第1章　第2節

次世代シークエンサを使った バイオインフォマティクスツール概要

株式会社キアゲン　　宮本 真理

次世代シークエンサ（next generation sequencer, NGS）の登場によりゲノム解析が従来よりも身近なものとなり，それにあわせて解析方法と解析ツールが開発されてきた.解析ツールはコマンドラインを利用したものから，グラフィカルなユーザーインターフェースを備えたものなど，これまでバイオインフォマティクスをあまり行ってこなかった研究者にも活用しやすいようなツールが無償・有償含めて開発されている.次世代シークエンサはゲノム資源としてのゲノム配列決定，品種改良のためのマーカーの作成などへ活用されている.ここでは，果樹ゲノム解析へ利用出来るバイオインフォマティクスツールについて紹介する.

1）ゲノムアセンブリ

ゲノムアセンブリ（genome assembly）とはシークエンサから得られるDNA断片をつなぎ合わせて長い配列を得ることである．以前はヒト一人分のゲノムを解読するために大量な資金と時間が必要であった．たとえばヒトゲノムプロジェクトでは，開始当時の1990年に30億塩基を読むために30億ドル15年が必要と見積もられた．開始翌年の1991年には必要な予算は27億ドルと見積もられプロジェクトは2年前倒しの2003年に終了した（National Human Genome Research Institute. “The Human Genome Project Completion： Frequently Asked Questions”, URL1-2-1）．次世代シークエンサの出現によりそのコス

Mari MIYAMOTO　✉Mari.Miyamoto@qiagen.com
An overview of bioinformatics tools using Next Generation Sequencers

トも時間も大幅に減少し，最新のIllumina社のHiSeq X Ten システムでは，ヒトゲノムを1000ドル，2日で読むことが可能となっている（Illumina社，URL1-2-2）．現在ドラフトゲノムが公開されている果樹の一部を表1-2-1に示す．本書第1章のゲノム解読された果樹のリストとともに参照していただきたい．

ゲノムアセンブリの方法は，従来のサンガー法により解読された配列と，次世代シークエンサを使って得られた方法では大きく異なる．サンガー法では，得られる配列の長さが数千塩基で，一回のランで出力されるデータ量もそれほど多くなく，シークエンサで得られた配列の断片の中で重なり合う箇所を元につなげていくという方法がとられていた．次世代シークエンサの中でも比較的スループット（throughput）が小さく，平均リード長が400塩基程度と長いRoche 454シークエンサ（Roche社）にはこの方法は利用可能であったが，現在最もよく利用されているIllumina社のシークエンサに特徴づけられるような短く大量なリードのアセンブリには，この方法は向かず，グラフ理論（graph theory）を利用したアルゴリズムVelvet（URL1-2-3, Zerbino and Birney, 2008）が開発された．現在次世代シークエンサの次の世代として第3世代シークエンサが登場している．第3世代シークエンサとは、1分子レベルでDNAを読む技術を持つシークエンサであり，Pacific Bioscience社のPacBio RS II シークエンサが第3世代シークエンサにあたる．PacBio RS II シークエンサは、SMRT（Single Molecule Real Time）テクノロジーという技術を使い、1分子のDNAを鋳型としてシークエンスを行う（Eid *et al.* 2009）．PacBio RS II シークエンサから得られるデータは，Illumina社やRoche社のリードとは異なり，リードの長さが平均で20 kbを超え（Pacific Bioscience社，PacBioRS II Brochure, URL1-2-4, 2015-11-28），第2世代シークエンサーとは異なるアセンブリ手法が開発された（Chin *et al.* 2013）．

次世代シークエンサの普及と共にいくつかのゲノムアセンブラ（assembler）は哺乳類のようなゲノムサイズが大きく，二倍体のアセンブリも行えるレベルとなっている（Luo *et al.* 2012）．次世代シークエンサ用のゲノムアセンブリとして比較的ゲノムサイズが大きなものにも適用可能なアセンブラ

