Chapter 2
Functional Genomics

Hoe-Han Goh, Chyan Leong Ng, and Kok-Keong Loke

Abstract Functional genomics encompasses diverse disciplines in molecular biology and bioinformatics to comprehend the blueprint, regulation, and expression of genetic elements that define the physiology of an organism. The deluge of sequencing data in the postgenomics era has demanded the involvement of computer scientists and mathematicians to create algorithms, analytical software, and databases for the storage, curation, and analysis of biological big data. In this chapter, we discuss on the concept of functional genomics in the context of systems biology and provide examples of its application in human genetic disease studies, molecular crop improvement, and metagenomics for antibiotic discovery. An overview of transcriptomics workflow and experimental considerations is also introduced. Lastly, we present an in-house case study of transcriptomics analysis of an aromatic herbal plant to understand the effect of elicitation on the biosynthesis of volatile organic compounds.

Keywords Crop genomics · Genomic medicine · Metagenomics · Pharmacogenomics · RNA-Seq · Sequencing · Transcriptomics

2.1 Introduction

Functional genomics is a field of molecular biology that integrates genomic and transcriptomic data to describe gene (and protein) functions and interactions. Genomics is a study of the function and structure of genome, which comprise the complete set of all genes, regulatory sequences, and non-coding regions within an organism's DNA. This discipline in genetics relies on sequencing and bioinformatics approach to sequence, assemble, and analyse all the gene coding and non-coding

H.-H. Goh (✉) · C. L. Ng · K.-K. Loke
Institute of Systems Biology, Universiti Kebangsaan Malaysia (UKM), Bangi, Malaysia
e-mail: gohhh@ukm.edu.my; clng@ukm.edu.my

W. M. Aizat et al. (eds.), *Omics Applications for Systems Biology*,
Advances in Experimental Medicine and Biology 1102,
https://doi.org/10.1007/978-3-319-98758-3_2

sequences and how these genetic components interact to produce an organism and all its functions. Conversely, transcriptomics is the study of the transcriptome—the complete set of RNA transcripts (including mRNA, rRNA, tRNA, and other non-coding RNA) that are produced by the genome, under specific circumstances or in a specific cell—using high-throughput methods. Sometimes, genomics is used as an umbrella term that encompasses genome-wide studies in many subdisciplines, including transcriptomics, proteomics, metabolomics, bioinformatics, systems biology, and synthetic biology. Hence, genomics provides not only a suite of methods and analytical techniques but also a perspective to study an organism as a whole.

On the other hand, functional genomics focus on the dynamic regulation of gene expression and protein-protein interactions, to elucidate DNA function at the levels of genes, transcripts, and proteins in a genome-wide context. The term "genomics" was first coined in 1986 by geneticist Tom Roderick during a meeting on the mapping of human genome, 66 years after the word "genome" was used by the German botanist Hans Winkler [1]. In the year 2000, genome sequences of the first model flowering plant *Arabidopsis thaliana* [2] and insect fruit fly *Drosophila melanogaster* [3] were published. The year after, two independent draft human genome sequences were reported 1 day after another in Feb 2001 [4, 5], followed by the mouse genome in 2002 [6]. The International Human Genome Project initiated in 1990 was officially completed in April 2003, 5 years after sequencing started in 1998, 2 years ahead of schedule. This provides a reference human genome with composite representative sequence derived from several selected individuals of nearly 100 anonymous donors. Since then, more genomes from individuals of different nations were sequenced [7]. The field has advanced to the cataloguing of regulatory elements and epigenomic mapping. These efforts, like genome sequencing, are not done by individual labs but continued as international collaborations, i.e. the ENCODE Consortium [8] and the Roadmap Epigenomics Mapping Consortium [9].

These genome projects not only drove the emergence of new methods for genome-wide investigations but also provided a framework for global views of biology through the advent of sequencing techniques [10]. This propelled myriads of genome sequencing projects of non-model organisms, including the "Genome 10K Project" for vertebrates [11]. Cancer (epi)genomics is another ongoing research hotpot [12]. This is made possible by a parallel development in bioinformatics tools and resources, such as GO tools for the unification of biology [13] and KEGG [14]. High-throughput data generation demands development in large-scale statistical analyses, such as that for genome-wide association studies (GWAS) [15], while clustering has become an integral tool to partition a large dataset into more easily digestible conceptual pieces [16]. Furthermore, visualisation of genome data is of paramount to comprehend emerging patterns [17], such as Integrative Genomics Viewer [18] and Circos [19]. Therefore, genomic transformation of biology into a data-intensive field has recruited many engineers, physicists, mathematicians, and computer scientists into biological research.

2.1.1 Different Aspects in Genomic Research

Current field of genomic research has bloomed into diverse scopes from molecular, cellular systems to population level. Molecular genomics like structural genomics [20], glycogenomics [21], toxicogenomics [22], chemogenomics [23], and pharmacogenomics [24] study particular genomic characteristics focused on molecular biology aspects. Cellular genomics like single-cell genomics [25] investigates cellular behaviour in the context of genomics content. Higher level of research scope encompasses complex interactions between multiple genomics such as epigenomics [12], metagenomics [26], comparative genomics [27], phylogenomics [28], GWAS [29], and translational research of genomic medicine [30]. Therefore, genomic research has span across the continuum of basic and applied research, which can be classified into comparative, functional, and translational genomics (Fig. 2.1).

Transcriptomics and proteomics are key parts of functional genomics. These varied genomic platforms allow researchers to address global, general, and specific questions in biology with respect to the genome under study. For example, GWAS have revealed variations in human genome with numerous single nucleotide polymorphisms (SNPs) that are linked to disease risk [31].

