Transcriptome and Proteome Data Reveal Candidate Genes for Pollinator Attraction in Sexually Deceptive Orchids

Khalid E. M. Sedeek1, Wei Hong Qi2, Monica A. Schauer3,4, Alok K. Gupta1,2, Lucy Poveda2, Shuqing Xu1,2*
Zhong-Jian Liu5,6,7, Ueli Grossniklaus3,4, Florian P. Schiestl1, Philipp M. Schlüter1*

1 Institute of Systematic Botany & Zürich-Basel Plant Science Centre, University of Zurich, Zürich, Switzerland, 2 Functional Genomics Centre Zurich, Zürich, Switzerland, 3 Institute of Plant Biology & Zürich-Basel Plant Science Centre, University of Zurich, Zürich, Switzerland, 4 Centre for Model System Proteomes, Zürich, Switzerland, 5 Shenzhen Key Laboratory for Orchid Conservation and Utilization, The Orchid Conservation & Research Centre of Shenzhen, Shenzhen, China, 6 The Centre for Biotechnology and BioMedicine, Tsinghua University, Shenzhen, China, 7 College of Forestry, South China Agricultural University, Guangzhou, China

Abstract

Background: Sexually deceptive orchids of the genus Ophrys mimic the mating signals of their pollinator females to attract males as pollinators. This mode of pollination is highly specific and leads to strong reproductive isolation between species. This study aims to identify candidate genes responsible for pollinator attraction and reproductive isolation between three closely related species, O. exaltata, O. sphegodes and O. garganica. Floral traits such as odour, colour and morphology are necessary for successful pollinator attraction. In particular, different odour hydrocarbon profiles have been linked to differences in specific pollinator attraction among these species. Therefore, the identification of genes involved in these traits is important for understanding the molecular basis of pollinator attraction by sexually deceptive orchids.

Results: We have created floral reference transcriptomes and proteomes for these three Ophrys species using a combination of next-generation sequencing (454 and Solexa), Sanger sequencing, and shotgun proteomics (tandem mass spectrometry). In total, 121 917 unique transcripts and 3531 proteins were identified. This represents the first orchid proteome and transcriptome from the orchid subfamily Orchidoideae. Proteome data revealed proteins corresponding to 2644 transcripts and 887 proteins not observed in the transcriptome. Candidate genes for hydrocarbon and anthocyanin biosynthesis were represented by 156 and 61 unique transcripts in 20 and 7 genes classes, respectively. Moreover, transcription factors putatively involved in the regulation of flower odour, colour and morphology were annotated, including Myb, MADS and TCP factors.

Conclusion: Our comprehensive data set generated by combining transcriptome and proteome technologies allowed identification of candidate genes for pollinator attraction and reproductive isolation among sexually deceptive orchids. This includes genes for hydrocarbon and anthocyanin biosynthesis and regulation, and the development of floral morphology. These data will serve as an invaluable resource for research in orchid floral biology, enabling studies into the molecular mechanisms of pollinator attraction and speciation.


Copyright: © 2013 Sedeek et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the Swiss National Science Foundation (Project 31003A_130796 to PMS), and the University Research Priority Programme Functional Genomics/Systems Biology of the University of Zurich (to UG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The orchids (Orchidaceae) are one of the most species-rich plant families, and their remarkable floral diversity and pollination biology have long fascinated evolutionary biologists [1,2]. It has been estimated that about one third of orchids are pollinated by deception, i.e. without rewarding their pollinators [2,3]. For example, Ophrys L., a Euro-Mediterranean genus of sexually deceptive orchids, is mostly pollinated by male insects, primarily solitary bees [4]. These orchids mimic the visual, tactile, and olfactory signals of the females of their pollinators, so that male insects are attracted and try to copulate with the flower labellum (a modified petal). During these so-called ‘pseudo-copulations’ the pollinia (pollen packets) become attached to the bees and are transferred during subsequent visits of the males to other flowers [5,6,7]. Numerous behavioural studies have shown that the Ophrys-pollinator relationship is highly specific: each Ophrys species is usually pollinated by only one (or very few) insect species [4,5,6,8,9,10]. It has also been shown that floral odour is the key factor in attracting specific pollinators and eliciting male mating...
behaviour [11,12,13]. In addition to odour, flower colour (including UV) and morphology (shape, size and texture) including epidermal structure (e.g. trichomes) also contribute to successful pollination [4,14,15,16,17]. Nonetheless, colour signals are of less importance than floral odour in a group of solitary bee-pollinated species [10] similar to those analysed in the present study.

