Rorippa aquatica
染色体レベルゲノムアセンブリデータベースの構築

坂 本 智 昭*
木 村 成 介*

要 旨

Rorippa aquatica は北米原産のアブラナ科植物であり、周囲の環境に応答して葉の形態を変化させる異形葉性を示す。また、葉断片から完全な植物体を形成することが出来るという特性を持つ。このように興味深い特性を示す本種であるが、それらの根底にある分子機構は完全には明らかになっていない。それらを解明するための手法として次世代シーケンサーを用いた様々な網羅的な解析が行われている。ゲノムアセンブリでは染色体レベルのゲノム配列情報とゲノム上に存在する遺伝子の情報が取得され、これにより全遺伝子を対象とした網羅的解析が可能になった。しかしながら、ゲノム・遺伝子情報は膨大であり情報の検索・取得を簡易に行うことが出来る環境が求められた。そこで本研究ではゲノム・遺伝子情報を統合して保管する *R. aquatica* 染色体レベルゲノムアセンブリデータベースの構築とデータを簡易に検索・取得することを可能にするインターフェイスの整備を試みた。

キーワード：*Rorippa aquatica*、ゲノムアセンブリ、データベース、SQLite、Perl/CGI

1. はじめに

Rorippa aquatica は北米原産のアブラナ科植物である。この植物は陸上でも水中でも生育可能な水陸両生植物であり、水中では陸上とは異なる形態の葉を形成する。このように環境応答性の異形葉性を示すが、水中と陸上といった環境の違いだけでなく、温度や光といった要素も葉の形態に影響を与えることが明らかになっている（Nakayama et al. 2018）。また、本種は外部から植物ホルモンを添加せずとも葉断片からの完全な植物体の再生を行うことが可能である（Amano et al. 2020）。

このように興味深い特性を示す本種であるが、モデル植物ではないため本種および同属の近縁種でのゲノム情報および網羅的な遺伝子情報が得られていなかった。そのため、次世代シーケンサーデータを用いて de novo トランスクリプトームによる網羅的遺伝子情報の取得、de novo ゲノムアセンブリおよび遺伝子アノテーションが行われてきた。その結果、Hi-Cseq データを加えてのゲノムアセンブリとスキャフォールディングにより染色体レベルのゲノム配列とゲノム上に存在する遺伝子の情報が得られ、そのゲノム情報をもとに *R. aquatica* の様々な形質における分子的な解析が進められている。

* 京都産業大学生命科学部

しかし、得られたゲノム情報は多岐にわたり膨大な数の遺伝子情報を含んでおり、特定の遺伝子を標的とした解析にあたってはより簡易に情報に到達できる環境が必要とされた。そこで本研究では *R. aquatica* 染色体レベルゲノムアセンブリによって得られたゲノムおよび遺伝子アノテーション情報を網羅的かつ統合的に収容するデータベースの構築を試みた。さらにデータベース内の情報へのアクセスを簡易に行うためのインターフェイスの整備を行った。

2. *R. aquatica* ゲノムデータベースの構築

ゲノムアセンブリと遺伝子アノテーションから得られたデータはテキストベースで保存されており、そのままでは検索が煩雑であり複数の異なる形式のデータ間の関連付けが困難であった。そのため、それらのデータを一括して管理するためのデータベースを構築した。データベース言語には管理・移植の簡易さから SQLite を選択した。単一のデータベースファイルを作成し、それぞれの情報ごとに以下のようなテーブルを設定し、必要なデータのインポートを行った。

2.1 ゲノム配列データ

ゲノムアセンブリにより 15 本の染色体配列と 2043 本の染色体に組み込まれなかった断片配列が得られている。その配列をゲノム配列データテーブルとしてデータベースに収容した。

表 1　ゲノム配列データテーブルフォーマット

カラム番号	カラム名	カラム内容詳細
1	SeqName	ゲノム配列名。本テーブル内では同一名のデータは含まれない。RaCh01 から RaChr15 までの染色体配列名と Ra_scaffold_XXXXX（XXXXX は 5 桁の整数からなる配列番号）と名付けられた断片配列名が含まれる。
2	Length	塩基配列の長さ。
3	Sequence	各配列の全長塩基配列