2.1.2 Functional Genomics in the Context of Systems Biology

The advancement of genomics provided a critical boost to systems biology by facilitating the prediction of complex systems' behaviours, properties, and active processes. For example, the human genomic network permits the gene prediction of the

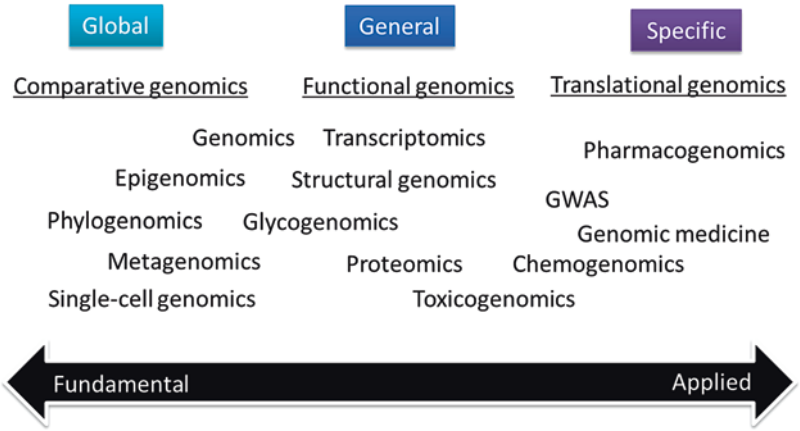

Fig. 2.1 A continuum of diverse research fields in genomics in addressing different levels of biological questions

best drug targets and guides the design of new therapies in treating complex diseases [24]. The ultimate aim is to produce more predictive, preventive, personalised, and participatory (P4) medicine for everyone (further described in Sect. 2.2.1).

Genomics in the broadest sense include both structural and functional aspects. The genome assembly and read mapping are considered structural, whereas the analyses of read abundance and exon usage are functional. These different aspects raise the question "to what extent genomic analyses qualify as systems biology?" [32]. For example, genome assembly is critical to genomic analysis with some of the greatest algorithmic challenges, but the resulting assembly on its own provides little direct insight about the biological system without further analysis.

To address this question, we apply the definition of systems biology as the study of interactions between system parts which involves (i) experimental perturbation, (ii) quantitative measurement, (iii) data integration, and (iv) modelling [33]. For instance, while genome survey solely for genome size and heterozygosity estimation would not fall within the realm of systems biology, the comparative analysis of genome sequences from different cancer cell types to study the genetic variations would qualify. The identification of mutations causing specific cancers represents a systems approach of finding one part in the system which affects the whole system's behaviour.

Reference genome assembly and annotation, as well as comparative genomics, do not shed light on system behaviour on its own, but serve as blueprints for systems-level analysis, such as gene regulatory network (GRN) inference. Functional genomics to study the genome "in action", such as tissue-specific gene expression and the dynamics of transcriptional regulation, are generally within the systems biology framework. Lastly, various genome-wide experiments involving chromatin immunoprecipitation-sequencing (ChIP-seq) interactomics, RNA-seq differentially expressed gene (DEG) analysis, and populational genome variation analysis also fall under the operational definition of systems biology. These analyses typically associate called peaks, expression levels, or variants of specific genes to infer functional enrichment in pathways. Figure 2.2 illustrates how genomics can fit into the context of systems biology through integration with other "omics" platforms.

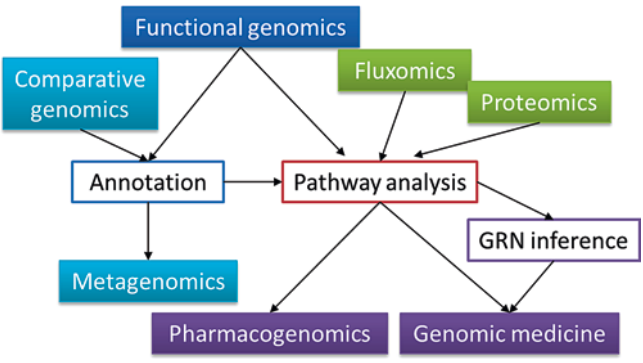

Fig. 2.2 Genomics in the context of systems biology

The integration of genomic data annotation, through functional and comparative genomic analyses, with that of proteomics and fluxomics allows systems-level pathway analysis, which helps in GRN inference. This contributes towards model development in pharmacogenomics and genomic medicine to identify drug targets. Metagenomics expands the study system beyond a single organism towards a community understanding of associated microbes at a functional level.

2.2 Applications of Functional Genomics

Over the past decades, genome sequencing technology has evolved from the first-generation Maxam-Gilbert and Sanger sequencing to current next-generation sequencing (NGS) methods of sequencing by synthesis and single-molecule real-time sequencing [34]. Genome data analyses, such as sequence mapping, assembly, genome annotation, and pathway mapping, have also undergone great advances with the advent of supercomputers, database development, and bioinformatics tools. While genome sequences only provide one-dimensional view on genetic compendium of a cell, when combined with systems biology, it can provide multidimensional insights and understanding on the dynamics of biological processes. In this section, example genomic applications in the studies of human, plants, and microbes are presented.

2.2.1 Comparative Genomics in the Study of Human Genetic Variation

A genome library provides an overview of the genetic makeup of a single organism. Comparison of multiple genome libraries from a single species provides insights into the genetic variation. The 1000 Genomes Project Consortium revealed the genetic variation that encompassed 26 human population and 2,504 individuals throughout five continents [7]. More than 88 million different variants were found, with approximately 96% of them represent SNPs followed by short insertions and deletions (indels) and structural variants. The comparative genomics analysis also reported that every individual harbours four to five million genetic variant sites with more than 99% of them are SNPs.

With the accessibility of genome sequencing, various diseases caused by single-gene mutations (Mendelian or monogenic diseases), such as cystic fibrosis, fragile X syndrome, and Huntington disease, can be easily identified within the genome of individuals or families with risk of inheritance. It is known that many genetic diseases are caused by single nucleotide variants that affect protein function through amino acid substitution [35]. By genome analysis, one can also identify SNPs with indirect association to the disease phenotype and important in the disease develop-

ment. For instance, many SNPs in the non-coding regions are known to play regulatory roles in gene transcription and expression [36]. Some SNPs found in transcriptional elements have been identified to be associated with β-thalassemia [37], tumour formation [38], melanoma [39], and retinal vasculature defects [40]. Genome analysis has also helped to identify loci which are responsible for disease susceptibility and severity, such as for type 2 diabetes, coronary heart disease, systemic lupus erythematosus, hypertriglyceridemia, or even infectious diseases like trypanosomiasis, malaria, and Lassa fever [41, 42].

