The orchid genomic toolkit

Barbara GRAVENDEEL[1,2,*], Diego BOGARÍN[1], Anita DIRKS-MULDER[1,2], Richa KUSUMA WATI[1], Dewi PRAMANIK[1]

[1] Naturalis Biodiversity Center
[2] University of Applied Sciences Leiden Darwinweg 2, 2333 CR Leiden, The Netherlands
[*] **barbara.gravendeel@naturalis.nl**

The application of genomic studies to orchids enables increasingly detailed discoveries of the evolution of both species and organs. We discuss different evolutionary questions that can be addressed using such approaches, and indicate optimal sequencing and data-handling solutions for each case. With this article we hope to promote the diversity of genomic research capitalizing the fascinating natural history of orchids that can now even be completed within the timeframe of a single PhD project.

Why sequence orchid genomes?

Sequencing methods keep diversifying and their costs continue to drop (van Dijk *et al*., 2014). This opens up new perspectives for orchid studies previously hampered by the lack of well-established genetic model species. Nowadays, the **genomes** (see glossary in Box 1) of organisms with no pre-existing genetic resources can be mined quite easily (Matz, 2017). And if optimal approaches are followed, new insights in orchid phylogenomics or gene regulation of for instance floral characters underlying adaptations against pollinators can be obtained in the timeframe corresponding with a single PhD project.

Orchid reference data currently available

Genomic studies are all organized in a similar way: anonymous **reads** are sequenced from an organism, and these reads are subsequently matched with a reference, either a genome or **transcriptome**. The reads can be short and the reference sequences do not need to be very accurate; they should only be sufficiently correct to allow unambiguous matching with reads produced from the study subject. Currently, genome sequences are available for a total of eighteen orchid species belonging to all five subfamilies of which further details are listed under Resources.

Previous orchid phylogenetic studies required labor-intensive marker development coupled with polymerase chain reactions of a single **locus** and **first generation DNA sequencing**. These techniques brought many important first insights in orchid evolution such as the most basal phylogenetic position of the Apostasioideae, followed by Vanilloideae and Cypripedioideae and a more derived position of the Orchidoideae and Epidendroideae subfamilies (Chase *et al*., 2003), the first estimate of the age of origin of the family (Ramirez *et al*., 2007) and the first set of developmental genes responsible for perianth formation as described in the orchid code model (Mondragon-Palomino and Theissen, 2011) that can be considered an expansion of the ABCDE model (Coen and Meyerowitz, 1991) and floral quartet model (Theissen and Saedler, 2001). **Next generation DNA sequencing** techniques, enabling high-throughput parallel sequencing of short DNA molecules of multiple samples, added new insights in the most important drivers of orchid evolution such as epiphytism, evolution of pollinia, and CAM photosynthesis (Givnish *et al*., 2015).

Despite major breakthroughs, though, several nodes in orchid phylogenies remained unresolved with first and next generation sequencing techniques, due to processes such as incomplete lineage sorting, hybridization, or gene duplication, that cause reticulate patterns among the relationships of species. Such cases can now be resolved with **Anchored Hybrid Enrichment** (Figure 1), an innovative next generation sequencing technique, producing data from hundreds of loci of potentially hundreds of individuals for both deep and shallow phylogenetic analyses in a single run (Lemmon *et al*., 2017). For this technique, (i) **probes** are first designed for target enrichment of ca. 500 loci in highly conserved regions in

Box 1. Glossary

Anchored Hybrid Enrichment: a method using conserved probes to recover a large number of innovative phylogenetic markers from chloroplast, mitochondrial and nuclear genomes
Annotation: the process of identifying the locations of the coding regions in a genome and determining what those regions do
Assembly: the process in which short DNA or RNA fragments are merged into longer fragments in an attempt to reconstruct the original sequence
Bioinformatics: the application of computational biology to handle the rapidly growing repository of genomic data
Coverage: the number of times that a given nucleotide in a sequence is sequenced
Datamonkey: a public server for analysis of sequence data
***de novo*:** lacking a reference
First generation sequencing: methods such as the Sanger technique for sequencing short individual DNA and RNA molecules
Genome: the complete set of genetic material present in a cell or organism
GitHub: an open source platform for software development
Locus: a fixed position on a chromosome where for instance a coding region is situated
Next generation sequencing: techniques such as Illumina or Ion Torrent for high-throughput parallel sequencing of short DNA or RNA molecules from multiple samples in a single run
Probe: a small fragment of DNA or RNA binding to a sequence that is complementary to its own
Read: a sequence of base pairs corresponding to a single DNA or RNA fragment
Script: the most basic part of a bioinformatics pipeline and written in a special programming language
Third generation sequencing: techniques such as Single Molecule Real Time (SMRT) sequencing by PacBio or nanopore sequencing by Oxford Nanopore Technologies for sequencing long individual DNA and RNA molecules
Transcriptome: the expressed part of a genome