2.2　遺伝子情報データ

ゲノム配列データと近縁種のシロイヌナズナの遺伝子情報やこれまでに得られた *R. aquatica* RNAseq データを基にゲノム上の 46200 の遺伝子座が推定されている。この遺伝子情報は GFF3 形式にて出力されている。このデータをデータベース用に遺伝子の各転写産物（isoform）毎に整理しデータベースにインポートした。遺伝子名および isoform 名は次のような命名規則により命名されている。遺伝子名はゲノム配列名と各配列毎の遺伝子番号から命名されている（例：RaChr04G10300）。各遺

伝子の isoform には番号が割り振られており、それを遺伝子名に付加したものを isoform 配列名としている（例：RaChr04G10300.1）。インポートされた isoform 情報は refflat 形式に準じた形式に変換されており、遺伝子の位置、エキソンの開始終止位置および染色体配列に対する遺伝子の向き（ストランド方向）といった情報を含んでいる。

また、遺伝子情報から得られる配列情報（転写産物配列、Coding Sequence（CDS）、翻訳アミノ酸配列）もゲノム配列と同様にそれぞれ別々のテーブルを作成しデータベースにあらかじめ登録しデータアクセスを迅速に行えるようにした。

表 2　遺伝子配列情報データテーブルフォーマット

カラム番号	カラム名	カラム内容詳細
1	isoformName	Isoform 配列名。
2	Length	配列の長さ。
3	Sequence	各配列の全長配列。

2.3　Orthogroup データ

アブラナ科ではモデル植物であるシロイヌナズナの遺伝子情報に基づいた研究がもっとも進んでいる。シロイヌナズナの遺伝子研究における成果は、*R. aquatica* における分子生物学研究でも応用できるものであると考えられ、シロイヌナズナと *R. aquatica* の相同遺伝子を比較しての議論が必要になると思われる。しかし、シロイヌナズナの遺伝子が *R. aquatica* のどの遺伝子と相同であるかは明らかになっていなかった。そこでシロイヌナズナを含む全遺伝子情報が既知のアブラナ科植物のアミノ酸情報を取得し、アミノ酸配列の相同性を基に全遺伝子を Orthogroup（相同遺伝子（ortholog）から構成される遺伝子グループ）に分類した。この分類結果からシロイヌナズナと *R. aquatica* のデータを抽出し Orthogroup データテーブルとしてデータベースにインポートした。さらに、過去の *R. aquatica* の遺伝子データとの紐づけのために 2017 年時点でのゲノム・遺伝子データの Orthogroup 情報も同様に追加した。

表3　Orthogroup データテーブルフォーマット

カラム番号	カラム名	カラム内容詳細
1	Orthogroup	Orthogroup 名（例：OG0008513）。
2	GeneID	遺伝子 ID。種毎に形式は異なる。
3	Species	種名。*Arabidopsis thaliana*、*R. aquatica* および旧データとして *R. aquatica*（2017）が含まれる。

2.4　blastp 検索結果データ

遺伝子推定によってゲノム上の遺伝子位置、転写産物配列およびそれらがコードするアミノ酸配列が明らかになった。しかしながら、それらの機能的特徴は明らかではない。そこで公開データベースに登録されている遺伝子情報との比較により、*R. aquatica* 遺伝子の特徴づけ（アノテーション）を行った。比較の方法として *R. aquatica* 遺伝子のコードするアミノ酸配列と既知のアミノ酸配列データベース Non-redundant protein sequences from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq（nr）を用いて blastp によるアミノ酸配列 - アミノ酸相同性検索を用いた。xml 形式で出力された各遺伝子の blastp 検索結果から検索にヒットした遺伝子名およびその遺伝子が属する種名、遺伝子間の相同性に関わる数値（e-value 等）を抽出し、blastp 検索結果データテーブルとしてインポートした。

3. データベースアクセスインターフェイスの整備

このように現時点での *R. aquatica* ゲノムおよび遺伝子アノテーション情報からデータベースを構築した。しかしながら、データベース単体のみでは情報を検索・閲覧するためにデータベース言語である SQLite への理解とコマンドライン上での操作が必要とする。そのため、それぞれのデータ取得を簡易に行うために、インターネットブラウザからアクセス可能な Web インターフェイスを構築した。本インターフェイスは試験段階であるため、研究室内の PC にサーバーを設置し、研究室内のローカルネットワーク内で運用している。Web インターフェイスは3種の検索・情報取得のためのインターフェイスからなるトップページ（図1）とそれらに入力された情報に基づいて検索・情報取得したデータを表示する検索結果ページから構成されている。トップページは HTML による静的な画面表示を、一方で検索結果画面は入力されたキーワードに基づいて Perl/CGI による動的な生成によって Web ページの表示を行う方法を選択した。検索結果を動的に生成させることで今後のデータベースへの情報の追加や変更に対して容易に対応することを可能にした。トップページは以下のように3種の検索・情報取得インターフェイスから構成されている。