This comparative genomics has extended our understanding on human population history at the molecular level and helps in linking disease phenotypes with genetic variants. In the foreseeable future, more comprehensive genomics data and analysis will be available to provide guidance in disease prevention, diagnosis, and treatment according to the personal genetic profile, hence moving us towards P4 medicine era.

2.2.2 Plant Functional Genomics for Crop Improvement

Food security is one of the biggest challenges in this century as the current 7.3 billion world population is projected to reach 9.7 billion by 2050 (UN DESA Report 2015). Along with the impact of climate change and water scarcity, crop productions need urgent improvement with new technologies to overcome upcoming challenges. One of such technologies is to apply functional genomics in place of time-consuming and laborious traditional plant breeding.

Since the fully sequenced genome of model plant *Arabidopsis thaliana* [2], more than 100 plant genome sequences are now available, especially important crop plants such as rice [43, 44], maize [45], soybean [46], potato [47], and bread wheat [48]. Other important commodity plants such as African oil palm [49] and rubber [50] are also sequenced. The availability of crop genome information allowed the identification of genes related to important traits, including yield, disease resistance, and stress tolerance. Crop breeders are now able to accelerate hybrid-breeding programme via marker-assisted selection with genotyping-by-sequencing to produce higher-quality crops [51].

One of the recent examples on how genomic studies can benefit the plantation industry is the oil palm genome study of *Elaeis guineensis* and *Elaeis oleifera* [49] with the identification of *MANTLED* locus responsible for the mantled phenotype through epigenome-wide association studies [52]. The methylation of *Karma* long interspersed nuclear element (LINE) retrotransposon was found to be associated with clones of normal fruit yield compared to mantled clones with hypomethylation [52]. It is therefore useful for screening somaclonal epigenetic alterations during in vitro cloning to cull mantling at plantlet stage to prevent commercial and land use losses.

Apart from the development of molecular markers for the selection of superior traits, precision genome engineering is now possible to improve crops using genome

editing tools such as transcription activator-like effector nuclease (TALEN) and CRISPR/Cas9 system [53]. Genetically modified crops with disease and pest resistance as well as higher yields could become more acceptable with the recent genome editing techniques to alter specific gene in a more precise manner without introducing foreign DNA. In the year 2014, hexaploid bread wheat with resistance to powdery mildew was generated by simultaneously introducing three targeted mutations into homoeoalleles of mildew resistance locus o (Mlo) using both TALEN and CRISPR/Cas9 technologies [54]. Genome editing with CRISPR/Cas9 is now possible in major crops such as sorghum, rice, maize, and soybean [55–58]. Despite some current technical challenges like the ineffective delivery method [59], genome editing tools which are relatively cheap and easy to apply will revolutionise crop improvement.

2.2.3 Metagenomics: A New Approach for Antibiotic Discovery

The discovery of antibiotic penicillin by Alexander Fleming in 1928 from *Penicillium rubens* mould saved millions of lives in fighting bacterial infection. Since then various kinds of antibiotics were discovered from bacteria during the prosperous age of antibiotic discovery (1940–1990), primarily from cultured bacteria and sensitivity testing with candidate compounds. A new class of antibiotic is hard to find nowadays in current bacterial collections. For instance, the discovery rate for new antibiotics in actinomycetes is only ~1%, which includes the genus *Streptomyces* that contributed to over 80% of antibiotics discovered so far [60]. Furthermore, the pressing need for new antibiotics is compounded by emerging multiresistant pathogenic bacteria.

Metagenomics has revolutionised microbiology in uncultured microflora [26]. It is estimated that 99% of bacterial species are not yet cultured in the laboratory, [61] and 70% of prokaryotic phyla that exist in seawater, freshwater, and soil are unexplored [62, 63]. While studying these uncultured bacteria remains a big challenge, metagenomics has made it possible to decode genomes for not one but a community of bacteria. This genome information opens up a new source for searching novel bioactive metabolite candidates as antibiotics or drugs. For the past two decades, metagenomics approaches have successfully identified novel antibiotics or new derivatives of known antibiotics (Table 2.1). For example, a new antibiotic teixobactin that could inhibit cell wall synthesis without causing resistance in pathogenic bacteria was discovered [64].

More importantly, many antibiotic resistance-related genes and pathways were also discovered through metagenomics studies [71]. Metagenomics also revealed events of antibiotic resistance gene exchange between environmental bacteria and clinical pathogens [72]. Moreover, single-cell and metagenomics studies coupled with bioinformatics, metabolomics, and chemical analysis by Wilson and colleagues discovered unique bioactive chemical compounds (polyketides and peptides) from two uncultivated phylotypes of marine sponges [73]. These findings encourage further exploitation of uncultivated bacteria for drug discovery via systems biology approach.