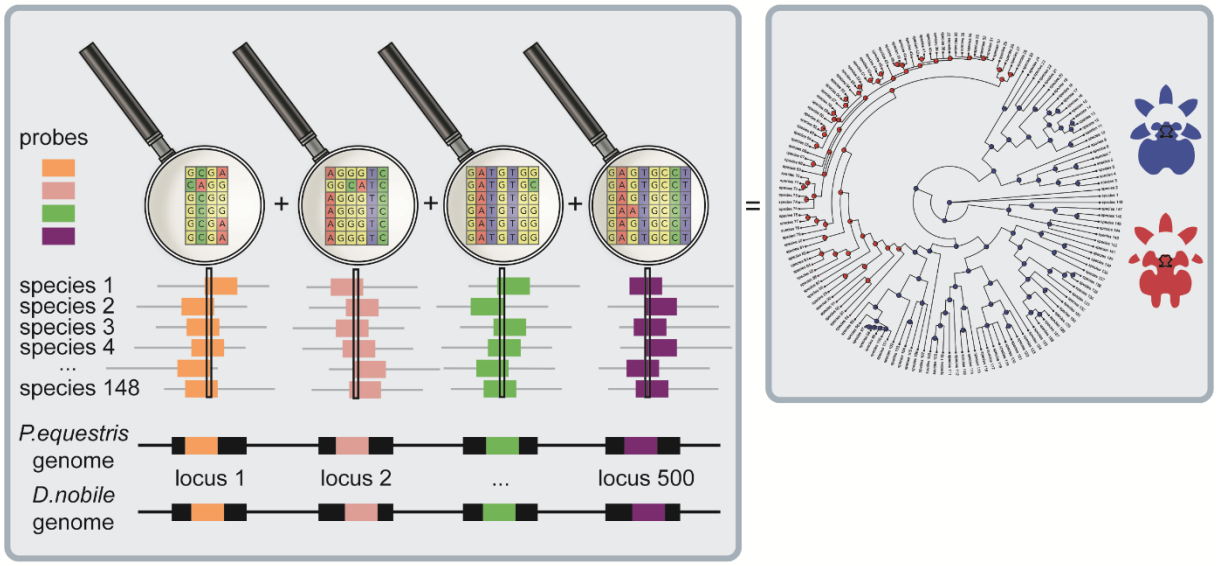

Figure 1. Graphical summary of the Anchored Hybrid Enrichment method. Color legends: black = reference genomes; orange, pink, green and purple = probes; grey = reads obtained from regions captured with probes in the genomes of the study species; blue = flowers with similar shaped sepals; red = flowers with differently shaped sepals. Illustration by Bas Blankevoort.

