# Advanced Applications of Next-generation Sequencing Technologies to Orchid Biology

Chuan-Ming Yeh[1†], Zhong-Jian Liu[2,3,4†] and Wen-Chieh Tsai[5,6,7*]

[1]Division of Strategic Research and Development, Graduate School of Science and Engineering, Satitama University, Saitama, Japan.
[2]Shenzhen Key Laboratory for Orchid Conservation and Utilization, The National Orchid Conservation Center of China and The Orchid Conservation and Research Center of Shenzhen, Shenzhen, China.
[3]The Center for Biotechnology and BioMedicine, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China.
[4]College of Arts, College of Landscape Architecture, Fujian Agriculture and Forestry University, Fuzhou, China.
[5]Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan, Taiwan.
[6]Department of Life Sciences, National Cheng Kung University, Tainan, Taiwan.
[7]Orchid Research and Development Center, National Cheng Kung University, Tainan, Taiwan.

*Correspondence: tsaiwc@mail.ncku.edu.tw
†These authors contributed equally

https://doi.org/10.21775/cimb.027.051

## Abstract

Next-generation sequencing (NGS) technologies are revolutionizing biology by permitting transcriptome sequencing, whole-genome sequencing and resequencing, and genome-wide single nucleotide polymorphism profiling. Orchid research has benefited from this breakthrough, and a few orchid genomes are now available; new biological questions can be approached and new breeding strategies can be designed. The first part of this review describes the unique features of orchid biology. The second part provides an overview of the current NGS platforms, many of which are already used in plant laboratories. The third part summarizes the state of orchid transcriptome and genome sequencing and illustrates current achievements. The genetic sequences currently obtained will not only provide a broad scope for the study of orchid biology, but also serves as a starting point for uncovering the mystery of orchid evolution.

## Introduction

The Chinese have been cultivating fragrant *Cymbidium* species since 500 BC. The earliest book on record about orchids is Shen Nung Pen Tsao Ching, published during the Han dynasty. This book refers to well-known orchids used as popular medicines, including Dendrobium, Gsatrodia, and Bletilla. It is generally agreed that the term orchid was first used by the Greek philosopher Theophrastus in his inquiry into plants (Arditti, 1992). Orchid cultivation and growth became popular in the late eighteenth century in Europe. Voyages around the world were sponsored by the wealthy to collect orchids, herbarium species, and other exotic plants. Merchants, government officials, sea captains, plant collectors, explorers, privateers, and other travellers began sending plants to their home countries soon after they discovered them. Some of these orchids were sent to botanical gardens; others reached private growers. Subsequently, the landed gentry, the wealthy, and commercial firms started

to accumulate orchid collections. In 1794, 15 epiphytic orchids were cultivated at Kew Gardens in London. To satisfy the needs of growers, large numbers of collectors were sent to faraway places. These collectors destroyed millions of plants, discovered many new species, and suffered and died from diseases and deprivation, but sent many orchids to England. By about 1820, it became possible to heat greenhouses with hot water flowing through pipes. These advances permitted growers to simulate what they considered to be appropriate conditions for orchid culture – heat and humidity. Improved methods such as lower temperature, better ventilation and potting contributed to higher survival of the orchids and became even more popular.

The family Orchidaceae is the largest family of flowering plants and the number of species may exceed 25,000 (Atwood, 1986). Like all other living organisms, present-day orchids have evolved from ancestral forms as a result of selection pressure and adaptation. They show a wide diversity of epiphytic and terrestrial growth forms and have successfully colonized almost every habitat on earth. Factors promoting orchid species richness include specific interaction between the orchid flower and pollinator (Cozzolino and Widmer, 2005), sequential and rapid interplay between drift and natural selection (Tremblay *et al.*, 2005), obligate interaction with mycorrhiza (Otero and Flanagan, 2006), and epiphytism which is true for most of all orchids and probably two-thirds of the epiphytic flora of the world. The radiation of the orchid family has probably taken place in a comparatively short period as compared with that of most flowering plant families, which had already started to diversify in the Mid-Cretaceous period (Crane *et al.*, 1995). The time of origin of orchids is in dispute, although Dressler suggests that they originated 80 to 40 million years ago (Mya; late Cretaceous to late Eocene) (Dressler, 1981). Recently, the origin of the Orchidaceae was dated with a fossil orchid and its pollinator. The authors showed that the most recent common ancestor of extant orchids lived in the late Cretaceous (76 to 84 Mya) (Ramírez *et al.*, 2007). They also suggested the largest orchid subfamilies, which together represent >95% of living orchid species, began to diversify early in the Tertiary (65 Mya) (Ramírez *et al.*, 2007).

According to molecular phylogenetic studies, Orchidaceae comprises five subfamilies: Apostasioideae, Cypripedioideae, Vanilloideae, Orchidoideae and Epidendroideae. The Apostasioideae is considered the sister group to other orchids. Vanilloideae diverged just before Cypripedioideae. Both subfamilies have relatively low numbers of genera and species. Most of the taxonomic diversity in orchids is in two recently expanded sister subfamilies: Orchidoideae and especially Epidendroideae (Górniaka *et al.*, 2010). Orchids are known for their diversity of specialized reproductive and ecological strategies (Tsai *et al.*, 2014). For successful reproduction, the production of labellum and gynostemium (a fused structure of androecium and gynoecium) to facilitate pollination is well documented and the co-evolution of orchid flowers and pollinators is well known (Schiestl *et al.*, 2003). In addition, the especially successful evolutionary progress of orchids may be explained by mature pollen grains packaged as pollinia, pollination-regulated ovary/ovule development, synchronized timing of micro- and mega-gametogenesis for effective fertilization, and the release of thousands or millions of immature embryos (seeds without endosperm) in a mature capsule (Yu and Goh, 2001). However, despite their unique developmental reproductive biology, as well as specialized pollination and ecological strategies, orchids remain under-represented in molecular studies relative to other species-rich plant families (Peakall, 2007). The reasons may be associated with the large genome size, long life cycle, and inefficient transformation system of orchids (Hsiao *et al.*, 2011b).

During the last 30 years DNA sequencing has completely changed our vision of biology and particularly plant biology. It has been possible to characterize a large number of genes by their nucleotide sequences, thus providing a shortcut to the corresponding protein sequences and their functions. Information on gene polymorphisms has facilitated genetic mapping, gene cloning and the understanding of evolutionary relationships and has allowed for the initiation of biodiversity studies. The most popular sequencing method has been the Sanger method (Sanger *et al.*, 1977a). When combined with the use of robotics, bioinformatics, computer databases and instrumentation, the method has allowed for sequencing larger DNA fragments and, finally, complete genomes. As a result, a series of landmark genomes was

obtained, such as *Caenorabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Homo sapiens* and *Oryza sativa* (International Human Genome Sequencing Consortium, 2004; International Rice Genome Sequencing Project, 2005; The Arabidopsis Genome Initiative, 2000). The deciphering of these genomes led to the era of functional genomics and completely modified biological investigation. However, this technology remained tedious and expensive. These limiting factors stimulated the development and commercialization of next-generation sequencing (NGS) technologies, as opposed to the automated Sanger method, which is considered a first-generation technology. When coupled with the appropriate computational algorithms, the development of NGS technologies has opened new avenues on a genome-wide scale to radically alter our understanding of biology that could not be answered with classical sequencing.

## Sequencing platforms

Sanger sequencing, developed by Sanger and his colleagues in 1977 based on the chain-termination method, is predominantly employed for DNA sequencing in the following 30 years (Sanger *et al.*, 1977b). It was first commercialized by Applied Biosystems to launch the automatic sequencing machine, AB370, in 1987. By adopting capillary electrophoresis, the sequencing became faster and more accurate (Liu *et al.*, 2012; Illumina 2016). Although Sanger sequencing was applied to complete the genome projects of human and several model organisms including the first sequenced plant, *Arabidopsis thaliana*, it took plenty of time, cost and resources (The Arabidopsis Genome Initiative, 2000; International Human Genome Sequencing Consortium, 2004; van Dijk *et al.*, 2014). Therefore, National Human Genome Research Institute (NHGRI) initiated a funding programme which aimed to reduce the cost of human genome sequencing to 1000 US dollars within 10 years. It prompted NGS development to create sequencers with fast, cheap, easy-to-operate and accurate features (Liu *et al.*, 2012; van Dijk *et al.*, 2014).

The 454 Genome Sequencer employing the pyrosequencing method is the first commercial NGS platform released by 454 Life Sciences in 2005 (Margulies *et al.*, 2005), followed by the Genome Analyser in 2006 from Solexa, which was purchased by Illumina 1 year later (Liu *et al.*, 2012). The SOLiD (Sequencing by Oligo Ligation Detection) platform is the third NGS technology developed by Applied Biosystems (now Thermo Fisher) in 2007 (Valouev *et al.*, 2008). These Solexa (Illumina) and SOLiD sequencers generated short reads with only 35-bp lengths comparing to the 110-bp reads by the 454 system. However, their numbers of reads (30 and 100 million reads, respectively) are much larger than that of the 454 sequencer (200 thousand reads). In 2010, PGM (Personal Genome Machine), the first NGS platform developed by semiconductor technology which generates 100-bp reads, was released by Ion Torrent (now Thermo Fisher). Without using optical-sensing device, the detection for sequencing by Ion Torren PGM is based on measuring the change of pH during nucleotide incorporation (Rothberg *et al.*, 2011; Liu *et al.*, 2012). These unique features make PGM a smaller size with higher speed and lower cost (Liu *et al.*, 2012; van Dijk *et al.*, 2014).

The sequencing approaches of these short-read NGS platforms can be classified to the sequencing by ligation (SBL) employed by SOLiD and the sequencing by synthesis (SBS) adopted by Illumina, 454 and Ion Torren. The SBS can be further distinguished to cyclic reversible termination (CRT) for Illumina and single-nucleotide addition (SNA) for 454 and Ion Torren (Goodwin *et al.*, 2016). Basically, the NGS systems employing the SBS approaches are similar to the first-generation sequencing in that the fluorescently labelled dNTPs (deoxyribonucleotide triphosphates) are incorporated into a DNA template by DNA polymerase. The incorporated nucleotides are then detected by fluorophore excitation during sequencing (Voelkerding *et al.*, 2009; Illumina, 2016). For the SBL approach, instead of DNA polymerase, it relies on DNA ligase to determine the underlying sequence of the template DNA. The fluorescently labelled interrogation probes hybridize to the template DNA followed by ligation and imaging. The identity of the bases in the interrogation probes is indicated by the emission spectrum of the fluorophore (Shendure *et al.*, 2005; Goodwin *et al.*, 2016). The major improvements of NGS leading to high throughput and sequencing rate include preparation of NGS libraries in a cell free system instead of bacterial cloning, parallel proceeding of a large number of sequencing

reactions and direct detection of sequencing output without performing electrophoresis (van Dijk *et al.*, 2014; Chaitankar *et al.*, 2016). Compared to the conventional sequencing, second-generation sequencing technologies generate vast amounts of shorter reads, ranging from 35 to 700 bp, making it possible to produce massive sequencing data in a much shorter time (Goodwin *et al.*, 2016). The current NGS systems can complete human genome sequencing within one day, which consumed 15 years before (Illumina, 2016).