表1-2-1 解読された果樹ゲノムの一部

通称	学名	シークエンステクノロジー	ゲノムサイズ (Mb)	アセンブリサイズ	スキャフォールド	遺伝子数	染色体数	倍数性	アセンブルツール	発表された年
Dates	*Ziziphus jujuba*	Illumina	444	438	5898 (>100bp)	32,808	12	2	SOAPdenovo	2014
Pear	*Pyrus bretschneideri*	Illumina BAC-to-BAC	527	512	2103 (>100bp)	42,812	17	2	SOAPdenovo SSPACE	2013
Sweet orange	*Citrus sinensis*	Illumina	367	320	16890 (>500bp)	29,445	9	2	SOAPdenovo OPERA	2013
Apple	*Malus domestica*	Sanger+454	742	598	163	95,216 /57,386	17	2	論文に明記なし	2010
Grape	*Vitis vinifera*	Sanger	475	487	3,514	30,434	19	2	Arachne	2007
Peach	*Prunus persica*	Sanger	265	227	391 (>1Mbp)	27,852	8	2	Arachne2	2013
Papaya	*Carica papaya*	Sanger	372	271	17,764	24,746	9	2	Arachne2	2008

は，SOAPdenovo（URL1-2-5, Li *et al.* 2010；Luo *et al.* 2012），Velvet, ALLPATHS-LG（URL1-2-6, Gnerre *et al.* 2011）や，倍数性（ploidy）が高い場合にも対応しているPlatanus（URL1-2-7, Kajitani *et al.* 2014）などが挙げられる．これらのツールはコマンドラインで利用し，使用するメモリやコンピュータ資源が比較的大きなものが必要となる．商用ツールでは，CLC Genomics Workbench（QIAGEN社），CLC Assembly Cell（QIAGEN社），NextGENe（Softgenetics社）などがある．CLCのアセンブラは他のツールと比較し使用するメモリが少ないこと（Jüemann *et al.* 2014），CLC Genomics Workbench はグラフィカルなインターフェースが付き，Windows, Mac, Linux いずれのOSでも実施できることがメリットとなる．

果樹など植物のゲノムではゲノムサイズだけでなく，高い倍数性，ヘテロ性（heterogeneity），また繰り返し領域（repeat region）の多さなど，アルゴリズムが苦手とする要素が多くゲノムアセンブリは難しくなる（Vinson *et al.* 2005；Velasco *et al.* 2007；Zheng *et al.* 2013）．そのような場合は，利用しているツールのパラメータの検討，（VelvetではK-mer, CLCではword sizeや

bubble sizeといったアセンブリの結果に影響を与えるパラメータ．ツールのマニュアルに記載がある），対象とするゲノムの何倍量のリードでアセンブリを行うかなどを検討し，利用可能であれば他のゲノムアセンブラを試すなどのトライアンドエラーが必要となる．またリードのクオリティがアセンブリ結果にも影響するため，クオリティのチェックも重要である．リードのクオリティのチェックは，FastQC（URL1-2-8, Andrews 2010）によりレポートの作成を行い，FASTX-Toolkit（URL1-2-9）などでクオリティの低い箇所を除去する操作を行う．サンガー法で得られた配列のアセンブリには，Phrap（URL1-2-10, Bastide *et al.* 2007）やCAP3（URL1-2-11, Huang *et al.* 1999）などがあり，有償ではCLC Main Workbenchなどがある．

2）トランスクリプトアセンブリ

全ゲノム解析は以前よりも身近にはなったが，研究対象が表現系に関係する場合，タンパク質をコードしている領域に絞って解析をすすめる場合も多く，その方法も開発された．対象を発現領域に絞ったアセンブリは，トランスクリプト *de novo* アセンブリと呼ばれ，mRNAから相補的DNA（complementary DNA, cDNA）を作成し，シークエンスとアセンブリを行う．使われるツールとしては，Trinity（URL1-2-12, Grabherr *et al.* 2011）やOases（URL1-2-13, Schulz *et al.* 2012），SOAPdenovoTrans（URL1-2-14, Xie *et al.* 2014）など，トランスクリプト用のアセンブリツールが用いられる．ゲノム用と異なるアセンブラを使う理由はスプライスバリアント（splice variant）による複数のトランスクリプトを考慮できるような工夫がされている点である．