In Ophrys orchids, floral odour mimics the sex pheromone produced by the female of the pollinators [11,12,19]. This pheromone is a mixture of cuticular alkane and alkene hydrocarbons produced by the flower labellum: specifically alkanes (saturated straight-chain hydrocarbons) with different carbon chain length (C_{21}-C_{31}) and alkenes (monounsaturated hydrocarbons) with different carbon chain length (C_{7}, C_{9}, or C_{12}) [17,20]. The relative amounts of alkanes and alkenes differ significantly among Ophrys species, thereby producing different pseudo-pheromone odour bouquets that attract different species of male bees as their pollinators [13,12,22]. These hydrocarbons are therefore crucial for pollinator-mediated reproductive isolation among Ophrys species [22], and thus play an important role in pollinator-mediated speciation in these orchids [20,21,23,24].

Because of their strong pollinator-mediated reproductive isolation and the relatively well-understood chemical ecology of their highly specific pollination, Ophrys orchids provide an excellent system for studying pollinator-driven speciation and for identifying reproductive ‘barrier genes’ [17,25], that is, genes directly involved in reproductive isolation [26]. Three closely related and sympatric Ophrys species, O. exaltata, O. sphaegodes and O. garganica (Fig. S1) are investigated in this study. They are genetically compatible and crossable, but are strongly isolated from each other by pollinator-mediated, odour-based reproductive isolation, whereas post-pollination reproductive barriers are weak [22]. These species produce different odour bouquets: O. exaltata produces high levels of 7-alkenes, whereas O. sphaegodes and O. garganica produce high levels of 9- and 12-alkenes in different proportions and carbon chain lengths [22]. Therefore, genes underlying these floral odour differences are candidate barrier genes, or possibly even speciation genes, among the study species. Alkanes and alkenes are expected to be derived from very-long-chain fatty acid (VLCFA) biosynthesis in epidermal cells of the flower labellum [17,20,27,28].

Although acyl-ACP (acyl carrier protein) desaturases that introduce a double-bond into alkene precursors have previously been identified as barrier genes among O. sphaegodes and O. exaltata [20,24], other genes responsible for odour differences, such as hydrocarbon chain length differences among O. sphaegodes and O. garganica, are still unknown.

The study of candidate genes in Ophrys orchids has so far been hindered by the lack of sequence resources. Currently, there is no genome sequence publically available for any orchid, and there are no comprehensive genome, transcriptome, or proteome resources for sexually deceptive orchids. A small number of orchid expressed sequence tags (ESTs) obtained by Sanger sequencing are available [29,30,31,32], including 277 ESTs from Ophrys [33]. Although transcriptomes of the Phalaenopsis and Oncidium ‘Gower Ramsey’ orchids have recently been released [34,35], these orchids are from the subfamily Epidendroideae and are only distantly related to Ophrys (subfamily Orchidoideae). Recently, next-generation sequencing such as 454 pyrosequencing has been widely used for de novo sequencing and EST analyses. These technologies have proven effective for expanding the available sequence information not only for model species [36,37] but also for non-model species [38,39] such as Ophrys, the large genome size of which (~10 Gbp [40]) makes transcriptome sequencing a good choice for gene discovery. Moreover, shotgun proteomics by tandem mass spectroscopy (MS/MS) has recently been successfully used for the discovery of the protein components of various biological systems for which no prior information was available [41].

The current study aims to aid progress in orchid biology by (1) uncovering candidate genes for specific pollinator attraction and pollinator-mediated reproductive isolation among three Ophrys species; (2) providing a benchmark reference transcriptome from the orchid subfamily Orchidoideae, and (3) providing the first proteomic data for orchids. We address these questions by use of a systems biology approach, combining high throughput next-generation sequencing technologies (454 pyrosequencing, together with Solexa and Sanger sequencing) and MS/MS shotgun proteomics, in three sexually deceptive Ophrys species.