Table 2.1 New antibiotics or known antibiotic derivatives identified from metagenomics approaches

Antibiotics	Function	Reference
Indirubin	Antimicrobial activity	MacNeil et al. [65]
Turbomycin A and B	Broad-spectrum antibiotic activity against Gram-negative and Gram-positive bacteria	Gillespie et al. [66]
Palmitoylputrescine	Antibacterial activity against *Bacillus subtilis*	Brady and Clardy [67]
Teixobactin	Inhibit bacterial cell wall synthesis	Ling et al. [64]
Antimicrobial peptides buwchitin	Antibacterial activity against *Enterococcus faecalis*	Oyama et al. [68]
Modified chloramphenicol (*Cm*) derivatives 1-Acetyl-3-propanoylchloramphenicol 1-Acetyl-3-butanoylchloramphenicol 3-Butanoyl-1-propanoylchloramphenicol	Inhibitors for methicillin-resistant *Staphylococcus aureus* (MRSA)	Nasrin et al. [69]
Malacidins	Calcium-dependent antibiotics against Gram-positive bacteria	Hover et al. [70]

2.3 Transcriptomics Workflow

Here we describe main aspects of general workflow for transcriptomics analysis based on RNA-sequencing (RNA-seq) with focus on protein-coding mRNA analysis, which include experimental considerations, sequencing approaches, and data analysis. Comprehensive descriptions are beyond the scope of this chapter; readers are recommended to refer to other recent literature for further understanding on other aspects of transcriptomics [74], such as small RNA-seq, degradome-seq [75], translatome-seq [76], targeted RNA-seq [77], single-cell RNA-seq [78], and epi-transcriptomics [79]. News on the latest development in the field of transcriptomics can be obtained by following the RNA-Seq Blog (https://www.rna-seqblog.com/).

2.3.1 Experimental Considerations

Various aspects need to be considered when designing a transcriptomic experiment, some of which are listed in Table 2.2. The first and the most critical aspect is the experimental design in addressing a specific biological question, which not only determine the strategy of all the downstream transcriptomic analyses but also key to the validity of the experimental results. This includes the purpose of the study on whether it is for expression/differential expression study, study of alternative splicing events (gene isoforms), or discovery of novel transcripts.

Table 2.2 Different aspects of transcriptomic experimental considerations

Aspect	Considerations
Experimental design	Purpose of study Expression and differential expression Isoform discovery and alternative expression Novel transcripts (coding/non-coding) Sample acquisition: purity, quantity, quality, storage Number of replicates (biological and technical)
RNA isolation	Conventional methods or kits Qualitative and quantitative measurements QC (gel electrophoresis, RIN) DNase treatment RNA fragmentation
Library construction	Enrichment: Total RNA, polyA+, polyA-, or ribo-depletion library Size selection (before and/or after cDNA synthesis) Small RNAs (microRNAs)? cDNA fragmentation A narrow fragment size distribution or a broad one Amplification 5′ or 3′ mRNA tags Adapter/index barcoding/multiplexing Stranded or non-stranded library Exome captured or un-captured Library normalisation
Sequencing	Sequencing platform to use: cost, accuracy, read length, time, throughput, applications Depth of sequencing Single-end or paired-end sequencing (Illumina) Spike-ins (ERCC)
Analysis	Raw read preprocessing: de-multiplexing, adapter, sequence quality, trimming Sample QC: correlation analysis/PCA Referenced or de novo pipeline Read alignment: mapping, assembly, or both Level of quantification: transcript/isoform, gene, exon Quantification measure: count, RPKM/FPKM, TPM Differential expression analysis: parametric or non-parametric Alternative splicing analysis: splicing event, isoform expression Functional analysis: annotation (NR, Swiss-Prot, Pfam, GO, COG, etc.), overrepresentation/enrichment analysis (GSEA), pathway analysis (IPA, KEGG, etc.) Visualisation: genome browser, sashimi plot, splice graph etc. Integration: pathway mapping, multi-omics, *etc*.

Different purposes require different strategies of RNA-seq. For example, differential expression analysis of gene with high transcript abundance will not require as high sequencing depth as for the analysis of gene expression with very low abundance and the same applied to the profiling of common and rare transcripts. This also influences the number of biological replicates required for the necessary statistical power to achieve the targeted significance level of differentially expressed genes (DEGs). In general, more biological replicates at lower sequencing depth are

better than fewer samples at higher depth with equivalent total number of reads [80]. Pseudo-replication of technical replicates should be avoided to identify true DEGs with biological significance. Furthermore, power analysis is recommended to estimate the minimum number of biological replicates under the budget constraint of RNA-seq experiments [81].

In most cases, obtaining sufficient amount of good quality RNA from limited samples can be a bottleneck for increasing the number of biological replicates. This is related to the second aspect of consideration which is RNA isolation in deciding the most suitable method of extracting high-quality (RIN > 8) RNA for cDNA library preparation. Which library preparation method to choose will depend on the purpose of study on whether the target is only the protein-coding RNA (mRNA), non-coding RNA, small RNA (size selection), or total RNA (no selection). Furthermore, conventional mRNA-seq of whole fragmented transcript can be replaced by tag-based sequencing [82] such as 5′ or 3′ end sequencing if only interested in the expression of annotated genes [83]. This also improves gene quantification with higher sensitivity for rare transcript without the need of gene length normalisation during downstream analysis as tagged-based sequencing avoids the situation of longer transcripts crowding out shorter transcripts at low abundance. Most of the cDNA library kit is currently strand-specific to provide a more accurate estimate of transcript expression and genome annotation.

The choice of sequencing platform will be described in the next section, but briefly, it will depend on the purpose of study with considerations on the cost, accuracy, read length, required throughput (depth), and application. For example, Ion Torrent will be suited for transcript variant analysis of model organisms but not suitable for de novo transcriptome profiling. Lastly, downstream analyses after gathering all the sequencing data will be major considerations depending on the biological questions to be answered. In general, these include preprocessing of raw reads by trimming poor reads, QC, and examination of consistency among the biological replicates through correlation or multivariate analysis such as principal component analysis (PCA). This will require some informed judgement on which is the most suitable software/tools or even version to choose based on intended goal. The rule of thumb is to use the most established and up-to-date version of software/tools relevant to the topics of study as described in the latest literature. It is important to report the version of software/tools and database used throughout the data analysis for reproducibility as different versions might influence the outcome of result.