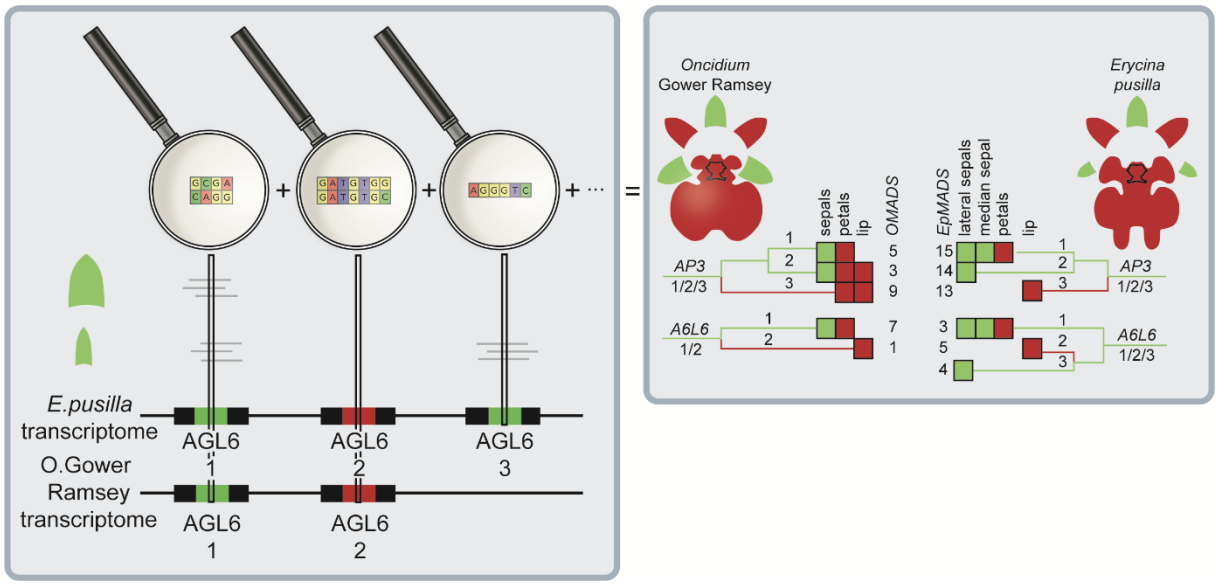

Figure 2. Graphical summary of transcriptome analyses. Color legend: black = transcriptomes; green and red = *AGL6* locus of which *E. pusilla* possesses three copies in its genome but *O.* Gower Ramsey only two; grey = reads obtained from *AGL6* gene copies expressed in differently shaped sepals. Illustration by Bas Blankevoort.

the genomes of reference species for which such data are available like for instance *Phalaenopsis equestris* and *Dendrobium nobile*; (ii) subsequently, enrichment of the genomes of the study species with these probes, called anchored enrichment, capable of recovering a large number of unlinked loci in either the chloroplast, mitochondrial or nuclear genome, is carried out; (iii) loci sequenced from the study species are then processed for **annotation** using an automated **script** in a **bioinformatics** pipeline of which further details are listed under Resources and (iv) used to reconstruct separate gene lineage trees, which can be used to infer a species tree. Due to the large number of gene lineage trees produced from both biparentally inherited nuclear loci and maternally inherited chloroplast and mitochondrial ones, hybridization patterns can be detected and resolved. This technique can therefore increase resolution of shallow orchid clades that remain unresolved with other sequencing techniques (Bogarín *et al*., in press), even when limited or no genomic resources are available for the target species themselves. Another added value of this technique is that DNA extracted from herbarium specimens can be processed as well (Hart *et al*., 2016), increasing sampling options for endangered or rare orchid species. Resolving the resolution of shallow orchid clades opens up the possibility of tracing character state evolution so that evolutionary informative characters can be separated from repeatedly evolving ones. The next challenge is to recover the genetic basis of evolutionary informative characters. This can be done with a transcriptome analysis (Figure 2), which is a cost-efficient alternative to whole-genome sequencing for gene expression studies. Ideally the transcriptome should be collected from the organ of which the genetic basis is sought, preferably harvested from buds and mature tissue, to find out which genes are involved in the early stages of development and which ones in the later stages. The standard way to generate a ***de novo*** transcriptome is to first produce RNA sequences with high **coverage** and then apply (i) data filtering to remove for instance contaminating sequences from endophytic bacteria and fungi, (ii) **assembly**, (iii) quality assessment and (iv) annotation tools using dedicated bioinformatics pipelines, of which further details are listed under Resources.

Massive parallel sequencing of short RNA molecules added important additional insights in the genetic basis of the orchid sepals, petals and lip, as described in the Perianth Code model (Hsu *et al*., 2015; Gravendeel and Dirks-Mulder, 2015) and Oncidiinae model (Dirks-Mulder *et al*., 2017). It also provided the first glimpses of the genetic basis of other orchid organs in the third and fourth floral whorls such as the stamen and stelidia (Dirks-Mulder *et al*., 2017). Ongoing developments in **third generation sequencing** technologies will soon enable retrieval of even higher quality transcriptomes by sequencing longer reads that are expected to ultimately encompass full-length transcripts (Hoang *et al.*, 2017).

Concluding remarks

In the past decade, several new genomic techniques were developed, accompanied by innovative bioinformatics tools. These methods now enable addressing fundamental evolutionary questions for any orchid species within a relatively short timeframe. The last remaining recalcitrant shallow nodes in orchid phylogenies can be resolved relatively fast when applying a team approach as recommended in Box 2. These resolved phylogenies could then be used for tracing character evolution to unravel the full genetic basis of the highly specialized organs that make orchids such fascinating subjects for evolutionary studies.