Results and Discussion

Sequencing and Assembly

Three normalised cDNA libraries were constructed from three different Ophrys species, O. exaltata, O. garganica, and O. sphaegodes. For all three libraries, different tissues (mostly of floral origin) were pooled (see Materials and Methods; Fig. 1). These libraries were 454 pyrosequenced, resulting in a total of 71.3 Mbp of sequence data after processing (Table S1), with approximately 80% of the reads between 100 and 450 bp in length (Fig. 2A). All the high quality reads generated in this study are available in the Sequence Read Archive (SRA) of the National Centre for Biotechnology Information (NCBI) with the accession number SRA060767. Additional sequencing of O. sphaegodes flower labella yielded 1.7 Mbp of Sanger (dbEST library LIBEST_028084; dbEST IDs: 77978749-77979571; GenBank accessions JZ163765-JZ164587) and 2.5 Gbp of Illumina Solexa (SRA060767) data (Table S1).

Solexa reads were assembled into contigs, whereas 454 and Sanger reads of the three species were separately assembled into isoforms (transcripts) (Table 1; Fig. 2B), thereby retaining putative gene/transcript relationships. This process left unassembled high-quality singleton reads, which can be considered to be either rare transcripts or artefacts. However, cross-validation of singleton reads (Table 2) by mapping of Solexa data suggested that the

![Figure 1. Flow chart of transcriptome and proteome analysis.](Image)
majority of singleton reads are not artefacts but represent real transcripts. Pooled 454 and Sanger assembly of all species together increased the effectiveness of the assembly (Table 1). Also, hybrid assemblies combining short (e.g. Solexa) and long (454 or Sanger) reads can improve de novo assembly of genomes and transcriptomes (for average read lengths, see Table S1) [42,43], and short reads can improve coverage and the mining for lowly expressed genes [42]. Therefore, our pooled 454 and Sanger assembly was further merged with the O. sphegodes Solexa assembly to produce a final assembly, which we refer to as the Ophrys reference transcriptome (Table 1). Overall, this assembly contained 51 795 contigs and isotigs, with an average contig/isotig length of 441 bp, and 70 122 singleton reads with an average length of 285 bp (Table 1, Fig. 2C). The proportion of transcripts shared among species was determined by mapping the 454 reads of each species to the assembled contigs and isotigs present in the reference transcriptome (Fig. 3A). The majority of transcripts were shared among three or at least two species (Fig. 3B), the highest number of transcripts being shared among O. garganica and O. sphegodes (Fig. 3A), suggesting a closer overall transcriptome similarity among these species, which is consistent with overall floral similarities (see Fig. S1).

Figure 2. Sequence length distributions. (A) Distribution of 454 read lengths after filtering and adapter removal for the three Ophrys species. (B) Contig length distribution (20 bp windows) for the pooled 454 assembly of the three Ophrys species. (C) Distribution of contig/isotig lengths in the Ophrys reference transcriptome (20 bp windows).

Table 1. Sequence assembly summary.

<table>
<thead>
<tr>
<th>Assembly Type</th>
<th>O. exa. 454</th>
<th>O. gar. 454</th>
<th>O. sph. 454</th>
<th>Pooled 454 + Sanger</th>
<th>Reference transcriptome (Pooled 454 + Sanger + Solexa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of contigs/isotigs</td>
<td>2205</td>
<td>4172</td>
<td>3815</td>
<td>51 230</td>
<td>9175</td>
</tr>
<tr>
<td>Bases in assembly (nt)</td>
<td>1 022 978</td>
<td>2 334 132</td>
<td>1 925 958</td>
<td>19 166 533</td>
<td>5 777 134</td>
</tr>
<tr>
<td>Average contig/isotig length (nt)</td>
<td>463</td>
<td>559</td>
<td>504</td>
<td>616</td>
<td></td>
</tr>
<tr>
<td>N50 contig/isotig length (nt)</td>
<td>498</td>
<td>594</td>
<td>549</td>
<td>616</td>
<td></td>
</tr>
<tr>
<td>Longest contig/isotig length (nt)</td>
<td>1860</td>
<td>2777</td>
<td>1811</td>
<td>4273</td>
<td></td>
</tr>
<tr>
<td>Number of singleton reads</td>
<td>1 35 567</td>
<td>42 261</td>
<td>43 228</td>
<td>290</td>
<td>289</td>
</tr>
<tr>
<td>Average singleton read length (nt)</td>
<td>273</td>
<td>289</td>
<td>289</td>
<td>289</td>
<td>289</td>
</tr>
</tbody>
</table>