General workflow of transcript reconstruction will be described in Sect. 2.3.2. For more details on transcriptomic analysis and experimental considerations, we can also refer to a good website, RNA-seqlopedia (https://rnaseq.uoregon.edu/), or a recent review [84]. There are many specialised analysis software/tools, either proprietary such as Ingenuity Pathway Analysis (IPA) or open-source Web Gene Ontology Annotation Plot (WEGO), which are developed for various downstream analyses that will require readers to be aware about the latest development in the field as many intensively used software/tools are regularly updated. However, the most up-to-date version is not necessarily the best option; it is therefore important to understand the detailed functions of a software/tool and changes made by the

updates. Older versions are generally still available for download if required. Many of these software/tools are available on GitHub, such as Trinity, with detailed documentation. There is currently a trend towards DockerHub, which contains Trinity and all dependent software used for downstream analyses within the Trinity framework. This will allow easy implementation of analysis pipeline in the server or for cloud computing. The reproducibility of analysis will be improved by the recent efforts in moving transcriptomics analysis towards a customisable automated pipeline in cloud computing [85–87].

2.3.2 Data Acquisition

Nowadays, transcriptomic analysis is largely dependent on data generated from RNA-sequencing technology, especially for non-model organisms. However, many studies on model organisms such as human and rice still apply the established Affymetrix microarray approach [88] which will not be covered in this section. Some of the current sequencing platforms used in RNA-seq is summarised in Table 2.3. Roche 454 sequencing platform is not included due to the termination of its development for application in the field of transcriptomics. In general, there are two categories of RNA-seq platforms, which are short-read sequencers such as Illumina and Ion Torrent with generally <400 bp and long-read sequencers such as

Table 2.3 Summary on the different state-of-the-art sequencing platforms for transcriptomics

Platform	Read statistics	Description	Model	Advantages
Illumina	25–300 bp (SE/PE) ~7 h–6 days ~0.6–3000 Gb	Sequencing by synthesis, fluorescence	MiSeq NextSeq HiSeq Novaseq	High-throughput deep sequencing for gene quantification and de novo analysis
Ion Torrent	200–400 bp (SE) 2–4.4 h ~0.03–10 Gb	Sequencing by synthesis, proton release	Proton PGM	Rapid sequencing for multiplex-PCR products, especially for mutation analysis and variant detection
Pacific Biosciences	>40 kb (SE or circular consensus) 3 h ~0.5–10 Gb	Single-molecule sequencing by synthesis, real-time fluorescence	RS II Sequel	Long read for full-length transcript sequencing and isoform detection
Oxford Nanopore	Variable depends on library (1D/2D reads), ~10–950 kb ~6 h to few days ~5 Gb	Single-molecule sequencing by synthesis, real-time electrical current	MinION GridION	Long read with minimal library preparation if high quality; direct RNA-seq is possible

Read throughput, run time, and yield per run are based on minimum and maximum values from different models. *PE* paired-end read, *SE* single-end read

Pacific Biosciences (PacBio) and Oxford Nanopore with >10 kb of reads in real time.

The choice of platform will depend on the purpose of experimental study as described above. It is now possible to generate a full-length high-quality transcriptome reference with PacBio isoform sequencing (Iso-Seq) [89] with unprecedented confidence in the identification of novel transcripts and allele-specific gene expression. However, Iso-Seq is still limited by relatively high cost and lower throughput of sufficient read depth for statistical gene expression analysis. On the other hand, if the study is only concerning with the expression of known genes in annotated genome or organism with high-quality transcriptome, 3′ mRNA-sequencing approach that generates only one fragment per transcript such as QuantSeq [90] can greatly increase the depth of sequencing with more sample multiplexing to improve statistical power at lower cost with simplified analysis. Therefore, it is foreseeable that both sequencing platforms will remain relevant for transcriptomics study, respectively, for transcriptomics profiling and differential expression analysis. To date, Illumina platform is still leading the field of transcriptomics in simultaneously providing the most cost-effective option for both applications.

2.3.3 Data Analysis

For RNA-seq data analysis, the major difference which distinguishes between different analysis pipelines depends on the method chosen for transcript reconstruction (Fig. 2.3). There are two main strategies of transcriptome assembly, namely, align-then-assemble or assemble-then-align. These two strategies can be combined to construct a more comprehensive reference transcriptome.

For reference-guided or ab initio assembly, the reference can be based on an annotated genome or a transcriptome reference. aware aligners or gapped mappers are used when mapping reads to genome to account for mapping across exon junctions. Novel transcripts or gene structures can be discovered from non-annotated transcripts, which can then be functionally annotated. If not interested in novel transcripts, reads can be mapped to a reference transcriptome using unspliced or ungapped mappers for more accurate mapping. Transcript identification and quantification can be achieved simultaneously, such as using Cufflinks.

When no reference is available, a reference-free or de novo assembly can be performed using two different types of assembly algorithms, namely, overlap-layout-consensus (OLC) and De Bruijn graphs (DBG). DBG is based on a much faster *k*-mer indexing approach that works well with short reads compared to a more computing intensive OLC that infers consensus sequences based on a layout of all the reads and overlaps information. DBG approach is currently more popular because most RNA-seq studies are using Illumina short-read sequencing. OLC can be useful for longer sequences generated from Sanger or 454 sequencing. The assembled contigs or transcripts are then used as a reference for read mapping to