アセンブリによりトランスクリプト作成後は，アノテーション（annotation）付けのための遺伝子予測や，トランスクリプトの発現量の目的でRNA-seqが行われる．順次これらについて述べていく．

3）遺伝子予測, アノテーション

遺伝子予測（gene prediction）とは塩基配列からゲノム配列上の遺伝子領

域・構造を予測することで，ドラフトゲノム作成時や，トランスクリプトのアセンブリを行った後などに行われる．予測では，繰り返し領域を特定することが最初のステップとなる．これはこの後の遺伝子予測のツールの検索対象からはずすためである．繰り返し領域の推定には，ナツメ（*Ziziphus jujuba*）のドラフトゲノムが発表された論文（Liu *et al.* 2014）ではRepbase（URL1-2-15, Jurka *et al.* 2005），Tandem repeats finder（URL1-2-16, Benson 1999），RepeatMasker（URL1-2-17）が使われている．

アノテーション付けは，公開されている生物種のデータベースを使いBLAST（URL1-1-5, Altschul *et al.* 1990）により，相同性検索を行いアノテーションを付ける方法と，新規で予測を行う*ab initio* gene predictionに分けられる．新規に遺伝し予測を行うツールでは，Augustus（URL1-2-18, Stanke *et al.* 2004）やGlimmerHMM（URL1-2-19, Majoros *et al.* 2004）がスイカ（*Citrullus lanatus*）やパパイヤ（*Carica papaya* L.）のアノテーション付けで用いられている（Guo *et al.* 2012; Ming *et al.* 2008）．また一連のアノテーション付けを行うパイプラインも提案されており，PASA（URL1-2-20, Haas 2003）やMAKER（URL1-2-21, Cantarel *et al.* 2008）などが公開されている．

その他，BLAST検索結果を利用して，Gene Ontology（GO）（Ashburner *et al.* 2000）の検索を行い機能予測をすることもよく行われる解析のひとつである．GOとは遺伝子の生物学的機能を共通辞書としてまとめたデータベースで，BLASTの検索結果とGOを紐付けることで，どのような機能を持っているか調べることが出来る．BLAST検索結果にGOを付加するツールとしては，BioBam社がBlast2GO（URL1-2-22, Conesa *et al.* 2008）を提供している．Blast2GOは，BLASTの検索結果に基づいてGOの情報をアノテーション付けし，機能分類を行えるツールである．Blast2GOには有償版のBlast2GO pro と無償版があるが，いずれもグラフィカルなインターフェースで利用でき，多くの論文でも活用されている（Hipp *et al.* 2014 ; He *et al.* 2013 ; Blanca *et al.* 2011）．

4）発現解析

発現解析（gene expression）では，マイクロアレイ（microarray）を利用する方法と，次世代シークエンサからのデータを利用するRNA-Seq法がある．次世代シークエンサのランニングコストの急速な低下に伴い，RNA-Seq法がマイクロアレイに置き換わる勢いで利用されている．

マイクロアレイの場合，データはマイクロアレイのチップやガラスの基板上に張り付けられたプローブと呼ばれる短いDNA断片に対して相補的に，mRNAから作成したcDNA断片をハイブリダイズしたものを蛍光スキャナーで読み取り，その蛍光量を発現量へ変換し，データ解析を行う（本書第1章第5節に詳述）．一方，次世代シークエンサでは，mRNAから作成されたcDNA断片を網羅的にシークエンサで読み発現量を得る．ゲノム既知，または利用できる近縁種のゲノムがある場合は，ゲノム配列へマッピングし，トランスクリプトごとにマップされたリードの数を発現量とする．ゲノム解読が行われておらず，近縁種のゲノムを利用できない場合は，前述に記載したアセンブリの方法でアセンブリを行い，アセンブリされた結果をトランスクリプトと仮定し，その配列へリードをマッピングし，マップされたリード数を発現量とする．