Although these short-read sequencing platforms have created a new sequencing era in the past decade, the short read lengths give rise to a new obstacle in genome assembly, such as improper discard of repeated sequences and assembling in wrong locations or orientations. It is required to develop novel computational algorithms for data analysis (Baker, 2012; van Dijk *et al.*, 2014). However, it is increasingly obvious that genomes are highly complex because of having many long repetitive elements, copy numbers and structural variations. The short-read NGS technologies are insufficient to resolve these complex elements (Goodwin *et al.*, 2016). 'Very long, very high-quality reads will do wonders for assembly, and fix many of these issues', says Adam Felsenfeld, the director of the Large-Scale Sequencing Program at the NHGRI (Baker, 2012). Although we are still not there yet, the so-called third-generation sequencing approaches can now provide an alternative choice (Goodwin *et al.*, 2016). The read lengths of the current long-read sequencing can reach to several kilobases, allowing to span complex or repetitive regions with a single continuous read. In addition, it is also helpful to identify the precise connectivity of exons and discern gene isoforms in transcriptomic research because it can span entire mRNA transcripts. No PCR amplification step before sequencing is one of the main characteristics of the third-generation sequencing technologies. The other is that the sequencing reaction can be detected in real time whether in PacBio by fluorescence or in Nanopore by electric current (Liu *et al.*, 2012). Although the more expensive cost and relatively lower throughput than second-generation sequencing currently limit their widespread application, ultra-long-read sequencing technologies with high outputs and low prices could be expected in the near future. In this review article, the current development and application of second- and third-generation sequencing platforms along with their benefits and drawbacks are introduced and discussed (Table 3.1).

## Roche 454

The 454 Life Sciences, purchased by Roche in 2007, released the first NGS platform, the 454 Genome Sequencer, in 2005 (Margulies *et al.*, 2005). The sequencing workflow is initiated by preparation of a sequencing library from the source nucleic acids. First, the long DNA or RNA molecules are fragmented into a suitable size (around 50 to 500 bp). The fragments are next fused with specific adapters followed by a size selection step to enrich molecules with a desired size and to remove the free adapters. The templates are adhered to micro-beads through adaptors and amplified by emulsion PCR (http://454.com/products/technology.asp; Egan *et al.*, 2012; van Dijk *et al.*, 2014). This sample preparation process is similar to the subsequently developed NGS platforms, Ion Torrent and SOLiD. For the 454 pyrosequencing, the templates is denatured and incubated with sequencing reagents including DNA polymerase, dNTP and several enzymes. Pyrophosphates are released and converted to ATP after the appropriate dNTPs are incorporated into the new strand by DNA polymerase. ATP in turn reacts with luciferase to release oxyluciferin fluorescence detected by a CCD (charge-coupled device) camera (Egan *et al.*, 2012; Goodwin *et al.*, 2016).

The Roche 454 currently offers two pyrosequencing platforms including the bench top GS Junior+ (the upgraded version of GS Junior) and FLX+ system (www.454.com). The advantages of the 454 platforms include the relatively fast run times and the long read lengths with maximum of 1 kb (Table 3.1). The generated long reads are helpful for mapping to a reference genome, *de novo* genome assembly or metagenomics applications. However, it possesses the drawbacks of the relatively low throughput, high reagent cost and high error rates in homopolymer repeats. In addition, an announcement that Roche will shut down 454 and stop the supporting services for the sequencing platform should be noticed (van Dijk *et al.*, 2014).

**Table 3.1** Comparison of performance of the current second- and third-generation sequencing platforms

| Company | Platform | Sequencing chemistry | Maximum output | Maximum reads per run | Maximum read length | Run time |
|---|---|---|---|---|---|---|
| Roche (454) | GS Junior+ | SBS (SNA) | 70 Mb | ~0.1 million | 1000 bp | 18 hours |
| | GS FLX+ | SBS (SNA) | 700 Mb | ~1 million | 1000 bp | 23 hours |
| Illumina | MiniSeq | SBS (CRT) | 8 Gb | 25 million | 2 × 150 bp | 4–24 hours |
| | MiSeq | SBS (CRT) | 15 Gb | 25 million | 2 × 300 bp | 4–55 hours |
| | NextSeq | SBS (CRT) | 120 Gb | 400 million | 2 × 150 bp | 12–30 hours |
| | HiSeq | SBS (CRT) | 1.5 Tb | 5 billion | 2 × 150 bp | 7 hours – 6 days |
| | HiSeq X | SBS (CRT) | 1.8 Tb | 6 billion | 2 × 150 bp | <3 days |
| Thermo Fisher | SOLiD 5500 | SBL | 320 Gb | ~1.4 billion | 50 or 75 bp | 10 days |
| | Ion PGM | SBS (SNA) | 2 Gb | 5.5 million | 200 or 400 bp | 4–7.3 hours |
| | Ion Proton | SBS (SNA) | 10 Gb | 80 million | 200 bp | 2–4 hours |
| | Ion S5 | SBS (SNA) | 15 Gb | 80 million | 200 or 400 bp | 2.5–4 hours |
| SeqLL (Helicos) | Heliscope | SMS | 35 Gb | ~1 billion | 55 bp | 8 days |
| Pacific Biosciences | PacBio RS II | SMRT | 1 Gb* | ~55,000 | 60 kb | 0.5–6 hours |
| | Sequel | SMRT | 7 Gb* | ~370,000 | 60 kb | 0.5–6 hours |
| Oxford Nanopore | MinION | SMRT | 42 Gb | 4.4 million | 230–300 kb | <2 days |
| | PromethION | SMRT | 12 Tb | 1250 million | 230–300 kb | <2 days |

The data are obtained from the homepages of each company and two recent review articles (Chaitankar *et al*., 2016; Goodwin *et al*., 2016). *Output per SMRT cell (Number of SMRT cell is 1–16). bp, base pairs; CRT, cyclic reversible termination; Gb, gigabase pairs; kb, kilobase pairs; Mb, megabase pairs; SBL, sequencing by ligation; SBS: sequencing by synthesis; SMRT, single-molecule real-time; SMS, single-molecule sequencing; SNA: single-nucleotide addition; Tb, terabase pairs.

## Illumina

The Illumina sequencing platform, the most widely adopted in the industry, employs the CRT approach of SBS methodology (Liu *et al*., 2012; Goodwin *et al*., 2016). The sequencing steps include library preparation, cluster generation and SBS. The polynucleotide samples are randomly fragmented followed by adapter ligation at both 5′ and 3′ ends to prepare sequencing library. After adapter ligation, PCR amplification and gel purification are performed. The library is then loaded into a flow cell for cluster generation. The nucleotide fragments are immobilized on the flow cell surface through hybridization of oligos and adapters. By bridge amplification, each fragment is amplified into a clonal cluster. The flow cell is in turn incubated with the sequencing reagents and the four fluorescent dNTPs bound with reversible terminators. The incorporated bases are identified according to emission wavelength and intensity (Illumina, 2016).

Solexa, acquired by Illumina in 2007, launched its first sequencer (the Genome Analyser) which can generate roughly 1 gigabase (Gb) of data per sequencing run in 2006. Following the Genome Analyser, the Illumina sequencers with different application and sequencing power have been developed. The current platforms include MiniSeq, MiSeq, NextSeq, HiSeq Series and HiSeq X Series (www.illumina.com). The outputs range from 1.8 to 7.8 Gb for targeted sequencing studies by the benchtop MiniSeq system to 1.6 to 1.8 terabase (Tb) for population-scale studies by the HiSeqX Series (Table 3.1). The HiSeqX Ten, released in 2014, is currently the sequencer with the highest throughput. It was upgraded to sequence 1.8 Tb per run, leading to the possibility of sequencing over 45 human genomes in a single day. In addition, the cost was down to approximately US$1000 for one human genome. It takes genome sequencing entering a period to see the differences among thousands of people and discover the critical genes causing cancer or other diseases (Illumina, 2016). However, the HiSeq X Ten is a set of 10 HiSeq X system with a price of US$10 million. In addition, the low cost claimed by Illumina is an average that can be reached by running full capacity of all HiSeq

X machines for one year. In other words, it is limited to be used only when large institutions carry out population-scale genome sequencing (Goodwin *et al.*, 2016). To overcome the limitations caused by short-read sequencing, Illumina recently released a Synthetic Long-Read Sequencing technology that can generate reads of around 10 kb. By using this technology, synthetically long fragments can be constructed from shorter sequencing reads generated by the HiSeq platform for accurate genome assembly and genome finishing (www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-synthetic-long-read.html; Li *et al.*, 2015a).