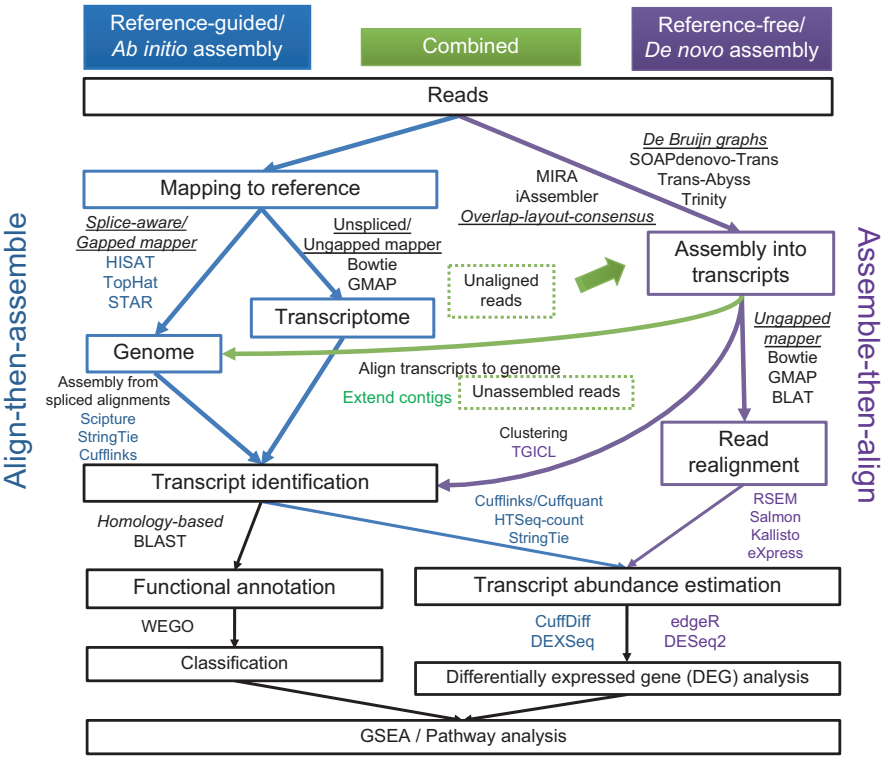

Fig. 2.3 Different analysis pipelines for transcript reconstruction from RNA-seq analysis based on a reference, de novo, or combined approach. Major differences in the approach are underlined. Some of the popular software/tools used for each step of analysis are listed and colour-coded according to different assembly approaches. Common analysis or general tools are in black font

estimate the transcript abundance. This quantification can be estimated at "transcript/isoform" or "gene" level. The transcripts generated from de novo assembly are often subjected to further clustering using software such as TGICL before further analysis.

For a more comprehensive assembly, reads that failed to align to the genome can be de novo assembled, whereas unassembled reads from de novo assembly can be used to scaffold and extend contigs based on the reference genome [91]. This combined approach helps to generate a comprehensive transcriptome that maximises the utilisation of sequencing reads. The final assembled transcriptome will serve as a reference for quantification of expression which can then be subjected to DEG analysis using various statistical software [80] that suit the experimental design or nature of the datasets. For functional annotation, assembled transcripts are BLAST searched against public databases, such as NR and Swiss-Prot, which can then be further categorised according to gene ontology (GO) or clusters of orthologous

group (COG). This functional information and results from DEG analysis can then be combined to answer biological questions based on gene set enrichment analysis (GSEA) or pathway analysis for an overview of affected metabolite pathways.

2.4 Case Study: Functional Genomics Study of *Polygonum minus*

Polygonum minus Huds. (syn. *Persicaria minor*) is rich in secondary metabolites with medicinal and pharmaceutical importance [92]. Functional genomics study of *P. minus* started in 2011 with the identification of cDNA for jasmonic acid-responsive genes in root by suppression subtractive hybridisation [93]. The first leaf, stem, and root expressed sequence tag (EST) library was established in 2012 [94]. The is followed by leaf transcriptome profiling of genes induced by salicylic acid and methyl jasmonate (MeJA) through cDNA-amplified fragment length polymorphism (AFLP) approach [95]. All of these studies relied on the low-throughput Sanger sequencing. Recently, de novo RNA-seq using a hybrid NGS approach was taken to construct a more comprehensive transcriptome profile from the leaf and root tissues, respectively, using Illumina sequencing and Roche 454 pyrosequencing [96, 97]. Table 2.4 summarises the statistics from EST library and NGS transcriptome, which shows the great improvement of currently available transcript sequences. Furthermore, DEG analysis of mRNA [98] and small RNA [99] transcriptomes in leaf treated with MeJA can help to understand the effect of elicitation on global gene reprogramming which resulted in the compositional changes of volatile organic compounds (VOCs) [36].

General workflow of RNA-seq analysis, particularly generating transcriptome profile, involves steps in the following order: raw reads preprocessing, filtering and trimming of low-quality reads and contaminant sequences, assembly and clustering, annotation, functional classification, and pathway mapping (Fig. 2.4).

Table 2.4 Statistics of *P. minus* EST library and NGS transcriptome

	EST library [94]	NGS transcriptome [96]
Sequence statistics		
Raw reads	7,292[a]	48,615,711[b]
Average read length (bp)	650	90
Processed reads	5,142	34,365,872
Assembly (*Unigenes*)	4,196	108,541
Functional analysis		
GO terms	2,024	52,796
EC assignment	200	482
KEGG pathway mapped	110	376

[a]Total count of leaf, root, and stem EST clones

[b]Leaf Illumina raw reads were combined with root 454 reads which were clipped to pseudo reads and digital normalised [96]