マイクロアレイの発現解析では，データ解析に利用できるツールは，統計解析のオープンソースツールであるR（URL1-1-23）上で利用できるBioconductor（URL1-1-24, Gentleman *et al.* 2004）を通して解析に必要なパッケージ群を利用することができる．詳細は東京大学 大学院農学生命科学研究科 門田幸二先生のウェブサイト（URL1-2-23）に使い方が具体例と共に公開されており，たいへん有用である．グラフィカルなインターフェースを持たせた商用ソフトウェアでは，GeneSpring（Agilent Technologies社），CLC Main Workbenchなどがある．

次世代シークエンサでの発現解析はmRNAからcDNAを作成し，シークエンスを行う．その後，リードをトランスクリプトへマッピングし発現量を計算する方法をRNA-seqといい，発現量の計算について，すでに解読されたゲノムが利用可能な場合と，そうでない場合について，図1-2-1に解析の流れを示し

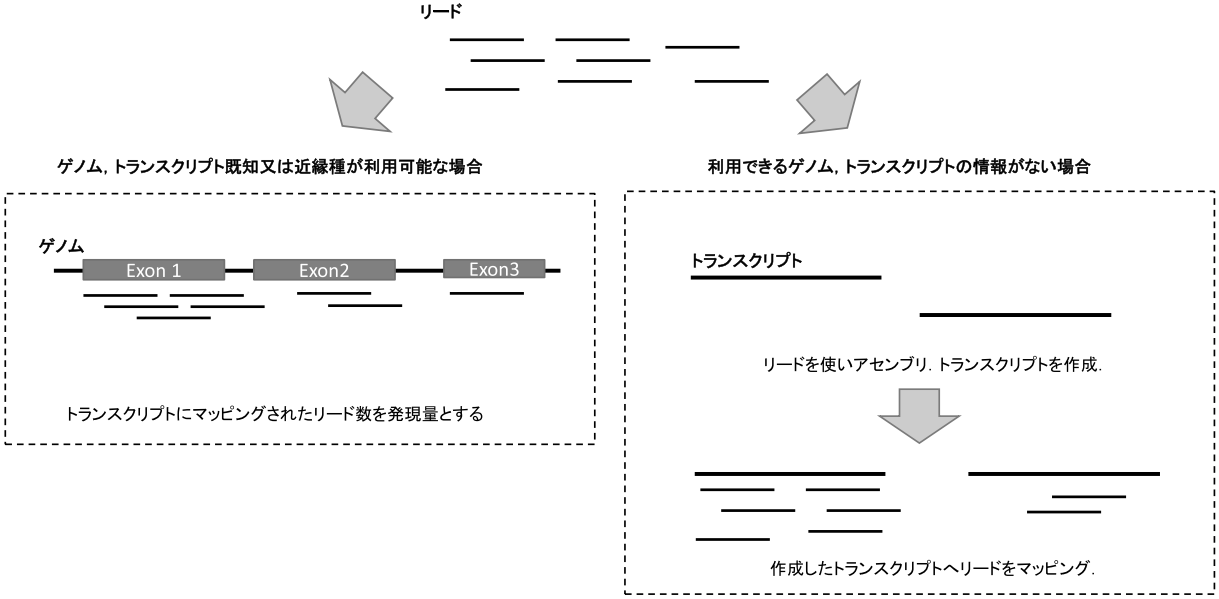

図1-2-1　ゲノム既知の場合の発現解析とゲノム未知の場合の発現解析の流れ

た．解読されたゲノムが利用できない場合は，前述のアセンブリ方法を使い，トランスクリプトを作成するというステップが必要となる．

発現差解析（differential gene expression）は，研究対象のサンプル群と，コントロールとなるサンプル群の比較により行われる．たとえばある薬剤を散布した群とそうでない群などである．それぞれのサンプルの発現量をRNA-seqにより計算した後，統計検定で群間の発現差をみることとなる．発現差解析を行うツールはBioconductorを通して利用できるedgeR（URL1-2-24, Robinson *et al.* 2012）や DESeq（URL1-2-25, Anders and Huber 2010 ; Love *et al.* 2014）がある．発現差解析のツールは比較論文が発表されてはいるが（Soneson and Delorenzi 2013 ; Nookaew *et al.* 2012 ; Kvam *et al.* 2012），自身のデータでポジティブ・コントロール，ネガティブ・コントロールを置き確認する事が重要である．次世代シークエンサ解析の発現解析においても，前述の門田先生のウェブサイトに使い方など詳細が具体例と共に記載されている．

5）変 異 検 出