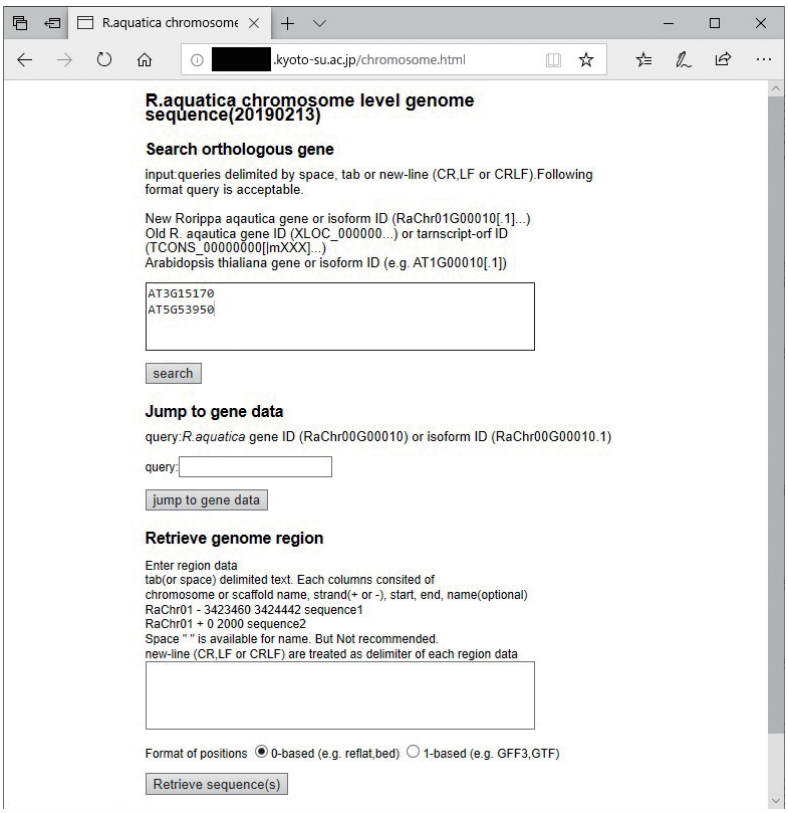


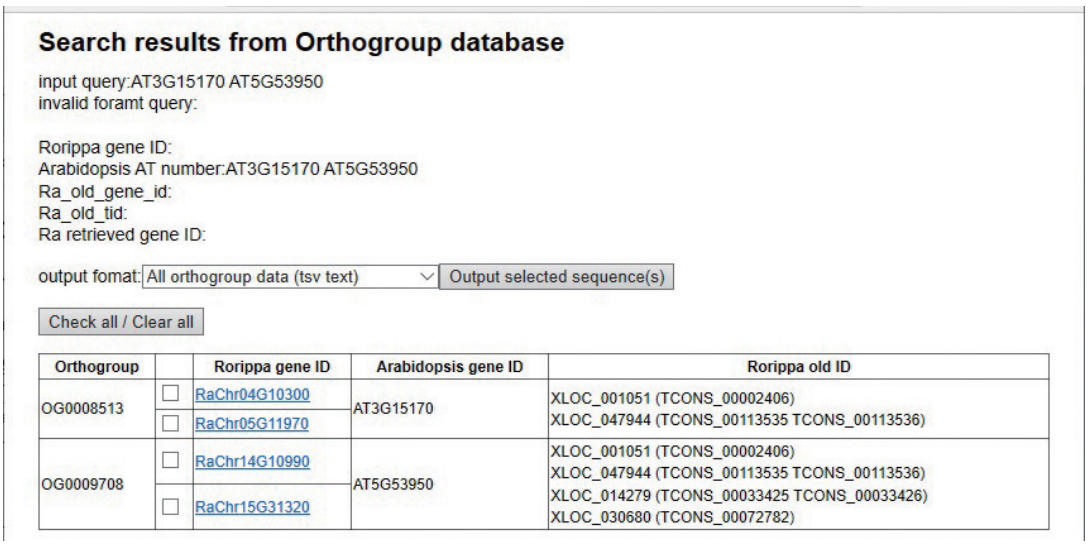
図１　検索インターフェイストップページ表示

3.1.1　Orthogroup 検索

orthogroup 検索（図１、上段）では遺伝子名や転写産物名からそれらと相同な *R. aquatica* の遺伝子のリストを取得することができる。検索文字列としてシロイヌナズナ遺伝子 ID および旧型式の *R. aquatica* 遺伝子・isoform 名が使用可能である。また、*R. aquatica* の遺伝子・isoform ID も使用可能である。これは *R. aquatica* ゲノム上に複数コピー存在する相同遺伝子を検索するためである。

3.1.2　Orthogroup 検索結果表示画面

Orthogroup 検索結果画面（図２）では検索に用いた遺伝子が属する Orthogroup を表示する。検索にヒットした Orthogroup 情報として Orthogroup 名、*R. aquatica* 遺伝子名、シロイヌナズナ TAIR 遺伝子 ID、旧 *R. aquatica* 遺伝子および isoform ID を表示する。*R. aquatica* 遺伝子名にはそれぞれの遺伝子情報ページへのリンクが追加されている。また、検索結果をテキスト形式で出力することも可能である。



Search results from Orthogroup database

input query:AT3G15170 AT5G53950
invalid foramt query:

Rorippa gene ID:
Arabidopsis AT number:AT3G15170 AT5G53950
Ra_old_gene_id:
Ra_old_tid:
Ra retrieved gene ID:

output fomat: All orthogroup data (tsv text) Output selected sequence(s)

Check all / Clear all

Orthogroup		Rorippa gene ID	Arabidopsis gene ID	Rorippa old ID
OG0008513	☐	RaChr04G10300	AT3G15170	XLOC_001051 (TCONS_00002406) XLOC_047944 (TCONS_00113535 TCONS_00113536)
	☐	RaChr05G11970		
OG0009708	☐	RaChr14G10990	AT5G53950	XLOC_001051 (TCONS_00002406) XLOC_047944 (TCONS_00113535 TCONS_00113536) XLOC_014279 (TCONS_00033425 TCONS_00033426) XLOC_030680 (TCONS_00072782)
	☐	RaChr15G31320		