### Thermo Fisher Scientific

The SOLiD sequencing technology based on SBL approach was developed by George Church and his colleagues. It was published in 2005 for the application in the resequencing of the *Escherichia coli* genome and was later improved and released by Applied Biosystems (now Thermo Fisher) in 2007 (Shendure *et al.*, 2005; Voelkerding *et al.*, 2009). The sequencing procedures include library and template preparation, bead deposition and sequencing. Two types of libraries including fragment or mate-paired library can be constructed according to different research applications. A mixture of short fragments flanked by adaptors are generated and attached to beads followed by emulsion PCR amplification. The beads with the amplified templates are then immobilized onto a glass slide or FlowChip by covalent attachment. SBL is begun by annealing a primer to the complementary adapter. In contrast to SBS mediated by polymerase (3′ hydroxyl group), the primer offer a 5′ phosphate group for ligation with one of the fluorescently labelled interrogation probes which are octamers consisting of two specific bases followed by six degenerate bases. There are 16 combinations for the first two specific bases in the interrogation probes, such as AA, AT and so on. In the first SBL step, these probes compete for hybridization with the templates followed by ligation with the primer and detection of the fluorescence signal. The fluorophore along with three bases is in turn cleaved from the probe, leaving a 5-bp fragment with a 5′ phosphate group for the next ligation of the interrogation probes. Multiple cycles of these processes are performed to complete one round of sequencing. The extension product is then denatured and a second round of sequencing is started with a new primer complementary to the n-1 position of the adapter. Sequencing by five rounds of these primer resets (until n-4) completes each read and each base of the template is sequenced twice (www.thermofisher.com/; Voelkerding *et al.*, 2009). The accuracy rate of the SOLiD system can reach 99.94% which is higher than most other NGS systems because each base is read by twice (Liu *et al.*, 2012; Goodwin *et al.*, 2016). However, the very short read lengths (75 bp), the much longer runtime (at least 6 days) and the less well-developed kits of sample preparation limit its wide application, such as for genome assembly and structural variant detection (van Dijk *et al.*, 2014; Goodwin *et al.*, 2016).

Another well-known NGS platform purchased by Thermo Fisher is Ion Torrent. Although the Ion Torrent system adopts SBS methods similar to most NGS technologies, it is a unique sequencing platform employing an integrated complementary metal oxide semiconductor (CMOS) and an ion-sensitive field-effect transistor (ISFET) as the detection system (Goodwin *et al.*, 2016). The detection is based on measurement of the pH change resulting from proton release during nucleotide incorporation but not fluorescence (Rothberg *et al.*, 2011). For sequencing, the chip is flooded with one nucleotide each time and the incorporated nucleotide is in turn detected. If the incorrect nucleotide is added, no voltage will be detected. In case two nucleotides are added, there will be double voltage (Liu *et al.*, 2012). For different research requirements, several types of chips and instruments are offered by Ion Torrent. Their throughputs range from ~50 Mb to 15 Gb and the runtimes are from 2 to 7 hours (Table 3.1; Goodwin *et al.*, 2016). The Ion PGM is the first commercial sequencer released by Ion Torrent, targeting to clinical applications and small labs. It possesses features of higher speed, lower cost and smaller size because it is not required to perform fluorescence labelling and camera scanning (Liu *et al.*, 2012). With the latest released 318 chips, the Ion PGM improves the output to over 1 Gb. As to the higher-throughput Ion Proton system, the adopted Proton-I chip is manufactured by the 110 nm CMOS technology to increase the number of wells to ~165 million. It can produce 60 to 80 million reads per run with an output of 10 Gb (Egan *et al.*, 2012; Buermans and den Dunnen,

2014). Aiming to develop NGS platform for clinical sequencing, Ion Torrent released its dedicated diagnostic instruments, the Ion PGM Dx and the Ion S5 series. The S5 series coupled with the Ion Chef library preparation and chip loading device could be one of the platforms with the simplest operation. In this combination system, argon required in other Ion Torrent instruments is unnecessary and the plug-and-play protocols have been established. However, the higher-throughput S5 devices along with the Ion Proton have limitations for elucidating long-range genomic or transcriptomic structure, because they cannot be applied for paired-end sequencing (Goodwin *et al.*, 2016).

## Helicos Biosciences

The Heliscope, released by Helicos Biosciences, is the first sequencer for single-molecule sequencing (SMS) derived from the technology developed by Braslavsky *et al.* (2003). It was considered as the interface between second- and third-generation sequencing. In this technology, DNA polymerase is used to acquire sequence information during synthesis of the complementary strand of a single DNA template. The SMS approach is attractive because it can directly sequence nucleic acids in an unbiased manner and prepare samples in a simple way. Without steps for cloning or PCR amplification, the GC-content and size biases appeared in other NGS could be avoided (Pushkarev *et al.*, 2009; Thompson and Steinmann, 2010). The sample preparation includes DNA fragmentation, addition of poly(A) tail at the 3′ end and fluorescence labelling of the final adenosine. The poly(A) tail of the DNA templates are hybridized to Poly(dT) oligonucleotides randomly immobilized on a flow-cell surface by covalent bonding. These random sequencing positions are recorded by the fluorescence of the captured DNA templates. Before sequencing gets started, the fluorescent labels are cleaved and the flow cells are incubated with DNA polymerase and one of the four Cy5-labelled dNTPs. The 'virtual terminator' included in each nucleotide can prevent a further incorporation. The excitation of Cy5 from the incorporated dNTP is in turn detected at 647 nm. The process is repeated to determine the next incorporation of nucleotides (Harris *et al.*, 2008; Goodwin *et al.*, 2016). In addition to single DNA and cDNA molecules, the Heliscope is also the first system that can directly sequence RNA without reverse transcription (Ozsolak *et al.*, 2009). It can prevent the biases from cDNA synthesis by using other RNA sequencing technologies. Moreover, both SMS of short and long RNAs can be done together without performing different sample manipulation steps (Ozsolak, 2016). However, Helicos BioSciences filed for Chapter 11 bankruptcy in 2012 and the properties were acquired by SeqLL in 2014. The SeqLL currently offers customized services for quantitative RNA and specialty DNA sequencing by using the True Single Molecule Sequencing technology (tSMS) of HeliScope Genetic Analysis System (http://seqll.com).

## Pacific Biosciences

The PacBio RS platform, released by Pacific Biosciences in 2010, is the first third-generation sequencing platform employing the Single-Molecule Real-Time (SMRT) sequencing technology. It enables parallel and real-time detection of thousands of single-molecule sequencing reactions (Eid *et al.*, 2009; Liu *et al.*, 2012). The SMRT technology was developed based on the zero-mode waveguide (ZMW) technology published at Science in 2003 (Levene *et al.*, 2003). In the conventional approaches, pico- to nanomolar concentrations of fluorophores are suitable for optical observation of dynamics of individual molecules. However, ligand concentration at micromolar is usually required for biological reactions that make it necessary to reduce sample volume by three orders of magnitude for optical observation of single molecules. ZMWs are tiny nanoholes with a diameter of 70 nm and depth of 100 nm in a metal film. It was successfully applied to observe activity of a single DNA polymerase molecule at micromolar concentrations with microsecond temporal resolution (Levene *et al.*, 2003; McCarthy, 2010). In the current PacBio RS II and Sequel systems, each SMRT cell consists of 150,000 and 1 million of ZMWs, respectively, with a single DNA polymerase at the bottom of each nanohole. The Sequel system thus can produce seven times as many reads as the PacBio RS II (www.pacb.com/products-and-services/pacbio-systems/). During sequencing by synthesis, the DNA polymerase incorporates one of the four nucleotides labelled by different fluorescent dye into the complementary strand of the template DNA. The signal is immediately

captured and recorded as a movie format by camera inside the sequencer for real-time observation. The dNTP-bound fluorophore is cleaved by DNA polymerase before the next incorporation of dNTP (McCarthy, 2010; Liu *et al.*, 2012).

Comparing to the second-generation sequencers, the PacBio platforms have several advantages including fast sample preparation (4 to 6 h), short run times (0.5 to 6 h) and long read lengths. Without PCR step in the sample preparation, the bias and error caused by PCR is reduced. In both of PacBio RS II and Sequel systems, half of the reads are over 20 kb with an average of 10 kb making PacBio ideal for genome assembly and improvement of the existing draft genomes (Liu *et al.*, 2012; www.pacb.com/products-and-services/pacbio-systems/). In addition, by using unique circular DNA templates, the ones shorter than 3 kb can be sequenced multiple times to generate a consensus read of insert, the so-called circular consensus sequence (Goodwin *et al.*, 2016). However, the PacBio platforms have drawbacks of relatively low throughput and high cost, currently limiting the range of applications (van Dijk *et al.*, 2014).

## Oxford Nanopore

Oxford Nanopore is another third-generation sequencing technology because it also sequences single molecules in a real-time manner (van Dijk *et al.*, 2014). Instead of monitor of incorporations or hybridizations of nucleotides employed by other sequencing technologies, the Nanopore platform can directly detect the nucleotide composition of single-stranded DNAs (Goodwin *et al.*, 2016). To carry out sequencing, the bases are identified by the change in electrical conductivity when a DNA molecule is transited through a tiny biopore with diameter in nanoscale (Clarke *et al.*, 2009; Liu *et al.*, 2012).The detection of single molecules based on the nanopore method has emerged from a PNAS paper published in 1996 (Kasianowicz *et al.*, 1996). It was reported that single-stranded DNA and RNA molecules can be driven by an electric field through an ion channel formed by *S. aureus* α-hemolysin across a lipid bilayer. When each polynucleotide molecule translocates through the channel, it can be detected by a transient decrease or block of ionic conductance due to occupy of the pore's volume. Therefore, the possibility of direct and rapid sequencing of single molecules of DNA or RNA by further improving this nanopore method was proposed and investigated (Kasianowicz *et al.*, 1996; Deamer and Akeson, 2000). It was proved that a single adenine nucleotide at a specific location can be identified by the characteristic reductions of ionic current in the α-hemolysin nanopore (Ashkenasy *et al.*, 2005). However, the technique still cannot discriminate each base because the polynucleotide translocation rate is too high. By using an exonuclease enzyme to cleave individual nucleotides from DNA and covalent attachment of an adapter molecule to the protein nanopore, continuous detection of unlabelled individual nucleotide has been achieved (Clarke *et al.*, 2009).

The first commercial nanopore sequencer for sequencing single DNA molecules is MinION released by Oxford Nanopore Technologies. It is an inexpensive portable device connecting to a PC or laptop by USB and capable of producing reads of up to 10 kb (van Dijk *et al.*, 2014; Brown and Clarke, 2016). The initial ASIC (application-specific integrated circuit) chip designed for MinION Mk1 flow cell has 512 individual channels enabling to sequence at ~70 bp per second. A new 3000-channel ASIC was developed for the new released MinION Mk1B (with an expected increase to 500 bp per second) and PromethION, an ultra-high-throughput platform possessing 48 individual flow cells with running at 500 bp per second (Goodwin *et al.*, 2016; https://nanoporetech.com/products/minion). Very recently the protocol for direct RNA sequencing by the MinIon device has been developed. It is currently the only platform available for directly sequencing the original RNA strands without cDNA synthesis and PCR reaction. Although the direct RNA sequencing method was firstly reported by Helicos in 2009, it depends on the synthetic copies of the native RNA strands through the SBS reaction. The RNA modifications cannot be detected by this approach (Garalde *et al.*, 2016).