変異検出（variant detection）とは，研究対象のサンプルを参照するゲノム配列と比較し，塩基配列の違いを検出する方法である．次世代シークエンサで使われる変異という単語はsingle nucleotide polymorphism（SNP）と混同されやすいが，SNPはある集団で1％以上の頻度で見られるものを指し，次世代シークエンサで見つかる一塩基変異はsingle nucleotide variant（SNV）といわれる．本章で記載する変異はこのSNVを指している．次世代シークエンサを使った変異の検出は，複数のステップで行われ，表1-2-2にその流れ，処理の目的，ツールを示した．ツールのURLについては表1－2－3に紹介した．大きな流れとしては，前処理，マッピング，変異検出，アノテーション付け・絞込みとなる．それぞれのステップごとに複数のツールを使い処理を行う．

表1-2-2　変異解析ステップ

ステップ	処理概要	処理詳細	処理目的	ツール名[z]	引用元
1	前処理	1.1 クオリティチェック	シークエンサエラーをどの程度含んでいるか，リード長の分布はどうなっているか，GC含有量などを確認する.	FastQC	
		1.2 トリミング	シークエンサエラーによる悪いクオリティを持った塩基の除去やシークエンスの際に付与したアダプターの除去	FASTX-Toolkit	
2	マッピング・変異検出	2.1 マッピング	リードを参照配列へアライメントする.	BWA	Li *et al*. 2009
		2.2 リアラインメント	マッピング後，より挿入や欠失を検出しやすいよう補正する.	GATK	McKenna *et al*. 2010
		2.3 重複除去	シークエンスされたリード中にPCRがかかりすぎた結果、同一配列を持つリードが大量に作成されることがある. そのような場合は，変異検出前に除去を行う.	Picard	
		2.3 変異検出	参照配列と異なる箇所を検出	GATK	
3	アノテーション付け・絞込み	3.1 アノテーション付け	遺伝子名やアミノ酸置換の有無を付加.	SnpEff	Cingolani *et al*. 2012
		3.2 絞込み	変異が見つかった領域のクオリティなどで絞込みを行う.	SnpEff	

[z]ツールについては代表的なものを記載. ツールのダウンロードURLは表1-2-3を参照.

表1-2-3　変異解析に利用されるオープンソースツール

ツール名	URL
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
FASTX-Toolkit	http://hannonlab.cshl.edu/fastx_toolkit/
BWA	http://bio-bwa.sourceforge.net/
GATK	https://www.broadinstitute.org/gatk/
Picard	http://sourceforge.net/projects/picard/
SnpEff	http://snpeff.sourceforge.net/