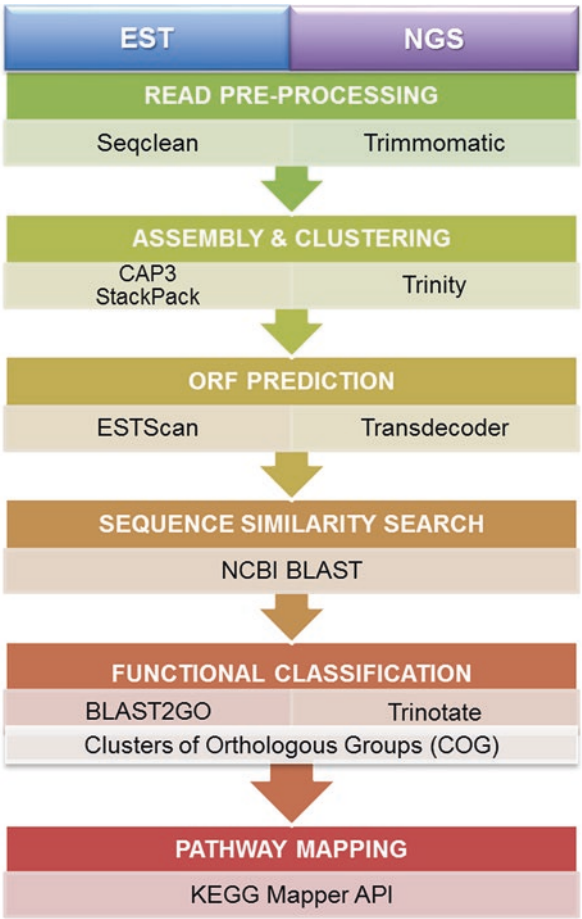

Fig. 2.4 Comparison of analysis workflow in transcriptomic studies of *P. minus*

In EST analysis, a preprocessing step was carried out using Seqclean and then an assembly step using CAP3 and StackPack, followed by open reading frame (ORF) prediction using ESTScan, whereas for RNA-seq analysis, a full Trinity analysis pipeline was followed for de novo assembly. Both required sequence similarity search using NCBI BLAST and functional classification using BLAST2GO based on Gene Ontology and Clusters of Orthologous Groups (COG). Lastly, pathway mapping was performed using KEGG Mapper.

Transcriptome profiling not only contributed to the identification of genes in response to emulated stresses but also allowed the discovery of genes involved in secondary metabolite biosynthesis. Several genes from secondary metabolite biosynthetic pathways were studied. One example is the functional characterisation of sesquiterpene synthase (*PmSTS*) which has been successfully expressed in both *Lactococcus lactis* [100] and *Arabidopsis thaliana* [101]. More recently, a recombi-

nant β-sesquiphellandrene synthase from *P. minus* was expressed and characterised [102]. Furthermore, the transcript sequence database serves as an important reference in proteomics study for protein identification. Increasing availability of genetic information on *P. minus* will help in future exploration of this plant for biotechnological applications.

References

1. Winkler H (1920) Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche. Verlag Von Gustav Fischer, Jena
2. Kaul S et al (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815
3. Adams MD et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195
4. Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
5. Craig Venter J et al (2001) The sequence of the human genome. Science 291:1304–1351
6. Waterston RH et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562
7. Auton A et al (2015) A global reference for human genetic variation. Nature 526:68–74
8. Harrow J et al (2012) GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 22:1760–1774
9. Ziller MJ et al (2013) Charting a dynamic DNA methylation landscape of the human genome. Nature 500:477–481
10. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486
11. Koepfli KP, Paten B, O'Brien SJ, Genome KC o S (2015) The genome 10K project: a way forward. Annu Rev Anim Biosci 3:57–111
12. Sandoval J, Esteller M (2012) Cancer epigenomics: beyond genomics. Curr Opin Genet Dev 22:50–55
13. Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
14. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32:D277–D280
15. Purcell S et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
16. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863–14868
17. Nielsen CB, Cantor M, Dubchak I, Gordon D, Wang T (2010) Visualizing genomes: techniques and challenges. Nat Methods 7:S5–S15
18. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192
19. Krzywinski M et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645
20. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96
21. Kersten RD et al (2013) Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc Natl Acad Sci U S A 110:E4407–E4416
22. Waters MD, Fostel JM (2004) Toxicogenomics and systems toxicology: aims and prospects. Nat Rev Genet 5:936–948

23. Kanehisa M et al (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354–D357
24. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690
25. Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630
26. Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685
27. Ellegren H (2008) Comparative genomics and the study of evolution by natural selection. Mol Ecol 17:4586–4596
28. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375
29. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108
30. McCarthy JJ, McLeod HL, Ginsburg GS (2013) Genomic medicine: a decade of successes, challenges, and opportunities. Sci Transl Med 5:189sr4
31. McCarthy MI et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369
32. Conesa A, Mortazavi A (2014) The common ground of genomics and systems biology. BMC Syst Biol 8:S1
33. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet 2:343–372
34. Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289
35. Wilson BJ, Nicholls SG (2015) The human genome project, and recent advances in personalized genomics. Risk Manage Healthc Policy 8:9–20
36. Shastry BS (2009) Single nucleotide polymorphisms. Springer, Berlin, pp 3–22
37. Orkin S, Antonarakis S, Kazazian H (1984) Base substitution at position-88 in a beta-thalassemic globin gene. Further evidence for the role of distal promoter element ACACCC. J Biol Chem 259:8679–8681
38. Bond GL et al (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119:591–602
39. Horn S et al (2013) TERT promoter mutations in familial and sporadic melanoma. Science 339:959–961
40. Madelaine R et al (2018) A screen for deeply conserved non-coding GWAS SNPs uncovers a MIR-9-2 functional mutation associated to retinal vasculature defects in human. Nucleic Acids Res 46:3517–3531
41. Janssens ACJW, van Duijn CM (2008) Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet 17:R166–R173
42. Gurdasani D et al (2015) The African genome variation project shapes medical genetics in Africa. Nature 517:327–332
43. Goff SA et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100
44. Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92
45. Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115
46. Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183
47. Xu X et al (2011) Genome sequence and analysis of the tuber crop potato. Nature 475:189–195
48. Brenchley R et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710