Box 2. Best Practices for Orchid Genomics

1. Apply different data-filtering settings to ensure robust results and report the exact settings used
2. Share new sequencing data, scripts and other bioinformatics tools with the orchid community in open-access repositories such as **GitHub**
3. Use partially overlapping probes in Anchored Hybrid Enrichment projects so that separate studies can be combined
4. Work on public servers such as **Datamonkey** for analyses that demand a lot of computational power

Resources

http://orchidstra2.abrc.sinica.edu.tw
https://github.com/naturalis/orchids
https://github.com/naturalis/orchid-transcriptome-pipeline/

References

Bogarín D., Pérez-Escobar O.A., Groenenberg D., Karremans A.P., Lemmon A.R., Lemmon A.M., Pupulin F., Smets E.F., Gravendeel B. In press. Anchored hybrid enrichment generated nuclear, plastid and mitochondrial markers resolve the *Lepanthes horrida* (Orchidaceae: Pleurothallidinae) species complex. *Mol. Phyl. Evol.*, 129: 27-47.

Chase, M.W., Barret, R.L., Cameron, K.N., Freudenstein, J.V. 2003. DNA data and Orchidaceae systematics: a new phylogenetic classification. *In*: *Orchid Conservation*. S.P.K. K. W. Dixon, R. L. Barrett and P. J. Cribb (Eds.). Kota Kinabalu, Sabah: Natural History Publications. pp. 69-89.

Coen E.S, Meyerowitz E.M. 1991. The war of the whorls: genetic interactions controlling flower development. *Nature*, 353: 31-37.

Dirks-Mulder A., Butôt R., van Schaik, P, Wijnands J.W.P.M., van den Berg R., Krol L., Doebar S., van Kooperen K., de Boer H., Kramer E.M., Smets E.F., Vos R.A., Vrijdaghs A., Gravendeel B. 2017. Exploring the evolutionary origin of floral organs of *Erycina pusilla*, an emerging orchid model system. *BMC Evol. Biol.*,17: 89.

Givnish T.J., Spalink D., Ames M., Lyon S.P., Hunter S.J., Zuluaga A., Iles W.J., Clements M.A., Arroya M.T., Leebens-Mack J., Endara L., Kriebel R., Neubig K.M., Williams N.H., Cameron K.M. 2015. Orchid phylogenomics and multiple drivers of their extraordinary diversification. *Proc. Roy. Soc. B ser.*, 282: 20151553.

Gravendeel B., Dirks-Mulder, A. 2015. Floral development: Lip formation in orchids unravelled. *Nature Plants*, 1: 15056 .

Hart M.L, Forrest L.L., Nicholls J.A., Kiddner C.A. 2016. Retrieval of hundreds of nuclear loci from herbarium specimens. *Taxon* 65: 1081-1092.

Hoang, N.V., Furtado A., Mason P.J., Marquardt A., Kasirajan L., Thirugnanasambandam, P.P., Botha F.C., Henry R.J. 2017. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics* 18: 395.

Hsu H.F., Hsu W.H., Lee Y.I., Mao W.T., Yang J.Y., Li J.Y., Yang C.H. 2015. Model for perianth formation in orchids. *Nature Plants*, 1: 15046.

Lemmon A.R., Emme S.A., Lemmon, E.M. 2017. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.*, 61: 727-744.

Matz M.V. 2017. Fantastic beasts and how to sequence them: ecological genomics for obscure model organisms. *Trends Genet.*, 34: 121-132.

Mondragon-Palomino M., Theissen G. 2011. Conserved differential expression of paralogous DEFICIENS- and GLOBOSA-like MADS-box genes in the flowers of Orchidaceae: refining the 'orchid code'. *Plant J.*, 66: 1008-1019.

Ramirez S.R., Gravendeel B., Singer R.B., Marshall C.N., Pierce N.E. 2007. Dating the origin of the Orchidaceae from a fossil orchid with its pollinator. *Nature*, 448: 1042-1045.

Theissen G., Saedler, H. 2001. Plant biology: Floral quartets. *Nature*, 409: 469-471.

van Dijk E.L., Auger H., Jaszczyszyn Y., Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet.*, 30: 418-426.