, Holt RA, Siddiqui A, Jones SJ, Marra MA, Ellis BE, Douglas CJ, Ritland K, Bohlmann J (2006) Genomics of hybrid poplar (*Populus trichocarpax deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar. *Mol Ecol* **15**: 1275-1297
- Ramirez SR, Basu C (2009) Comparative Analyses of Plant Transcription Factor Databases. *Current Genomics* **10**: 10-17
- Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365-386
- Sakurai T, Plata G, Rodriguez-Zapata F, Seki M, Salcedo A, Toyoda A, Ishiwata A, Tohme J, Sakaki Y, Shinozaki K, Ishitani M (2007) Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response. *BMC Plant Biol* **7**: 66
- Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Krasnov S, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller Y, Karsch-Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E, Ye J (2012) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **40**: D13-25
- Schindler S, Szafranski K, Hiller M, Ali GS, Palusa SG, Backofen R, Platzer M, Reddy ASN (2008) Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes. *BMC Genomics* **9**
- Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296**: 141-145
- Shahriari M, Richter K, Keshavaiah C, Sabovljevic A, Huelskamp M, Schellmann S (2011) The *Arabidopsis* ESCRT protein-protein interaction network. *Plant Molecular Biology* **76**: 85-96
- Soderlund C, Descour A, Kudrna D, Bomhoff M, Boyd L, Currie J, Angelova A, Collura K, Wissotski M, Ashley E, Morrow D, Fernandes J, Walbot V, Yu Y (2009) Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS Genet* **5**: e1000740
- Sojikul P, Kongsawadworakul P, Viboonjun U, Thaiprasit J, Intawong B, Narangajavana J, Svasti MR (2010) AFLP-based transcript profiling for cassava genome-wide expression analysis in the onset of storage root formation. *Physiol Plant* **140**: 189-198
- Sraphet S, Boonchanawiwat A, Thanyasiriwat T, Boonseng O, Tabata S, Sasamoto S, Shirasawa K, Isobe S, Lightfoot DA, Tangphatsornruang S, Triwitayakorn K (2011) SSR and EST-SSR-based genetic linkage map of cassava (*Manihot esculenta* Crantz). *Theor Appl Genet* **122**: 1161-1170
- Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, Shiraki T, Hayashizaki Y, Champe M, Pacleb J,

- Wan K, Yu C, Carlson J, George R, Celniker S, Rubin GM (2002) The *Drosophila* gene collection: identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res* **12**: 1294-1300
- Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S (2004) A *Populus* EST resource for plant functional genomics. *Proc Natl Acad Sci U S A* **101**: 13951-13956
- Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, Villarreal R, Van Montagu M, Sandberg G, Olsson O, Teeri TT, Boerjan W, Gustafsson P, Uhlen M, Sundberg B, Lundeberg J (1998) Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proc Natl Acad Sci U S A* **95**: 13330-13335
- Suzuki Y, Yamashita R, Nakai K, Sugano S (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* **30**: 328-331
- Taji T, Sakurai T, Mochida K, Ishiwata A, Kurotani A, Totoki Y, Toyoda A, Sakaki Y, Seki M, Ono H, Sakata Y, Tanaka S, Shinozaki K (2008) Large-scale collection and annotation of full-length enriched cDNAs from a model halophyte, *Thellungiella halophila*. *BMC Plant Biol* **8**: 115
- Tameling WI, Vossen JH, Albrecht M, Lengauer T, Berden JA, Haring MA, Cornelissen BJ, Takken FL (2006) Mutations in the NB-ARC domain of I-2 that impair ATP hydrolysis cause autoactivation. *Plant Physiol* **140**: 1233-1245
- Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33-36
- Tonukari NJ (2004) Cassava and the future of starch. *Electronic Journal of Biotechnology* **7**: 5-8
- 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 (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604
- Umezawa T, Sakurai T, Totoki Y, Toyoda A, Seki M, Ishiwata A, Akiyama K, Kurotani A, Yoshida T, Mochida K, Kasuga M, Todaka D, Maruyama K, Nakashima K, Enju A, Mizukado S, Ahmed S, Yoshiwara K, Harada K, Tsubokura Y, Hayashi M, Sato S, Anai T, Ishimoto M, Funatsuki H,

- Teraishi M, Osaki M, Shinano T, Akashi R, Sakaki Y, Yamaguchi-Shinozaki K, Shinozaki K** (2008) Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA Res* **15**: 333-346
- Utsumi Y, Tanaka M, Morosawa T, Kurotani A, Yoshida T, Mochida K, Matsui A, Umemura Y, Ishitani M, Shinozaki K, Sakurai T, Seki M** (2012) Transcriptome Analysis Using a High-Density Oligomicroarray under Drought Stress in Various Genotypes of Cassava: An Important Tropical Crop. *DNA Res*
- Wullschlegel SD, Jansson S, Taylor G** (2002) Genomics and forest biology: Populus emerges as the perennial favorite. *Plant Cell* **14**: 2651-2655
- Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang W** (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotechnol* **30**: 105-111
- Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, Imanishi T** (2008) Distribution and effects of nonsense polymorphisms in human genes. *PLoS One* **3**: e3393
- Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T** (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**: 67

謝辞

本研究の過程において、終始懇篤なるご指導とご鞭撻を賜り、本論文を纏めるに際しても、親身な助言と激励をいただきました東京大学大学院農学生命科学研究科 教授 篠崎和子博士に心より感謝致しますとともに、厚く御礼申し上げます。

独立行政法人 理化学研究所 環境資源科学研究センター センター長 篠崎一雄博士は、植物科学研究における情報科学の意義を早くに認識され、植物ゲノム研究に携わる機会を与えてくださいました。心よりお礼申し上げます。

本研究における試料作成、情報提供、議論にあたり、多大なるご協力とご教示をいただきました独立行政法人 森林総合研究所 コーディネーター 篠原健司博士、室長 楠城時彦博士、国際熱帯農業研究センター 上席研究員 石谷学博士に深くお礼申し上げます。理化学研究所ゲノム科学総合研究センター在職時に、シロイヌナズナのゲノム研究へのきっかけを与えてくださいました関原明博士、黒森崇博士、伊藤卓也博士に深くお礼申し上げます。当時、上司であり、理化学研究所 ゲノム科学総合研究センター 知識ベース研究開発チーム チームリーダーであった豊田哲郎博士には、大量に生産した解析データのデータベース構築にあたり、ご助言をいただきました。深くお礼申し上げます。

理化学研究所 環境資源科学研究センター 副チームリーダー 持田恵一博士、京都大学 助教 飯田慶博士、関東学院大学 助教 近藤陽一博士には、論文作成にとどまらず、公私にわたり温かな交流と大きな刺激を与えていただき心より感謝申し上げます。また、ご支援いただきましたすべての皆様に心より感謝申し上げます。

最後に私事ではありますが、温かく見守ってくれた両親と兄、そして日々支えてくれた妻真実と子供たちに感謝し、謝辞と致します。

平成 25 年 12 月

独立行政法人 理化学研究所 環境資源科学研究センター
統合ゲノム情報研究ユニット 櫻井哲也