6）SNPジェノタイピング

SNPのマーカーを作成する場合，全ゲノム配列を特定する必要はなく，形式に関連したSNPを見つけることが目的となるため，シークエンスを行う領域を絞ることで迅速に効率よく行う方法が考案された．代表的なものとしては，制限酵素（restriction enzyme）を利用したrestriction site associated DNA Sequence（RAD-seq, Davey *et al.* 2010）とgenotyping-by-sequence（GBS）（Elshire *et al.* 2011）が挙げられる．RAD-seqでは制限酵素により切断された末端の一部を読み，そこから変異を探索し，異なる系統間での比較を行うことで形質に関連した変異を見つける方法である．RAD-seqでは，ゲノムが未知の場合にも利用でき，制限酵素により切断された末端から一部のみのシークエンスを行うため，シークエンスの量を減らす事ができ，マップされるリード数，カバレッジ（coverage）を高くすることが出来るため，コストを減らしながらも精度の向上も期待できる有用な方法として注目されている（Hipp *et al.* 2014; Zhou *et al.* 2015）．RAD-seqの解析ツールとしては，Stacks（URL1-2-26, Catchen *et al.* 2011）がある．

7）SNPの応用

SNPマーカーが作成できると，それを活用した応用が行える．量的形質座（Quantitative Trail Locus, QTL）との関連を見つけるためのQTLマッピングや，より高解像度に行えるGenome Wide Association Study（GWAS），予測を目的としたGenomics Selection（GS）も近年果樹への応用が行われている（Kumar *et al.* 2013 ; Kumar *et al.* 2012）．GWASでは形質とSNPの関連をさぐり，どのSNPが有用な形質に関連しているかを解析するのに対し，GSでは，どのSNPが有用か，という点よりも，どれが有用な形質を引き継いだ個体かを予測することに焦点を当てている．GWAS解析では，PLINK（URL1-2-27, Purcell *et al.* 2007）が最も有名で，アレル頻度の計算や統計検定などのツールのセットとなっている．GSは機械学習を背景にした方法であるが，これまでの解析手法で紹介したような広く使われているツールが存在するわけではなく，統計の知識を持ったバイオインフォマティシャンとの共同研究にて活用されることが望ましいと考える．

8）統合解析ツールとビューア

ひとつの解析を完了するには，複数のツールを組み合わせて行う事が多く，さらにサンプルが多検体となる場合は，解析を何度も繰り返し行うこととなる．またそれぞれの解析ツールも設定が決まれば同じ操作を繰り返すだけの作業となるため，簡便に間違いなく実施できるよう，解析ツールをまとめて実行できるツールが用意されており統合解析ツールとよばれ，ここでいくつか紹介する．その 1 つ目は，国立遺伝学研究所が提供しているDDBJ Read Annotation Pipeline（URL1-1-27, Nagasaki *et al.* 2013）が統合解析環境にあたる．これはインターネットを経由してデータを転送し，ウェブブラウザ により解析をおこなうサイトであり，本章で紹介した *de novo* アセンブリや，変異解析，RNA-seq解析に必要なツール群が実行できる．コマンドライン での解析操作が難しく，うまく活用できない場合や，解析に必要なメモリ，CPUの大き

なコンピュータが利用できない場合に有用である．

2つ目はGalaxy（URL1-1-25, Goecks *et al.* 2010；Blankenberg *et al.* 2010；Giardine *et al.* 2005）であり，こちらもウェブベースでの統合解析環境を提供しているサイトで，DDBJ Read Annotation Pipeline同様に変異解析やRNA-seqなどに必要なツールが利用可能である．GalaxyはWeb経由でデータを外部に送ることが難しい場合や、インターネットの速度が遅く、ウェブへのアップロードに時間がかかる場合などに備えて，研究者の施設内でも解析環境を構築することが可能となっている．

商用ソフトウェアの多くは，統合解析環境を提供しており，1つのソフトウェアで複数の処理が行えるようになっている．商用ソフトウェアとしては，CLC Genomics Workbench, StrandNGS（Strand Genomics社）などがある．利用可能なツール内容については，製品により異なるため確認が必要である．