Without performing PCR and fluorescent labelling steps before sequencing, the Nanopore system can reduce costs and increase sequencing speeds (Clarke *et al.*, 2009; Laver *et al.*, 2015). In addition, except for exonuclease, it is not required to use polymerase and ligase, making Oxford Nanopore less temperature sensitive than other platforms (Liu *et al.*, 2012). Because sequence quality is high in the long reads sequenced by the Nanopore system,

it benefits to *de novo* sequencing, long-range haplotype mapping and the high-resolution analysis of chromosomal structure variation (Clarke *et al.*, 2009; Laver *et al.*, 2015).

## An overview of current of orchid genome project

### EST and BAC

Genomics studies for the orchids are just in their infancy. A survey of the literature revealed that genome size data for Orchidaceae are comparatively rare, representing just 327 species (Leitch *et al.*, 2009). Nevertheless, they reveal that Orchidaceae are currently the most variable angiosperm family with genome sizes ranging 168-fold (1C = 0.33–55.4 pg). Large scale sequence analysis of orchid genomes was first revealed by bacterial artificial chromosome (BAC) end sequences analysis in *Phalaenopsis* orchid (Hsu *et al.*, 2011). This work offers the first insights into the composition of the Phalaenopsis genome in terms of GC content, transposable elements present, protein-encoding regions, simple sequence repeats, and potential microsynteny between *Phalaenopsis* and other plant species (Hsu *et al.*, 2011). In addition to the nuclear genome, the entire chloroplast genome of *Phalaenopsis* orchid is also sequenced. The chloroplast genome of *P. aphrodite* subsp. *formosana* is about 150 kb, which encode 110 different known genes, including 74 protein-coding genes, four rRNA genes, 30 tRNA genes and two conserved reading frames of unknown function (Chang *et al.*, 2006). Furthermore, the transcripts of 74 protein-coding genes from the chloroplast genome of *P. aphrodite* subsp. *formosana* were used to study extensively the pattern of RNA editing in chloroplasts. A total of 44 editing sites are identified in the 24 transcripts of *P. aphrodite* chloroplast genes, and all are of the C-to-U conversion type (Zeng *et al.*, 2007). On the basis of the above information, the chloroplast genome of several orchids were sequenced, including *Oncidium* Gower Ramsey, *P. equestris*, *Erycina pusilla*, seven species in *Cymbidium*, *Dendrobium officinale* and *Cypripedium macranthos* (Wu *et al.*, 2010; Jheng *et al.*, 2012; Pan *et al.*, 2012; Yang *et al.*, 2013; Luo *et al.*, 2014). Further plastome sequencing of orchids will be necessary to clarify the diversity of chloroplast genomes and to improve our understanding of the relationships within the Orchidaceae.

Large-scale EST sequencing provides a gateway into the genome of organisms owing to the massive information buried in the genome-scale expression data. Before NGS technology has been developed, the most popular sequencing method has been the Sanger method applied to the EST sequencing project. A subtractive EST library was constructed from the pseudobulb of *O.* Gower Ramsey, and 1080 subtractive ESTs were obtained. Most ESTs were annotated as being involved in carbohydrate metabolism, in mannose, pectin and starch biosynthesis, transportation, and stress-related and regulatory function (Tan *et al.*, 2005). To study gene expression in *Phalaenopsis* reproductive organs, a cDNA library was constructed from mature flower buds of *P. equestris*; 5593 ESTs were sequenced and assembled into 3688 unigenes (Tsai *et al.*, 2006). In addition, a cDNA library has been constructed from scented *P. bellina* flower buds with the column removed; 2359 ESTs were sequenced and assembled into 1187 unigenes (Hsiao *et al.*, 2006). The set of floral scent-producing enzymes in the biosynthetic pathway from glyceraldehyde-3-phosphate to geraniol and linalool is recognized through these ESTs and distinguished by comparing their expression patterns in *P. bellina* and a scentless species, *P. equestris* (Hsiao *et al.*, 2006). A similar strategy was adopted for *Vanda* Mimi Palmer principally to mine any potential fragrance-related EST-SSRs as markers in the identification of fragrant vandaceous orchids endemic to Malaysia (Teh *et al.*, 2011).

#### Orchid transcriptomes generated by NGS technologies

The sudden rise of relatively low-cost and rapid NGS technologies is dramatically advancing our ability to comprehensively interrogate the nucleic-acid-based information in a cell at unparalleled resolution and depth (Delseny *et al.*, 2010). The technologies were rapidly adopted for orchid transcriptome analysis (Table 3.2). 206,960 ESTs were released from the pool containing *P. equestris*, *P. aphrodite*, and *P. bellina* and a total of 50,908 contig sequences were from six different tissues of *O.* Gower Ramsey (Chang *et al.*, 2011) by 454 technology respectively to expansively cover the *Phalaenopsis* and *Oncidium* orchid transcriptome and facilitate identifying sets of genes involved

**Table 3.2** Characteristics of findings in the literature for the application of next-generation sequencing to orchid transcriptomes

| Subfamily | Species | Sequencing platform | Tissue | Study aim | Reference |
|---|---|---|---|---|---|
| Apostasioideae | *Apostasia shenzhenica* | Illumina/Solexa | Mature flower buds | Study of floral development and evolutionary trends of orchid flowers | Tsai *et al*., 2013 |
| Apostasioideae | *Neuwiedia malipoensis* | Illumina/Solexa | Mature flower buds | | |
| Vanilloideae | *Vanilla shenzhenica* | Illumina/Solexa | Mature flower buds | | |
| Vanilloideae | *Galeola faberi* | Illumina/Solexa | Mature flower buds | | |
| Cypripedioideae | *Paphiopedilum armeniacum* | Illumina/Solexa | Mature flower buds | | |
| Cypripedioideae | *Cypripedium singchii* | Illumina/Solexa | Mature flower buds | | |
| Orchidoideae | *Habenaria delavayi* | Illumina/Solexa | Mature flower buds | | |
| Orchidoideae | *Hemipilia forrestii* | Illumina/Solexa | Mature flower buds | | |
| Epidendroideae | *Phalaenopsis equestris* | Illumina/Solexa | Mature flower buds | | |
| Epidendroideae | *Cymbidium sinense* | Illumina/Solexa | Mature flower buds | | |
| Vanilloideae | *Vanilla planifolia* | Illumina/Solexa; Roche/454 | Pod tissues, seeds | Study of biosynthetic routes to flavour components | Rao *et al*., 2014 |
| Cypripedioideae | *Paphiopedilum concolor* | Illumina Hiseq 2000 | Roots | Identify the genes that control root growth and development | Li *et al*., 2015b |
| Orchidoideae | *Ophrys* species | Roche/454; Illumina/Solexa | Flowers, labellums, leaves, flower organ from open flowers and buds | Identify genes responding for pollinator attraction | Sedeek *et al*., 2013 |
| Orchidoideae | *Orchis itali*ca | Illumina/Solexa (MiSeq) | Inflorescences | The roles of small RNAs on the flower development | Aceto *et al*., 2014 |
| Orchidoideae | *Orchis italica* | Illumina Hiseq 2500 | Florets of inflorescence before anthesis | Analysing transcripts potentially involved in flower development | Paolo *et al*., 2014 |
| Orchidoideae | *Serapias vomeracea* | Roche/454 | Protocorms | Investigate the molecular bases of the orchid response to mycorrhizal invasion | Perotto *et al*., 2014 |
| Orchidoideae | *Anoectochilus roxburghii* (Wall.) Lindl. | Illumina HiSeq 4000 | Dry seeds, seeds from asymbiotic or symbiotic germination | Study of seed germination process | Liu *et al*., 2015 |
| Orchidoideae | *Gastrodia elata* Blume | Illumina Hiseq 2000 | Vegetative tissues, corms, juvenile tubers | Address the gene regulation mechanism in gastrodin biosynthesis | Tsai *et al*., 2016 |
| Epidendroideae | *Phalaenopsis aphrodite* | Sanger: EST | Protocorms | Gene discovery and genomic annotation | Fu *et al*., 2011; Hsiao *et al*., 2011 |
| Epidendroideae | *Phalaenopsis equestris* | Sanger: EST | Mature flower buds | | |
| Epidendroideae | *Phalaenopsis bellina* | Sanger: EST | Mature flower buds without column | | |

**Table 3.2** Continued

| Subfamily | Species | Sequencing platform | Tissue | Study aim | Reference |
|---|---|---|---|---|---|
| Epidendroideae | *Phalaenopsis aphrodite; Phalaenopsis equestris; Phalaenopsis bellina* | Roche/454 | Mixed tissues | Gene discovery and genomic annotation | Fu *et al*., 2011; Hsiao *et al*., 2011 |
| Epidendroideae | *Phalaenopsis equestris* | Illumina/Solexa | Leaves | | |
| Epidendroideae | *Phalaenopsis aphrodite* | Roche/454; Illumina/Solexa | Leaves, stems, roots, young inflorescences, stalks, flower buds, flowers, germinating seeds | Investigate expressed genes involved in many biological processes of orchids | Su *et al*., 2011 |
| Epidendroideae | *Phalaenopsis aphrodite* | Illumina/Solexa | Leaves, stalks, flower buds | Study the roles of small RNAs on the regulation of flowering | An *et al*., 2011; An and Chan, 2012 |
| Epidendroideae | *Phalaenopsis aphrodite* | Illumina/Solexa | Leaves, roots, flowers, germinating seeds, young inflorescences | Identify species- and tissue-specific miRNAs | Chao *et al*., 2014 |
| Epidendroideae | *Phalaenopsis* Brother Spring Dancer 'KHM190' | Illumina Hiseq 2000 | Petals, sepals or labellums from flower buds of wild-type and peloric petal mutant plants | Study regulation of floral- organ development | Huang *et al*., 2015 |
| Epidendroideae | *Phalaenopsis* sp. | Illumina Hiseq 2000 | Explants | Examine *Phalaenopsis* leaf explant browning | Xu *et al*., 2015 |
| Epidendroideae | *Oncidium* 'Gower Ramsey' | Roche/454 | Leaves, pseudobulbs, young inflorescences, inflorescences, flower buds, mature flowers | Identify genes associated with flowering time | Chang *et al*., 2011 |
| Epidendroideae | *Oncidium* 'Gower Ramsey' | Illumina/Solexa | Roots with or without fungus | Study the roles of small RNAs on the interaction between root and the fungus | Ye *et al*., 2014 |
| Epidendroideae | *Erycina pusilla* | Illumina/Solexa | Roots, leaves, peduncles, flowers, capsules | Investigate photoperiod-dependent flowering genes | Chou *et al*., 2013 |
| Epidendroideae | *Erycina pusilla* | Illumina/Solexa | Roots, leaves, peduncles, flowers, capsules | Study the roles of small RNAs on the regulation of flowering | Lin *et al*., 2013 |
| Epidendroideae | *Cymbidium ensifolium* 'Tiegusu' | Illumina HiSeq 2000 | Flower buds, mature flower | Identify genes associated with floral development | Li *et al*., 2013 |
| Epidendroideae | *Cymbidium sinense* 'Qi Jian Bai Mo' | Illumina HiSeq 2000 | Plants in vegetative phase/floral differentiation phase/ reproductive phase | Identify genes associated with floral development | Zhang *et al*., 2013 |
| Epidendroideae | *Cymbidium hybridum* 'Golden Boy' | Illumina HiSeq 2000 | Roots with or without fungus | Study of orchid-mycorrhizal fungi interactions | Zhao *et al*., 2014 |