49. Singh R et al (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature 500:335–339
50. Rahman AYA et al (2013) Draft genome sequence of the rubber tree *Hevea brasiliensis*. BMC Genomics 14:75
51. He J et al (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484
52. Ong-Abdullah M et al (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525:533
53. Rinaldo AR, Ayliffe M (2015) Gene targeting and editing in crop plants: a new era of precision opportunities. Mol Breed 35:1–15
54. Wang Y et al (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat Biotechnol 32:947–951
55. Jiang W et al (2013) Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. Nucleic Acids Res 41:e188
56. Lawrenson T et al (2015) Induction of targeted, heritable mutations in barley and *Brassica oleracea* using RNA-guided Cas9 nuclease. Genome Biol 16:258
57. Svitashev S et al (2015) Targeted mutagenesis, precise gene editing and site-specific gene insertion in maize using Cas9 and guide RNA. Plant Physiol:00793.02015, 169(2):931–945
58. Li Z et al (2015) Cas9-guide RNA directed genome editing in soybean. Plant Physiol:00783.02015, 169(2):960–970
59. Gao C (2018) The future of CRISPR technologies in agriculture. Nat Rev Mol Cell Biol 39:1–2
60. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:R245–R249
61. Davies J (1999) Millennium bugs. Trends Genet 15:M2–M5
62. Rappé MS, Giovannoni SJ (2003) The uncultured microbial majority. Annu Rev Microbiol 57:369–394
63. Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol 6:431–440
64. Ling LL et al (2015) A new antibiotic kills pathogens without detectable resistance. Nature 517:455
65. MacNeil I et al (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. J Mol Microbiol Biotechnol 3:301–308
66. Gillespie DE et al (2002) Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 68:4301–4306
67. Brady SF, Clardy J (2004) Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water. J Nat Prod 67:1283–1286
68. Oyama LB et al (2017) Buwchitin: a ruminal peptide with antimicrobial potential against *Enterococcus faecalis*. Front Chem 5:51
69. Nasrin S et al (2018) Chloramphenicol derivatives with antibacterial activity identified by functional metagenomics. J Nat Prod 81:1321
70. Hover BM et al (2018) Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. Nat Microbiol 3:415
71. Li B et al (2015) Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J 9:2490–2502
72. Forsberg KJ et al (2012) The shared antibiotic resistome of soil bacteria and human pathogens. Science 337:1107–1111
73. Wilson MC et al (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506:58–62
74. Hrdlickova R, Toloue M, Tian B (2017) RNA-seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 8. https://doi.org/10.1002/wrna.1364

75. Ma X, Tang Z, Qin J, Meng Y (2015) The use of high-throughput sequencing methods for plant microRNA research. RNA Biol 12:709–719
76. Aviner R, Geiger T, Elroy-Stein O (2013) PUNCH-P for global translatome profiling: methodology, insights and comparison to other techniques. Translation 1:e27516
77. Li W et al (2015) Comprehensive evaluation of AmpliSeq transcriptome, a novel targeted whole transcriptome RNA sequencing methodology for global gene expression analysis. BMC Genomics 16:1069
78. Saliba A-E, Westermann AJ, Gorski SA, Vogel J (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 42:8845–8860
79. Dominissini D (2014) Roadmap to the epitranscriptome. Science 346:1192
80. Lamarre S et al (2018) Optimization of an RNA-seq differential gene expression analysis depending on biological replicate number and library size. Front. Plant Sci. 9:108
81. Ching T, Huang S, Garmire LX (2014) Power analysis and sample size estimation for RNA-seq differential expression. RNA 20:1684–1696
82. de Klerk E, den Dunnen JT, 't Hoen PAC (2014) RNA sequencing: from tag-based profiling to resolving complete transcript structure. Cell Mol Life Sci 71:3537–3551
83. Jamaluddin ND, Mohd Noor N, Goh H-H (2017) Genome-wide transcriptome profiling of *Carica papaya* L. embryogenic callus. Physiol Mol Biol Plants 23:357–368
84. Conesa A et al (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13
85. Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL (2015) Informatics for RNA sequencing: a web resource for analysis on the cloud. PLOS Comput Biol 11:e1004393
86. Nagasaki H et al (2013) DDBJ read annotation pipeline: a cloud computing-based pipeline for high-throughput analysis of next-generation sequencing data. DNA Res 20:383–390
87. Afgan E et al (2018) The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 46:W537–W544
88. Bair E (2013) Identification of significant features in DNA microarray data. Wiley Interdiscip Rev Comput Stat 5. https://doi.org/10.1002/wics.1260
89. An D, Cao HX, Li C, Humbeck K, Wang W (2018) Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. Genes 9:43
90. Moll P, Ante M, Seitz A, Reda T (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods 11:972
91. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671
92. Christapher P, Parasuraman S, Christina J, Asmawi MZ, Vikneswaran M (2015) Review on *Polygonum minus*. Huds, a commonly used food additive in Southeast Asia. Pharm Res 7:1–6
93. Gor MC et al (2011) Identification of cDNAs for jasmonic acid-responsive genes in *Polygonum minus* roots by suppression subtractive hybridization. Acta Physiol Plant 33:283–294
94. Roslan ND et al (2012) Flavonoid biosynthesis genes putatively identified in the aromatic plant *Polygonum minus* via expressed sequences tag (EST) analysis. Int J Mol Sci 13:2692–2706
95. Ee SF et al (2013) Transcriptome profiling of genes induced by salicylic acid and methyl jasmonate in *Polygonum minus*. Mol Biol Rep 40:2231–2241
96. Loke K-K et al (2016) RNA-seq analysis for secondary metabolite pathway gene discovery in *Polygonum minus*. Genomics Data 7:12–13
97. Loke KK et al (2017) Transcriptome analysis of *Polygonum minus* reveals candidate genes involved in important secondary metabolic pathways of phenylpropanoids and flavonoids. Peer J 2017. PeerJ 5:e2938
98. Rahnamaie-Tajadod R, Loke KK, Goh HH, Noor NM (2017) Differential gene expression analysis in *Polygonum minus* leaf upon 24h of methyl jasmonate elicitation. Front Plant Sci 8:109
99. Nazaruddin N et al (2017) Small RNA-seq analysis in response to methyl jasmonate and abscisic acid treatment in *Persicaria minor*. Genomics Data 12:157–158

100. Song AAL et al (2012) Overexpressing 3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGR) in the lactococcal mevalonate pathway for heterologous plant sesquiterpene production. PLOS ONE 7:e52444
101. Ee SF et al (2014) Functional characterization of sesquiterpene synthase from *Polygonum minus*. Sci World J 2014:840592
102. Ker DS et al (2017) Purification and biochemical characterization of recombinant *Persicaria minor* β-sesquiphellandrene synthase. PeerJ 5:e2961