統合解析ツールを使う場合は，出力されたファイルの内容を閲覧するためのビューア（viewer）が合わせて利用可能な場合が多いが，前述のコマンドラインツールなどを使う場合，閲覧用のビューアを別に用意する必要がある．次世代シークエンサの解析では，特にリードをマッピングした後のビューアが，閲覧したい結果のひとつとなる．このようなビューアとしては，Integrative Genomics Viewer（IGV，URL1-2-28, Thorvaldsdóttir *et al.* 2013）が汎用され無償で利用可能である．

9）配列解析ツールとスクリプト言語

バイオインフォマティクスは現在では広範な解析範囲を包括しており，前述のような次世代シークエンス解析以外にも，たとえばDNA配列の相補的な配列を得たり，タンパク質の翻訳配列を得るなどといった配列解析が必要となる場合があり，そのための解析ツールも用意されている．

配列解析に利用できるツールでは，European Bioinformatics Institute（EBI）から提供されているEMBOSS（URL1-2-29, Rice *et al.* 2000）が150を超える解析ツールを提供している．解析ツールの例としては，GC含有量（GC-content）

の計算からBLAST，アラインメントツールなど多岐にわたる．EMBOSSはWindows，Mac，Linuxのいずれでも利用可能である．配列解析はEMBOSS以外にも，BioPerl（URL1－2－30, Stajich *et al.* 2002），BioRuby（URL1－2－31, Goto *et al.* 2010），BioPython（URL1-2-32, Cock *et al.* 2009）を使って解析することが可能である．それぞれPerl, Ruby, Pythonといったスクリプト言語をベースとしている．

BioPerlは最も歴史が深く，非常に膨大なモジュールを持っており資料も豊富である．BioRubyは，日本で開発されたRubyを元にされており，BioRubyの開発にも日本人の研究者が携わっていることから，日本語での資料が比較的多く用意されてる．Ruby自体も非常にわかりやすい言語である．Pythonも非常にわかりやすい言語であるが日本語の資料は若干少な目かもしれない．これらのスクリプト言語は本人の好みに依存するところも多く，いろいろ試しながら自分が使いやすいものをまずは試してみるとよいと思われる．

引用文献

Altschul, S. F. *et al.*（1990）Basic local alignment search tool. Journal of Molecular Biology. 215：403－410.

Anders, S. *et al.*（2010）Differential expression analysis for sequence count data. Genome Biology. 11：R106.

Andrews, S.（2010）Babraham Bioinformatics－FastQC：A Quality Control tool for High Throughput Sequence Data.

Ashburner, M. *et al.*（2000）Gene Ontology：tool for the unification of biology. Nature Genetics. 25：25－29.

Bastide, M. *et al.*（2007）Assembling genomic DNA sequences with PHRAP. Current Protocols in Bioinformatics. 11－14.

Benson, G.（1999）Tandem repeats finder：A program to analyze DNA sequences. Nucleic Acids Research. 27：573－580.

Blanca, J. M. *et al.*（2011）Melon Transcriptome Characterization：Simple Sequence Repeats and Single Nucleotide Polymorphisms Discovery for High

Throughput Genotyping across the Species. The Plant Genome Journal. 4：118.

Blankenberg, D. *et al.*（2010）Galaxy：A Web-Based Genome Analysis Tool for Experimentalists. Current Protocol Molecular Biology. Chapter 19：Unit 19. 10. 1-21.

Cantarel, B. L. *et al.*（2008）MAKER：An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. 18：188-196.

Catchen, J. M. *et al.*（2011）Stacks：Building and Genotyping Loci *De Novo* From Short-Read Sequences. G. 1：171-182.

Chin, C. *et al.*（2013）Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods. 10：563-569.

Cock, P. J. *et al.*（2009）Biopython：Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 25：1422-1423.

Conesa, A. *et al.*（2008）Blast2GO：A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics. 1-12.

Davey, J. L. *et al.*（2010）RADseq：Next-generation population genetics. Brief Functional Genomics. 9：416-423.