**Table 3.2** Continued

| Subfamily | Species | Sequencing platform | Tissue | Study aim | Reference |
|---|---|---|---|---|---|
| Epidendroideae | *Cymbidium ensifolium* 'Tiegusu' | Illumina/Solexa | Flower bud | Identify miRNAs related to floral development | Li *et al*., 2015c |
| Epidendroideae | *Cymbidium ensifolium* 'tianesu' | Roche/454 | Sepals, petals, labellums, gynostemia from flower buds and mature flowers | Reveal genes associated with floral organ differentiation | Yang and Zhu, 2015 |
| Epidendroideae | *Cymbidium sinense* 'Dharma' | Roche/454 | Roots, leaves, pseudobulbs, flowers | Analyse molecular mechanism underlying leaf-colour variations | Zhu *et al*., 2015 |
| Epidendroideae | *Cymbidium sinense; Cymbidium atropurpureum; Cymbidium mannii* | Illumina HiSeq 2000 | Leaves | Explore the evolution and molecular regulation of CAM plants | Zhang *et al*., 2016c |
| Epidendroideae | *Dendrobium officinale* | Roche/454 | Stems | Study of alkaloid biosynthesis | Guo *et al*., 2013 |
| Epidendroideae | *Dendrobium officinale* | Illumina HiSeq 2000 | Juvenile and adult plants | Identify genes associated with polysaccharide synthesis | Zhang *et al*., 2016b |
| Epidendroideae | *Dendrobium officinale* | Illumina HiSeq 2500 | Flower, roots, leaves, stems | Study of the regulatory networks of the production and accumulation of the medicinal constituents | Meng *et al*., 2016 |

in a broad range of biological processes (Hsiao *et al.*, 2011a; Chang *et al.*, 2011). A total of 121,917 unique transcripts were obtained from the *Ophrys* species by using 454 pyrosequencing and Illumina (Solexa) technologies to identify genes responding for pollinator attraction (Sedeek *et al.*, 2013). To study the genes involved in alkaloid biosynthetic pathway and polysaccharide biosynthesis in *Dendrobium officinale*, an important traditional Chinese herb, 454 pyrosequencing and Illumina technology was respectively applied to generate plentiful ESTs (Guo *et al.*, 2013; Zhang *et al.*, 2016b). To provide a general resource for studying on the pod development of *Vanilla plantifolia*, one of the most valued flavour species for its flavour qualities and is therefore widely cultivated and used for the production of food additives, the combined 454/Illumina RNA-seq platforms produced high quality *de novo* transcriptome assembly for this non-model crop species (Rao *et al.*, 2014). In addition, to improve the horticultural value of *Phalaenopsis* and *Cymbidium*, transcriptome derived from browning leaf of *Phalaenopsis* explant (sequencing by Illumina HiSeq 2000), and variable colour of *Cymbidium* leaf (sequencing by 454 pyrosequencing) were investigated (Xu *et al.*, 2015; Zhu *et al.*, 2015). Orchids are unique among plants in that mycorrhizal symbioses with soil fungi are required throughout the life history stages, from seed germination to adulthood. To understand the molecular mechanism of orchid seed germination and the symbiotic orchid–fungus relationship, 454 and Illumina were adopted to explore transcriptomes derived from *Serapias vomeracea* (Perotto *et al.*, 2014), *C. hybridium* (Zhao *et al.*, 2014), *Anoectochilus roxburghii* (Liu *et al.*, 2015), and *Gastrodia elata* (Tsai *et al.*, 2016). In Orchidaceae, about 40% species adopt crassulacean acid metabolism (CAM) to fix carbon dioxide suggesting the Orchidaceae is the largest CAM clade (Silvera *et al.*, 2009). To illuminate the origin and evolution of CAM pathway, transcriptomes derived from leaves of CAM orchids *P. equestris, D. terminale* and *C. mannii* were sampled at different time interval and sequencing by Illumine HiSeq

2000 (Deng *et al.*, 2016; Zhang *et al.*, 2016c). To study the development of spectacular orchid flower morphology, developing floral transcriptomes originating from *Cymbidium* (Li *et al.*, 2013; Yang *et al.*, 2015; Zhang *et al.*, 2013), *Orchis* (De Paolo *et al.*, 2014), and *Phalaenopsis* (Huang *et al.*, 2015) were applied to identify genes associated with floral development. Recently, root transcriptome from *Paphiopedilum concolor* was also produced to explore genes involved in orchid root development (Li *et al.*, 2015b). The accumulated transcribed sequences could be directly used to develop microarray platform, and be the resource for phylogenetic analysis. For example, an oligomicroarray containing 14,732 unigenes based on the information of expressed sequence tags derived from *Phalaenopsis* orchids was developed and applied to compare transcriptome among different types of floral organs including sepal, petal and labellum (Hsiao *et al.*, 2013). 315 single-copy orthologous genes extracted from the transcriptomes of species covering five subfamilies of Orchidaceae were applied to reconstruct a more robust phylogeny of orchids, and the results indicated that this method is more efficient and reliable than methods based on a few gene markers for phylogenic analyses, especially for the holomycotrophic species or those whose DNA sequences have been difficult to amplify (Deng *et al.*, 2015).

Next-generation sequencing technologies are not only applied to characterize orchid transcriptomes but also used to systematically analyse small RNAs in orchids (Table 3.2). The roles of small RNAs were studied on the regulation of flowering in *P. aphrodite* and *E. pusilla* (An *et al.*, 2011; An and Chan, 2012; Lin *et al.*, 2013), flower development in *Orchis italica* (Aceto *et al.*, 2014; De Paolo *et al.*, 2014) and *Cymbidium ensifolium* (Li *et al.*, 2015c), and interaction between the fungus *Piriformospora indica* and the root of an *Oncidium* hybrid orchid (Ye *et al.*, 2014). Later, comprehensive collection of small RNAs derived from *P. aphrodite* (Chao *et al.*, 2014), and *D. officinale* (Meng *et al.*, 2016) were performed. These efforts provide valuable information about the composition, expression and function of small RNAs and will aid functional genomics studies of orchids.

Recently, OrchidBase has collected the transcriptome sequences from 11 Phalaenopsis cDNA libraries and flower tissue of 10 species distributed in five subfamilies of Orchidaceae (Fu *et al.*, 2011; Tsai *et al.*, 2013; Niu *et al.*, 2016). The EST sequences collected in OrchidBase were obtained through both deep sequencing with ABI 3730, Roche 454 and Illumina/Solexa. OrchidBase is freely available at http://orchidbase.itps.ncku.edu.tw/ and provides researchers with a high-quality genetic resource for data mining and efficient experimental studies of orchid biology and biotechnology. Another orchid transcriptomic database, Orchidstra (http://orchidstra.abrc.sinica.edu.tw), was constructed from the 233,924 unique contigs of the transcriptome sequences of *P. aphrodite* by use of a Roche 454 and Illumina/Solexa platform, and the genes of tissue-specific expression were categorized by profiling analysis with RNA-seq (Su *et al.*, 2011). Oncidium cDNA libraries for six different organs, including leaves, pseudobulbs, young inflorescences, inflorescences, flower buds and mature flowers, were generated from 50,908 contig sequences by use of the Roche 454 platform and were constructed into the OncidiumOrchidGenomeBase (http://predictor.nchu.edu.tw/oogb/) (Chang *et al.*, 2011). All this EST information will be very useful for gene annotation in genomic sequencing, specificity of orchids, and organization of the orchid genome. The plentiful collection of ESTs and BESs in *Phalaenopsis* makes them reasonable candidates for orchid whole-genome sequencing. The two native *Phalaenopsis* species in Taiwan, *P. equestris* and *P. aphrodite* subsp. *formosana*, are usually used as parents for breeding and have a relatively small genome size of 3.37 pg/2C and 2.80 pg/2C, respectively. The basic studies and genomics information collected have laid the groundwork for *P. equestris* to serve as a model orchid plant for whole-genome sequencing.

## Orchid genome project

With the quick development and lower cost of NGS, whole genome sequencing of non-model species, like orchids, can be realized. The first milestone is sequencing the tropical epiphytic orchid *Phalaenopsis equestris*, a frequently used parent species for orchid breeding (Cai *et al.*, 2015). The *P. equestris* genome is sequenced via a whole-genome shotgun strategy (Illumina technology) and the genome size is estimated to be 1.16 Gb contains with 29,431 predicted protein-coding genes. Analysing the *P. equestris* genome showed that repetitive DNAs,

mostly transposable elements (TEs), account for the majority of the genome, at 62%. The authors find evidence for an orchid-specific paleopolyploidy event that preceded the radiation of most orchid clades. This species is also the first whole-genome-sequenced CAM plant and a gene family (α carbonic anhydrase) involved in CAM pathway is found having an obvious expansion which suggests that gene duplication might have contributed to the evolution of CAM photosynthesis in *P. equestris*. In addition, genes located at the heterozygous regions might relate to self-incompatibility. Genes in type II MADS-box clades, including the E-class, C/D-class, B-class AP3 and AGL6 clades, are found contained more genes than other species. These expanded clades are involved in orchid floral organs which can support the unique evolutionary routes of these floral organ identity genes associated with the unique labellum and gynostemium innovation in orchids. Furthermore, the *Phalaenopsis* genome sequence was applied to identify MYB genes controlling floral pigmentation patterning (Hsu *et al.*, 2015), and TCP genes involved in ovule development (Lin *et al.*, 2016).