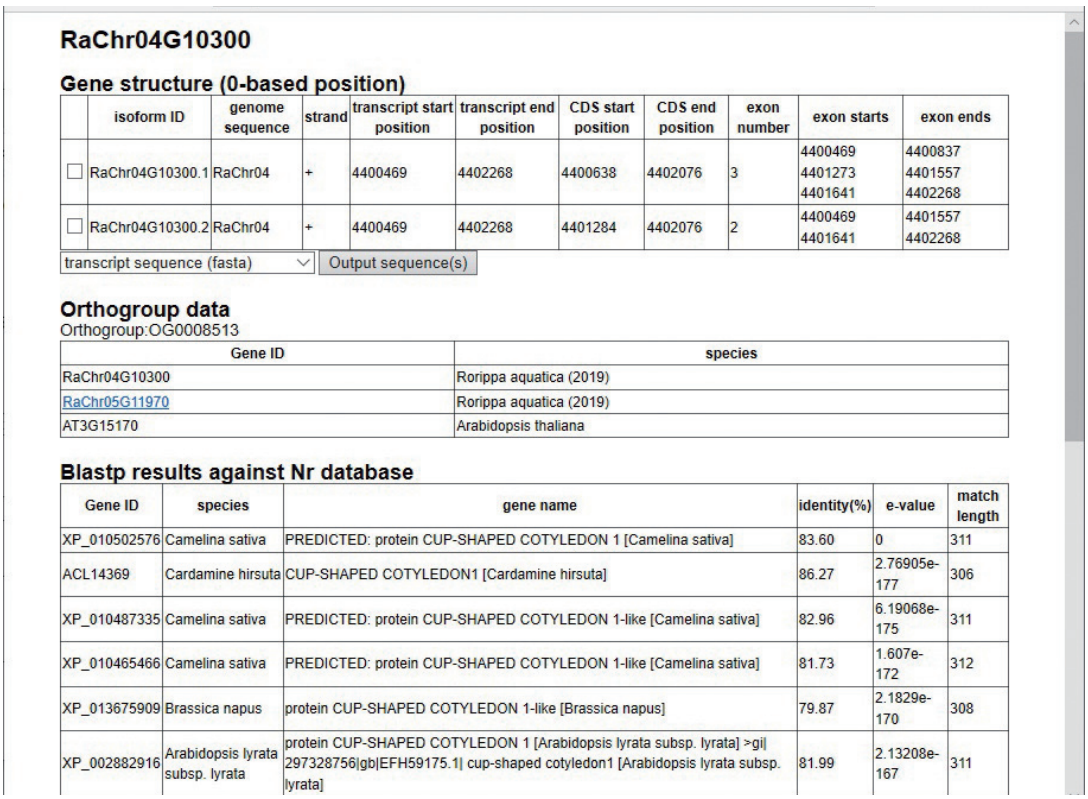
図 1　Orthoroup 検索結果表示

3.2　遺伝子情報表示画面

遺伝子情報画面（図 3）では *R. aquatica* の遺伝子情報が表示される。表示される遺伝子情報は以下の 3 種である。

表 4　遺伝子情報表示データ

遺伝子構造情報	遺伝子のゲノム配列上の位置情報を表示する。遺伝子情報にはエキソンの開始終止点、翻訳開始終止位置も含まれている。遺伝子が複数の isoform を持つ場合は isoform ごとの情報が表示される。
Orthogroup 情報	遺伝子が属する orthogroup の情報。Orthogroup 名とそれに属する *R. aquatica* 遺伝子およびシロイヌナズナ遺伝子の ID リストが表示される。
balstp 検索結果情報	各遺伝子の blastp 検索結果を表示する。上述の blastp 検索結果データから遺伝子 ID、種名、遺伝子詳細や相同性を示す数値を抽出して表示する。



RaChr04G10300

Gene structure (0-based position)

	isoform ID	genome sequence	strand	transcript start position	transcript end position	CDS start position	CDS end position	exon number	exon starts	exon ends
☐	RaChr04G10300.1	RaChr04	+	4400469	4402268	4400638	4402076	3	4400469 4401273 4401641	4400837 4401557 4402268
☐	RaChr04G10300.2	RaChr04	+	4400469	4402268	4401284	4402076	2	4400469 4401641	4401557 4402268

transcript sequence (fasta) Output sequence(s)

Orthogroup data

Orthogroup:OG0008513

Gene ID	species
RaChr04G10300	Rorippa aquatica (2019)
RaChr05G11970	Rorippa aquatica (2019)
AT3G15170	Arabidopsis thaliana

Blastp results against Nr database

Gene ID	species	gene name	identity(%)	e-value	match length