Eid, J. *et al.*（2009）Real-time DNA sequencing from single polymerase molecules. Science. 323：133-138.

Elshire, R. J. *et al.*（2011）A robust, simple genotyping-by-sequencing（GBS）approach for high diversity species. PLoS ONE. 6：1-10.

Gentleman, R. C. *et al.*（2004）Bioconductor：open software development for computational biology and bioinformatics. Genome Biology. 5：R80.

Giardine, B. *et al.*（2005）Galaxy：a platform for interactive large-scale genome analysis. Genome Research. 15：1451-1455.

Gnerre, S. *et al.*（2011）High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America. 108：1513-1518.

Goecks, J. *et al.*（2010）Galaxy：a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology. 11：R86.

Goto, N. *et al.*（2010）BioRuby：Bioinformatics software for the Ruby programming language. Bioinformatics. 26：2617-2619.

Grabherr, M. G. *et al.*（2011）Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 29：644-52.

Guo, S. *et al*. (2012) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nature genetics. 45:51-58.

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

He, N. *et al*. (2013) Draft genome sequence of the mulberry tree Morus notabilis. Nature Communications. 4:2445.

Hipp, A. L. *et al*. (2014) A framework phylogeny of the American oak clade based on sequenced RAD data. PLoS ONE. 9.

Stajich, E. *et al*. (2002) The Bioperl Toolkit : Perl Modules for the Life Sciences. Genome Research. 12:1611-1618.

Jünemann, S. *et al*. (2014) GABenchToB : A Genome Assembly Benchmark Tuned on Bacteria and Benchtop Sequencers. PLoS One. 9:e107014.

Jurka, J. *et al*. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research. 110:462-467.

Kajitani, R. *et al*. (2014) Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research. 24:1384-1395.

Kumar, S. *et al*. (2012) Genomic selection for fruit quality traits in apple (*Malus*× *domestica* Borkh.). PLoS ONE. 7:1-10.

Kumar, S. *et al*. (2013) Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics. 14:393.

Kvam, V. M. *et al*. (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American journal of botany. 99:248-56.

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

Li, R. *et al*. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research. 20:265-272.

Love, M. I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv. 1-21.

Luo, R. *et al*. (2012) SOAPdenovo2 : an empirically improved memory-efficient short-read *de novo* assembler. GigaScience. 1:18.

Majoros, W. H. *et al*. (2004) TigrScan and GlimmerHMM : Two open source *ab initio* eukaryotic gene-finders. Bioinformatics. 20:2878-2879.

McKenna, A. *et al*. (2010) The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 20 :

1297-303.

Ming, R. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature. 452 (7190):991-996.

Nagasaki, H. *et al.* (2013) DDBJ read annotation pipeline: A cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Research. 20:383-390.

Nookaew, I. *et al.* (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic acids research. 40:10084-10097.

Purcell, S. *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics. 81:559-575.

Rice, P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genetics. 16:276-277.

Robinson, M. *et al.* (2012) edgeR: differential expression analysis of digital gene expression data User's Guide.

Schulz, M. H. *et al.* (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 28:1086-1092.

Soneson, C. *et al.* (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 14:91.

Stanke, M. *et al.* (2004) AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Research. 32:309-312.

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

Velasco, R. *et al.* (2007) A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS One. 2:e1326.

Xie, Y. *et al.* (2014) SOAPdenovo-Trans: *De novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics. 30:1-7.

Zerbino, D. R. *et al.* (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research. 18:821-829.

Zheng, W. *et al.* (2013) High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. Nature Communications. 4:1-10.

Zhou, L. *et al*. (2015) Identification of domestication-related loci associated with flowering time and seed size in soybean with the RAD-seq genotyping method. Science Report. 5:9350.

Vinson, J. P. *et al*. (2005) Assembly of polymorphic genomes : Algorithms and application to Ciona savignyi. Genome Research. 15:1127-1135.