Having both ornamental value and a broad range of therapeutic effects, *Dendrobium officinale* is the other Orchidaceae plant which was sequenced by combining the second-generation Illumina Hiseq 2000 and third-generation PacBio sequencing technologies (Yan *et al.*, 2015). The assembled genome of *D. officinale* has a predicted 35,567 protein-coding genes. The number of predicted genes in *D. officinale* is higher than that in *Phalaenopis*. For example, the number of B-class MADS-box genes presented in *D. officinale* is much higher than that in *Phalaenopsis*. In *Phalaenopsis*, there are four members in B-class AP3-like subfamily, and one member in B-class PI-like subfamily. In contrast, there are 19 AP3-like genes and five PI-like genes presented in the *Dendrobium* genome. It is possible that the plants used for the whole genome sequencing are not native species, but hybrids. Later, another *Dendrobium* species, *Dendrobium catenatum* (鐵皮石斛), was whole genome sequenced by Illumina HiSeq 2000 platform (Zhang *et al.*, 2016a). The predicted 28,910 protein-coding genes are comparable with those of *Phalaenopsis*, and a whole genome duplication event could be share with *Phalaenopsis*. The expansion of many resistance-related genes in *Dendrobium* suggests a powerful immune system responsible for adaptation to a wide range of ecological niches. In addition, extensive duplication of genes involved in glucomannan synthase activities is likely related to the synthesis of medicinal polysaccharides. Expansion of MADS-box gene clades ANR1, StMADS11, and MIKC*, involved in the regulation of development and growth, suggests that these expansions are associated with the astonishing diversity of plant architecture in the genus *Dendrobium* (Zhang *et al.*, 2016a).

These complete genome sequences of Orchidaceae species will facilitate future research on the diversity and evolution of orchid plants. The genome sequences will also be an important resource for genetic engineering, such as molecular marker-assisted breeding and the production of transgenic plants, which are necessary to increase the efficiency of orchid breeding and aid orchid horticulture research.

## Future perspective

About 2500 years ago Confucius wrote 'Lan (orchid) that grows in the deep valleys never withholds its fragrances even without being appreciated'. Then, 300 years ago, Charles Darwin, in a letter to Joseph Hooker, wrote 'I never was more interested in any subject in my life, than in this of Orchids'. Because of the unique reproduction strategy in orchids, their origin has been a recurring question in botany and evolutionary biology since the nineteenth century. The study of orchid biology by using NGS technologies, although still young, has already offered new and exciting perspectives on this intriguing plant family. Recent advances in sequencing technologies and functional genomics methodologies have allowed studies on orchid biology to become a standard scientific research topic accessible to many investigators, which has in turn resulted in many exciting new discoveries. With the whole genome sequences of *P. equestris*, *D. catenatum*, and *D. officinale* available, the genetic blueprint of orchids provides a basic understanding of the genetic basis of orchids. Furthermore, the genome sequences of the primitive orchid *Apostasia* and one of the most popular aromatic orchids, *Vanilla*, will be available soon. The efforts by many scientists to use a plethora of genome information and genomics tools will lead to a promising understanding of the biological, physiological, molecular and genetic

mechanisms of orchids in years to come. In addition, we will have access to a greater portion of their genetic diversity, thus allowing orchid breeders to associate this diversity with phenotypic traits and to continue to engineer new varieties better adapted to a changing environment. Clearly a new era of orchid biology is now open because of the sequencing revolution.

## Acknowledgements

This work was supported by Grants 103-2313-B-006-001-MY3, 104-2321-B-006-025-, and 105-2321-B-006-026- from Ministry of Science and Technology, Taiwan, and was supported by the 948 Program of State Forestry Administration, China (No. 2011-4-53), Development Special Fund of Biological Industry of Shenzhen Municipality (No. JC201005310692A), Development Funds for Emerging Industries of Strategic Importance of Shenzhen (No. JCYJ20140402093332029, No. NYSW20140331010039), Fundamental Research Project of Shenzhen Municipality (No. JCYJ20150403150235943), and Forestry Science and Technology Innovation Fund Project of Guangdong Province (No. 2016KJCX025).

## References

Aceto, S., Sica, M., De Paolo, S., D'Argenio, V., Cantiello, P., Salvatore, F., and Gaudio, L. (2014). The analysis of the inflorescence miRNome of the orchid *Orchis italica* reveals a *DEF*-like MADS-box gene as a new miRNA target. PLOS ONE *9*, e97839. http://dx.doi.org/10.1371/journal.pone.0097839

An, F.M., and Chan, M.T. (2012). Transcriptome-wide characterization of miRNA-directed and non-miRNA-directed endonucleolytic cleavage using Degradome analysis under low ambient temperature in *Phalaenopsis aphrodite* subsp. *formosana*. Plant Cell Physiol. *53*, 1737–1750. http://dx.doi.org/10.1093/pcp/pcs118

An, F.M., Hsiao, S.R., and Chan, M.T. (2011). Sequencing-based approaches reveal low ambient temperature-responsive and tissue-specific microRNAs in *Phalaenopsis* orchid. PLOS ONE *6*, e18937. http://dx.doi.org/0.1371/journal.pone.0018937

(The reference is the same as The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–815. http://dx.doi.org/10.1038/35048692 which listed in Page 4)Arditti, J. (1992). Fundamentals of Orchid Biology. John Wiley, New York, USA.

Ashkenasy, N., Sánchez-Quesada, J., Bayley, H., and Ghadiri, M.R. (2005). Recognizing a single base in an individual DNA strand: a step toward DNA sequencing in nanopores. Angew. Chem. Int. Ed. Engl. *44*, 1401–1404. http://dx.doi.org/10.1002/anie.200462114

Atwood, J.T. (1986). The size of Orchidaceae and the systematic distribution of epiphytic orchids. Selbyana *9*, 171–186.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nature Met. *9*, 333–337. http://dx.doi.org/10.1038/nmeth.1935

Brown, C.G., and Clarke, J. (2016). Nanopore development at Oxford Nanopore. Nat. Biotechnol. *34*, 810–811. http://dx.doi.org/10.1038/nbt.3622

Buermans, H.P., and den Dunnen, J.T. (2014). Next generation sequencing technology: Advances and applications. Biochim. Biophys. Acta *1842*, 1932–1941.

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W., Chen, L.J., He, Y., Xu, Q., Bian, C., *et al.* (2015). The genome sequence of the orchid *Phalaenopsis equestris*. Nature Genet. *47*, 65–72. http://dx.doi.org/10.1038/ng.3149

Chaitankar, V., Karakülah, G., Ratnapriya, R., Giuste, F.O., Brooks, M.J., and Swaroop, A. (2016). Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Prog. Retin. Eye. Res. *55*, 1–31.

Chang, C.C., Lin, H.C., Lin, I.P., Chow, T.Y., Chen, H.H., Chen, W.H., Cheng, C.H., Lin, C.Y., Liu, S.M., Chang, C.C., *et al.* (2006). The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol. Biol. Evol. *23*, 279–291. http://dx.doi.org/10.1093/molbev/msj029

Chang, Y.Y., Chu, Y.W., Chen, C.W., Leu, W.M., Hsu, H.F., and Yang, C.H. (2011). Characterization of *Oncidium* 'Gower Ramsey' transcriptomes using 454 GS-FLX pyrosequencing and their application to the identification of genes associated with flowering time. Plant. Cell. Physiol. *52*, 1532–1545. http://dx.doi.org/10.1093/pcp/pcr101

Chao, Y.T., Su, C.L., Jean, W.H., Chen, W.C., Chang, Y.C., and Shih, M.C. (2014). Identification and characterization of the microRNA transcriptome of a moth orchid *Phalaenopsis aphrodite*. Plant. Mol. Biol. *84*, 529–548. http://dx.doi.org/10.1007/s11103-013-0150-0

Chou, M.L., Shih, M.C., Chan, M.T., Liao, S.Y., Hsu, C.T., Haung, Y.T., Chen, J.J., Liao, D.C., Wu, F.H., and Lin, C.S. (2013). Global transcriptome analysis and identification of a *CONSTANS*-like gene family in the orchid *Erycina pusilla*. Planta *237*, 1425–1441. http://dx.doi.org/10.1007/s00425-013-1850-z

Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. Nat. Nanotechnol. *4*, 265–270. http://dx.doi.org/10.1038/nnano.2009.12

Cozzolino, S., and Widmer, A. (2005). Orchid diversity: an evolutionary consequence of deception? Trends Ecol. Evol. *20*, 487–494.

Crane, P.R., Friis, E.M., and Pedersen, K.R. (1995). The origin and early diversification of angiosperms. Nature *374*, 27–33. http://dx.doi.org/10.1038/374027a0

De Paolo, S., Salvemini, M., Gaudio, L., and Aceto, S. (2014). *De novo* transcriptome assembly from inflorescence of *Orchis italica*: analysis of coding and non-coding transcripts. PLOS ONE *9*, e102155. http://dx.doi.org/10.1371/journal.pone.0102155

Deamer, D.W., and Akeson, M. (2000). Nanopores and nucleic acids: prospects for ultrarapid sequencing. Trends Biotechnol. *18*, 147–151.

Delseny, M., Han, B., and Hsing, Y.I.e. (2010). High throughput DNA sequencing: The new sequencing revolution. Plant. Sci. *179*, 407–422. http://dx.doi.org/10.1016/j.plantsci.2010.07.019

Deng, H., Zhang, G.Q., Lin, M., Wang, Y., and Liu, Z.J. (2015). Mining from transcriptomes: 315 single-copy orthologous genes concatenated for the phylogenetic analyses of Orchidaceae. Ecol. Evol. *5*, 3800–3807. http://dx.doi.org/10.1002/ece3.1642

Deng, H., Zhang, L.S., Zhang, G.Q., Zheng, B.Q., Liu, Z.J., and Wang, Y. (2016). Evolutionary history of PEPC genes in green plants: Implications for the evolution of CAM in orchids. Mol. Phylogenet. Evol. *94*, 559–564. http://dx.doi.org/10.1016/j.ympev.2015.10.007

Dressler, R.L. (1981). The orchids: Natural history and classification. Cambridge, Massachusetts, USA: Harvard University Press.