XP_010502576	Camelina sativa	PREDICTED: protein CUP-SHAPED COTYLEDON 1 [Camelina sativa]	83.60	0	311
ACL14369	Cardamine hirsuta	CUP-SHAPED COTYLEDON1 [Cardamine hirsuta]	86.27	2.76905e-177	306
XP_010487335	Camelina sativa	PREDICTED: protein CUP-SHAPED COTYLEDON 1-like [Camelina sativa]	82.96	6.19068e-175	311
XP_010465466	Camelina sativa	PREDICTED: protein CUP-SHAPED COTYLEDON 1-like [Camelina sativa]	81.73	1.607e-172	312
XP_013675909	Brassica napus	protein CUP-SHAPED COTYLEDON 1-like [Brassica napus]	79.87	2.1829e-170	308
XP_002882916	Arabidopsis lyrata subsp. lyrata	protein CUP-SHAPED COTYLEDON 1 [Arabidopsis lyrata subsp. lyrata] >gi|297328756|gb|EFH59175.1| cup-shaped cotyledon1 [Arabidopsis lyrata subsp. lyrata]	81.99	2.13208e-167	311

図３　遺伝子情報表示

3.2.2　データ取得

遺伝子情報ページでは情報の表示だけでなく、様々な遺伝子情報をデータベースから取得することが出来る。取得できるデータは以下の通りでありページ途中のリストボックスから選択して実行する。複数の isoform を持つ遺伝子から必要な情報のみを取得できるように、isoform 毎にチェックボックスを設けて情報の選択を可能にした。

表５　取得可能な遺伝子データ

配列データ	選択された isoform の配列データを fasta 形式で出力する。出力可能な配列データは転写産物、CDS、アミノ酸配列から選択できる。
Isoform 名リスト	Isoform 名のリスト
Isoform 構造情報	Isoform のゲノム配列上での位置情報を取得する。データはタブ区切りテキストからなる refflat 形式で出力される。