Egan, A.N., Schlueter, J., and Spooner, D.M. (2012). Applications of next-generation sequencing in plant biology. Am. J. Bot. *99*, 175–185. http://dx.doi.org/10.3732/ajb.1200020

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138. http://dx.doi.org/10.1126/science.1162986

Fu, C.H., Chen, Y.W., Hsiao, Y.Y., Pan, Z.J., Liu, Z.J., Huang, Y.M., Tsai, W.C., and Chen, H.H. (2011). OrchidBase: a collection of sequences of the transcriptome derived from orchids. Plant. Cell. Physiol. *52*, 238–243. http://dx.doi.org/10.1093/pcp/pcq201

Garalde, D.R., Snell, E.A., Jachimowicz, D., Heron, A.J., Bruce, M., Lloyd, J., Warland, A., Pantic, N., Admassu, T., Ciccone, J.,*et al.* (2016). Highly parallel direct RNA sequencing on an array of nanopores. http://dx.doi.org/10.1101/068809.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351. http://dx.doi.org/10.1038/nrg.2016.49

Górniak, M., Paun, O., and Chase, M.W. (2010). Phylogenetic relationships within Orchidaceae based on a low-copy nuclear coding gene, *Xdh*: Congruence with organellar and nuclear ribosomal DNA results. Mol. Phylogenet. Evol. *56*, 784–795. http://dx.doi.org/10.1016/j.ympev.2010.03.003

Guo, X., Li, Y., Li, C., Luo, H., Wang, L., Qian, J., Luo, X., Xiang, L., Song, J., Sun, C., *et al.* (2013). Analysis of the *Dendrobium officinale* transcriptome reveals putative alkaloid biosynthetic genes and genetic markers. Gene *527*, 131–138. http://dx.doi.org/10.1016/j.gene.2013.05.073

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., *et al.* (2008). Single-molecule DNA sequencing of a viral genome. Science *320*, 106–109. http://dx.doi.org/10.1126/science.1150427

Hsiao, Y.Y., Chen, Y.W., Huang, S.C., Pan, Z.J., Fu, C.H., Chen, W.H., Tsai, W.C., and Chen, H.H. (2011a). Gene discovery using next-generation pyrosequencing to develop ESTs for *Phalaenopsis* orchids. BMC Genomics *12*, 360. http://dx.doi.org/10.1186/1471-2164-12-360

Hsiao, Y.Y., Pan, Z.J., Hsu, C.C., Yang, Y.P., Hsu, Y.C., Chuang, Y.C., Shih, H.H., Chen, W.H., Tsai, W.C., and Chen, H.H. (2011b). Research on orchid biology and biotechnology. Plant. Cell. Physiol. *52*, 1467–1486. http://dx.doi.org/10.1093/pcp/pcr100

Hsiao, Y.Y., Huang, T.H., Fu, C.H., Huang, S.C., Chen, Y.J., Huang, Y.M., Chen, W.H., Tsai, W.C., and Chen, H.H. (2013). Transcriptomic analysis of floral organs from *Phalaenopsis* orchid by using oligonucleotide microarray. Gene *518*, 91–100. http://dx.doi.org/10.1016/j.gene.2012.11.069

Hsiao, Y.Y., Tsai, W.C., Kuoh, C.S., Huang, T.H., Wang, H.C., Leu, Y.L., Wu, T.S., Chen, W.H., and Chen, H.H. (2006). Comparison of transcripts in *Phalaenopsis bellina* and *Phalaenopsis equestris* (Orchidaceae) flowers to deduce the monoterpene biosynthesis pathway. BMC Plant Biol. *6*, 18. http://dx.doi.org/10.1186/1471-2229-6-14

Illumina. (2016). An introduction to next-generation sequencing technology. (www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf).

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945. http://dx.doi.org/10.1016/j.prrv.2009.04.003

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature *436*, 793–800. http://dx.doi.org/10.1038/nature03895

Jheng, C.F., Chen, T.C., Lin, J.Y., Chen, T.C., Wu, W.L., and Chang, C.C. (2012). The comparative chloroplast genomic analysis of photosynthetic orchids and developing DNA markers to distinguish *Phalaenopsis* orchids. Plant Sci. *190*, 62–73. http://dx.doi.org/10.1016/j.plantsci.2012.04.001

Kasianowicz, J.J., Brandin, E., Branton, D., and Deamer, D.W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. U.S.A. *93*, 13770–13773.

Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol. Detect. Quantif. *3*, 1–8. http://dx.doi.org/10.1016/j.bdq.2015.02.001

Leitch, I.J., Kahandawala, I., Suda, J., Hanson, L., Ingrouille, M.J., Chase, M.W., and Fay, M.F. (2009). Genome size diversity in orchids: consequences and evolution. Ann. Bot. *104*, 469–481. http://dx.doi.org/10.1093/aob/mcp003

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science *299*, 682–686. http://dx.doi.org/10.1126/science.1079700

Li, R., Hsieh, C.L., Young, A., Zhang, Z., Ren, X., and Zhao, Z. (2015a). Illumina Synthetic Long Read Sequencing Allows Recovery of Missing Sequences even in the 'Finished' *C. elegans* Genome. Sci. Rep. *5*, 10814. http://dx.doi.org/10.1038/srep10814

Li, D.M., Zhao, C.Y., Liu, X.R., Liu, X.F., Lin, Y.J., Liu, J.W., Chen, H.M., and Lǚ, F.B. (2015b). *De novo* assembly and characterization of the root transcriptome and

development of simple sequence repeat markers in *Paphiopedilum concolor*. Genet. Mol. Res. *14*, 6189–6201. http://dx.doi.org/10.4238/2015.June.9.5

Li, X., Jin, F., Jin, L., Jackson, A., Ma, X., Shu, X., Wu, D., and Jin, G. (2015c). Characterization and comparative profiling of the small RNA transcriptomes in two phases of flowering in *Cymbidium ensifolium*. BMC Genomics *16*, 622. http://dx.doi.org/10.1186/s12864-015-1764-1

Li, X., Luo, J., Yan, T., Xiang, L., Jin, F., Qin, D., Sun, C., and Xie, M. (2013). Deep sequencing-based analysis of the *Cymbidium ensifolium* floral transcriptome. PLOS ONE *8*, e85480. http://dx.doi.org/10.1371/journal.pone.0085480

Lin, C.S., Chen, J.J., Huang, Y.T., Hsu, C.T., Lu, H.C., Chou, M.L., Chen, L.C., Ou, C.I., Liao, D.C., Yeh, Y.Y., *et al.* (2013). Catalog of *Erycina pusilla* miRNA and categorization of reproductive phase-related miRNAs and their target gene families. Plant. Mol. Biol. *82*, 193–204. http://dx.doi.org/10.1007/s11103-013-0055-y

Lin, Y.F., Chen, Y.Y., Hsiao, Y.Y., Shen, C.Y., Hsu, J.L., Yeh, C.M., Mitsuda, N., Ohme-Takagi, M., Liu, Z.J., and Tsai, W.C. (2016). Genome-wide identification and characterization of TCP genes involved in ovule development of *Phalaenopsis equestris*. J. Exp. Bot. *67*, 5051–5066. http://dx.doi.org/10.1093/jxb/erw273

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. *2012*, 251364. http://dx.doi.org/10.1155/2012/251364

Liu, S.S., Chen, J., Li, S.C., Zeng, X., Meng, Z.X., and Guo, S.X. (2015). Comparative transcriptome analysis of genes involved in GA-GID1-DELLA regulatory module in symbiotic and asymbiotic seed germination of *Anoectochilus roxburghii* (Wall.) Lindl. (Orchidaceae). Int. J. Mol. Sci. *16*, 30190–30203. http://dx.doi.org/10.3390/ijms161226224

Luo, J., Hou, B.W., Niu, Z.T., Liu, W., Xue, Q.Y., and Ding, X.Y. (2014). Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the orchidaceae and development of molecular markers for phylogenetic applications. PLOS ONE *9*, e99016. http://dx.doi.org/10.1371/journal.pone.0099016

McCarthy, A. (2010). Third-generation DNA sequencing: pacific biosciences' single molecule real time technology. Chem. Biol. *17*, 675–676. http://dx.doi.org/10.1016/j.chembiol.2010.07.004

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Meng, Y., Yu, D., Xue, J., Lu, J., Feng, S., Shen, C., and Wang, H. (2016). A transcriptome-wide, organ-specific regulatory map of *Dendrobium officinale*, an important traditional Chinese orchid herb. Sci. Rep. *6*, 18864. http://dx.doi.org/10.1038/srep18864

Niu, S.C., Xu, Q., Zhang, G.Q., Zhang, Y.Q., Tsai, W.C., Hsu, J.L., Liang, C.K., Luo, Y.B., and Liu, Z.J. (2016). *De novo* transcriptome assembly databases for the butterfly orchid *Phalaenopsis equestris*. Sci. Data *3*, 160083. http://dx.doi.org/10.1038/sdata.2016.83

Otero, J.T., and Flanagan, N.S. (2006). Orchid diversity – beyond deception. Trends Ecol. Evol. *21*, 64–65.

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., and Milos, P.M. (2009). Direct RNA sequencing. Nature *461*, 814–818. http://dx.doi.org/10.1038/nature08390

Ozsolak, F. (2016). Attomole-level Genomics with Single-molecule Direct DNA, cDNA and RNA Sequencing Technologies. Curr. Issues. Mol. Biol. *18*, 43–48.

Pan, I.C., Liao, D.C., Wu, F.H., Daniell, H., Singh, N.D., Chang, C., Shih, M.C., Chan, M.T., and Lin, C.S. (2012). Complete chloroplast genome sequence of an orchid model plant candidate: *Erycina pusilla* apply in tropical *Oncidium* breeding. PLOS ONE *7*, e34738. http://dx.doi.org/10.1371/journal.pone.0034738

Peakall, R. (2007). Speciation in the Orchidaceae: confronting the challenges. Mol. Ecol. *16*, 2834–2837.

Perotto, S., Rodda, M., Benetti, A., Sillo, F., Ercole, E., Rodda, M., Girlanda, M., Murat, C., and Balestrini, R. (2014). Gene expression in mycorrhizal orchid protocorms suggests a friendly plant-fungus relationship. Planta *239*, 1337–1349. http://dx.doi.org/10.1007/s00425-014-2062-x

Pushkarev, D., Neff, N.F., and Quake, S.R. (2009). Single-molecule sequencing of an individual human genome. Nat. Biotechnol. *27*, 847–850. http://dx.doi.org/10.1038/nbt.1561

Ramírez, S.R., Gravendeel, B., Singer, R.B., Marshall, C.R., and Pierce, N.E. (2007). Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. Nature *448*, 1042–1045.