3.3 遺伝子情報ページへのジャンプ検索

遺伝子名が事前にわかっている場合には上述の *Rorippa* 遺伝子情報画面へは検索画面を通さずにジャンプ検索インターフェイス（図 1 中段）から直接移動することができる。キーワードとして *R. aquatica* 遺伝子名および isoform 名を使用可能である。

3.4 染色体上区間配列の取得

上述のとおり各遺伝情報ページからあらかじめ用意された遺伝子配列を取得することができる。しかし、プロモーター解析等では遺伝子情報として規定されていない領域の配列が必要になることが想定される。そこで、染色体上の任意の区間配列を抽出可能なインターフェイスを設置した（図 1 下段）。抽出には染色体名、ストランド方向、開始終止位置が必要であり、それらをタブまたはスペース区切りテキストとして入力する。入力された情報をもとに染色体区間配列を fasta 形式で出力する。また、オプションとして任意の配列名を入力することができる。配列名が入力されなかった場合は入力データから配列名を作成し出力データに付与する。改行文字で区切られた複数の位置情報データを入力することが可能であり、それぞれの配列を multi-fasta 形式で出力する。

4. 今後の展開

本 Web インターフェイスは、データベースに収容されたデータを Perl/CGI によって検索し、目的に合わせた形で動的な HTML 生成を行うことで必要なデータ表示を行っている。このことは Perl/CGI スクリプトの修正・変更によって表示の変更・機能の追加が可能であることを示している。現在、ゲノム・遺伝子アノテーション情報を基盤として *R. aquatica* の様々な生理現象を対象とした分子生物学解析が進められている。これらの解析から得られたデータをデータベースにフィードバックすることでより有用なデータベースが構築できると思われる。様々な解析間の連携を強め、多角的な視点からの知見を得ることで *R. aquatica* を用いた研究が発展することが期待される。

謝辞

本研究は，京都産業大学科研費再挑戦支援プログラム「植物の水環境への適応形質の進化と収斂」（課題番号 E1903）の研究の一部として実施された。また，研究の一部は，平成 27 年度私立大学戦略的研究基盤形成支援事業「植物における生態進化発生学研 究拠点の形成―統合オミックス解析による展開―（課題番号 S1511023）」，および，平成 30 科学研究費助成事業（科学研究費補助金）（新学術領域研究（研究領域提案型））「水陸両生植物の水環境への適応に寄与する環境応答統御システムの解析（課題番号 18H04787）」の支援を受けて実施した。

参考文献

Amano, R, Nakayama, H, Momoi, R, Omata, E, Gunji, S, Takebayashi, Y, Kojima, M, Ikematsu, S, Ikeuchi, M, Iwase, A, Sakamoto, T, Kasahara, H, Sakakibara, H, Ferjani, A and Kimura, S（2020）Molecular Basis for Natural Vegetative Propagation via Regeneration in North American Lake Cress, *Rorippa aquatica*（Brassicaceae）.. Plant Cell Physiol 61:353-369.

Nakayama, H, Sakamoto, T, Okegawa, Y, Kaminoyama, K, Fujie, M, Ichihashi, Y, Kurata, T, Motohashi, K, Al-Shehbaz, I, Sinha, N and Kimura, S（2018）Comparative transcriptomics with self-organizing map reveals cryptic photosynthetic differences between two accessions of North American Lake cress.. Sci Rep 8:3302.

Construction of chromosome level genome assembly database of *Rorippa aquatica*.

Tomoaki SAKAMOTO
Seisuke KIMURA

Abstract

Semi-aquatic plant *Rorippa aquatica* (Brassicaceae) showed various physiological traits, such as leaf form conversion in response to environmental conditions and regeneration from a piece of leaf without exogenous plant hormones treatment. The molecular mechanisms underlying these phenomenon were not identified yet. To identify them, various omics analysis with next generation sequencer data were performed. And then, chromosome level genome sequences and whole gene information were obtained. Although these information was useful for molecular analysis, huge amount of data was difficult to handle. We tried to construct genome database for integration of whole genomic and genetic information. Furthermore, graphical web interface environment was established to easy to search and access various data stored in database.

Keywords：*Rorippa aquatica*, Genome database, Genomic analysis, Chromosome level genome assembly, Next0generation sequencing technology