Rao, X., Krom, N., Tang, Y., Widiez, T., Havkin-Frenkel, D., Belanger, F.C., Dixon, R.A., and Chen, F. (2014). A deep transcriptomic analysis of pod development in the vanilla orchid (*Vanilla planifolia*). BMC Genomics *15*, 964. http://dx.doi.org/10.1186/1471-2164-15-964

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., *et al.* (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352. http://dx.doi.org/10.1038/nature10242

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison III, C.A., Slocombe, P.M., and Smith, M. (1977a). Nucleotide sequence of bacteriophage phi X174 DNA. Nature *265*, 687–695. http://dx.doi.org/10.1038/265687a0

Sanger, F., Nicklen, S., and Coulson, A.R. (1977b). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74*, 5463–5467.

Schiestl, F.P., Peakall, R., Mant, J.G., Ibarra, F., Schulz, C., Franke, S., and Francke, W. (2003). The chemistry of sexual deception in an orchid-wasp pollination system. Science *302*, 437–438. http://dx.doi.org/10.1126/science.1087835

Sedeek, K.E., Qi, W., Schauer, M.A., Gupta, A.K., Poveda, L., Xu, S., Liu, Z.J., Grossniklaus, U., Schiestl, F.P., and Schlüter, P.M. (2013). Transcriptome and proteome data reveal candidate genes for pollinator attraction in sexually deceptive orchids. PLOS ONE *8*, e64621. http://dx.doi.org/10.1371/journal.pone.0064621

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D.,

Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science *309*, 1728–1732.

Silvera, K., Santiago, L.S., Cushman, J.C., and Winter, K. (2009). Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. Plant Physiol. *149*, 1838–1847. http://dx.doi.org/10.1104/pp.108.132555

Su, C.L., Chao, Y.T., Alex Chang, Y.C., Chen, W.C., Chen, C.Y., Lee, A.Y., Hwa, K.T., and Shih, M.C. (2011). De novo assembly of expressed transcripts and global analysis of the *Phalaenopsis aphrodite* transcriptome. Plant. Cell. Physiol. *52*, 1501–1514. http://dx.doi.org/10.1093/pcp/pcr097

Tan, J., Wang, H.L., and Yeh, K.W. (2005). Analysis of organ-specific, expressed genes in *Oncidium* orchid by subtractive expressed sequence tags library. Biotechnol. Lett *27*, 1517–1528. http://dx.doi.org/10.1007/s10529-005-1468-8

Teh, S.L., Chan, W.S., Abdullah, J.O., and Namasivayam, P. (2011). Development of expressed sequence tag resources for *Vanda* Mimi Palmer and data mining for EST-SSR. Mol. Biol. Rep. *38*, 3903–3909. http://dx.doi.org/10.1007/s11033-010-0506-3

The *Arabidopsis* Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796–815. http://dx.doi.org/10.1038/35048692

Thompson, J.F., and Steinmann, K.E. (2010). Single molecule sequencing with a HeliScope genetic analysis system. Curr. Protoc. Mol. Biol. *Chapter 7*, Unit7.10. http://dx.doi.org/10.1002/0471142727.mb0710s92

Tremblay, R.L., Ackerman, J.D., Zimmerman, J.K., and Calvo, R.N. (2005). Variation in sexual reproduction in orchids and its evolutionary consequence: a spasmodic journey to diversification. Biol. J. Linn. Soc. *84*, 1–54. http://dx.doi.org/10.1111/j.1095-8312.2004.00400.x

Tsai, C.C., Wu, K.M., Chiang, T.Y., Huang, C.Y., Chou, C.H., Li, S.J., and Chiang, Y.C. (2016). Comparative transcriptome analysis of *Gastrodia elata* (Orchidaceae) in response to fungus symbiosis to identify gastrodin biosynthesis-related genes. BMC Genomics *17*, 212. http://dx.doi.org/10.1186/s12864-016-2508-6

Tsai, W.C., Fu, C.H., Hsiao, Y.Y., Huang, Y.M., Chen, L.J., Wang, M., Liu, Z.J., and Chen, H.H. (2013). OrchidBase 2.0: comprehensive collection of Orchidaceae floral transcriptomes. Plant. Cell. Physiol. *54*, e7. http://dx.doi.org/10.1093/pcp/pcs187

Tsai, W.C., Hsiao, Y.Y., Lee, S.H., Tung, C.W., Wang, D.P., Wang, H.C., Chen, W.H., and Chen, H.H. (2006). Expression analysis of the ESTs derived from the flower buds of *Phalaenopsis equestris*. Plant Sci. *170*, 426–432. http://dx.doi.org/10.1016/j.plantsci.2005.08.029

Tsai, W.C., Pan, Z.J., Su, Y.Y., and Liu, Z.J. (2014). New insight into the regulation of floral morphogenesis. Int. Rev. Cell. Mol. Biol. *311*, 157–182. http://dx.doi.org/10.1016/B978-0-12-800179-0.00003-9

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J.A., Costa, G., McKernan, K., *et al.* (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. Genome Res. *18*, 1051–1063. http://dx.doi.org/10.1101/gr.076463.108

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends Genet. *30*, 418–426. http://dx.doi.org/10.1016/j.tig.2014.07.001

Voelkerding, K.V., Dames, S.A., and Durtschi, J.D. (2009). Next-generation sequencing: from basic research to diagnostics. Clin. Chem. *55*, 641–658. http://dx.doi.org/10.1373/clinchem.2008.112789

Wu, F.H., Chan, M.T., Liao, D.C., Hsu, C.T., Lee, Y.W., Daniell, H., Duvall, M.R., and Lin, C.S. (2010). Complete chloroplast genome of *Oncidium* Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biol. *10*, 68. http://dx.doi.org/10.1186/1471-2229-10-68

Xu, C., Zeng, B., Huang, J., Huang, W., and Liu, Y. (2015). Genome-wide transcriptome and expression profile analysis of *Phalaenopsis* during explant browning. PLOS ONE *10*, e0123356. http://dx.doi.org/10.1371/journal.pone.0123356

Yan, L., Wang, X., Liu, H., Tian, Y., Lian, J., Yang, R., Hao, S., Wang, X., Yang, S., Li, Q., *et al.* (2015). The genome of *Dendrobium officinale* illuminates the biology of the important traditional chinese orchid herb. Mol. Plant *8*, 922–934. http://dx.doi.org/10.1016/j.molp.2014.12.011

Yang, F., and Zhu, G. (2015). Digital gene expression analysis based on *de novo* transcriptome assembly reveals new genes associated with floral organ differentiation of the orchid plant *Cymbidium ensifolium*. PLOS ONE *10*, e0142434. http://dx.doi.org/10.1371/journal.pone.0142434

Yang, J.B., Tang, M., Li, H.T., Zhang, Z.R., and Li, D.Z. (2013). Complete chloroplast genome of the genus *Cymbidium*: lights into the species identification, phylogenetic implications and population genetic analyses. BMC Evol. Biol. *13*, 84. http://dx.doi.org/10.1186/1471-2148-13-84

Ye, W., Shen, C.H., Lin, Y., Chen, P.J., Xu, X., Oelmüller, R., Yeh, K.W., and Lai, Z. (2014). Growth promotion-related miRNAs in *Oncidium* orchid roots colonized by the endophytic fungus *Piriformospora indica*. PLOS ONE *9*, e84920. http://dx.doi.org/10.1371/journal.pone.0084920

Yu, H., and Goh, C.J. (2001). Molecular genetics of reproductive biology in orchids. Plant Physiol. *127*, 1390–1393.

Zeng, W.H., Liao, S.C., and Chang, C.C. (2007). Identification of RNA editing sites in chloroplast transcripts of *Phalaenopsis aphrodite* and comparative analysis with those of other seed plants. Plant. Cell. Physiol. *48*, 362–368.

Zhang, G.Q., Xu, Q., Bian, C., Tsai, W.C., Yeh, C.M., Liu, K.W., Yoshida, K., Zhang, L.S., Chang, S.B., Chen, F., *et al.* (2016a). The *Dendrobium catenatum* Lindl. genome sequence provides insights into polysaccharide synthase, floral development and adaptive evolution. Sci. Rep. *6*, 19029. http://dx.doi.org/10.1038/srep19029

Zhang, J., He, C., Wu, K., Teixeira da Silva, J.A., Zeng, S., Zhang, X., Yu, Z., Xia, H., and Duan, J. (2016b). Transcriptome analysis of *Dendrobium officinale* and its application to the identification of genes associated with polysaccharide synthesis. Front. Plant Sci. *7*, 5. http://dx.doi.org/10.3389/fpls.2016.00005

Zhang, J., Wu, K., Zeng, S., Teixeira da Silva, J.A., Zhao, X., Tian, C.E., Xia, H., and Duan, J. (2013). Transcriptome analysis of *Cymbidium sinense* and its application to the identification of genes associated with floral development. BMC Genomics *14*, 279. http://dx.doi.org/10.1186/1471-2164-14-279

Zhang, L., Chen, F., Zhang, G.Q., Zhang, Y.Q., Niu, S., Xiong, J.S., Lin, Z., Cheng, Z.M., and Liu, Z.J. (2016c). Origin and mechanism of crassulacean acid metabolism in orchids as implied by comparative transcriptomics and genomics of the carbon fixation pathway. Plant J. *86*, 175–185. http://dx.doi.org/10.1111/tpj.13159

Zhao, X., Zhang, J., Chen, C., Yang, J., Zhu, H., Liu, M., and Lv, F. (2014). Deep sequencing-based comparative transcriptional profiles of *Cymbidium hybridum* roots in response to mycorrhizal and non-mycorrhizal beneficial fungi. BMC Genomics *15*, 747. http://dx.doi.org/10.1186/1471-2164-15-747

Zhu, G., Yang, F., Shi, S., Li, D., Wang, Z., Liu, H., Huang, D., and Wang, C. (2015). Transcriptome characterization of *Cymbidium sinense* 'Dharma' using 454 pyrosequencing and its application in the identification of genes associated with leaf color variation. PLOS ONE *10*, e0128592. http://dx.doi.org/10.1371/journal.pone.0128592