Individual assembly statistics are shown for data from different sequencing technologies (454, Sanger, Solexa) and different species (O. exaltata, O. garganica, and O. sphegodes), where ‘Pooled 454’ refers to 454 data pooled from all three species. N50 length denotes the length-weighted median contig length of a given assembly. Singleton reads from 454 and Sanger data only. Out of these, 92 are Sanger reads.
The combination of the 454 transcriptomes of the three species with the EST Sanger sequencing and Solexa data into a reference transcriptome represents the maximum amount of genetic information available to date for Ophrys. In total, 121 917 unique putative Ophrys transcripts were obtained and used for annotation and subsequent analysis. This is considerably more than the published Phalaenopsis orchid transcriptome (42 363 transcripts) [35], and comparable to that of Oncidiuim [34] in terms of contigs (50 908), although the latter transcriptome had a higher number of singleton reads (120 219). We note that our Ophrys reference transcriptome (assembled sequences and annotation provided in Supplementary Files S1 and S2, respectively) was generated mostly from floral tissues. This is the first comprehensive sequence resource both from orchids of the Orchidoideae subfamily, and for sexually deceptive orchids.

Proteomics Results

To obtain a more comprehensive understanding of Ophrys flowers and to further corroborate the authenticity of Ophrys transcripts, we performed large-scale shotgun proteomic analysis using MS/MS. Proteins were extracted from labellum tissue of each study species, digested and subjected to liquid chromatography-tandem mass spectrometry (LC-MS/MS) to obtain three transcripts, we performed large-scale shotgun proteomic analysis

Table 2. Cross-validation of NGS data sets.

<table>
<thead>
<tr>
<th>Reference data</th>
<th>Nref</th>
<th>Mapped by</th>
<th>Nmap</th>
<th>%age (mapping)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solexa-only contigs (O. sph.)</td>
<td>42 493</td>
<td>454 reads (3 spp.)</td>
<td>16 255</td>
<td>38.3%</td>
</tr>
<tr>
<td>454 reads (3 spp.) and Sanger singleton reads (O. sph.)</td>
<td>25 287 (92)</td>
<td>Solexa reads (O. sph.)</td>
<td>15 316 (52)</td>
<td>60.6% (56.5%)</td>
</tr>
<tr>
<td>454 singleton reads (O. exa.)</td>
<td>18 664</td>
<td>Solexa reads (O. sph.)</td>
<td>9752</td>
<td>52.3%</td>
</tr>
<tr>
<td>454 singleton reads (O. gar.)</td>
<td>26 171</td>
<td>Solexa reads (O. sph.)</td>
<td>16 225</td>
<td>62.0%</td>
</tr>
</tbody>
</table>

Summary of the proportion of a given sequence data set to which reads from another NGS data set can be mapped. Nref: number of sequences in the reference data set; Nmap: number of sequences in the reference data set that is mapped by the query data set; %age (mapping): Nmap expressed as a percentage. The term “3 spp. ” refers to data from all three orchid species.

In this row, values in brackets are for Sanger reads.

doi:10.1371/journal.pone.0064621.t002

whereas 16.6% of HQ spectra (275 HQ, 764 PD proteins) also matched proteins in those databases. An additional 204 putative HQ proteins (PD: 887) from 1222 HQ spectra not present in Ophrys transcriptome data (Table 3) could be identified using TAIR and SwissProt databases. Using the FunCat classification scheme [45] for spectra with a hit in the TAIR database, no significant difference (p>0.01; Fisher’s exact test) in any functional category could be found among the entire protein set and proteins present in the Ophrys transcriptome data. No orchid peptides matched translations from different strands of the same transcript (Table 3), further indicating the overall high quality of the assembly.

Like in other published studies [42,46,47,48], far fewer unique proteins (whether HQ or PD) could be identified than transcripts. Possible reasons include the complexity of the proteome and lower protein coverage as compared to transcript sequence data. The overlap among orchid-specific proteins for the three study species (Fig. 3C,D) was similar to the proportions of shared transcripts among species. All in all, proteomic corrobororation of sequence data supports the good quality of our reference transcriptome. Moreover, proteomic data allowed the identification of up to 887 (PD) proteins for which no transcripts were observed, possibly due to a short half-life of the corresponding transcripts.

Functional Annotation

All unique sequences were annotated using BLASTX based on sequence similarity searches against public NCBI non-redundant (nr) and UniProt databases [49,50]. Among all sequences, 44 034 (36.1%; 53.0% of contigs/isotigs and 23.6% of singleton reads) had at least one significant hit to existing genes in the databases. This is nearly twice the number of annotated Phalaenopsis (only 22 234 transcripts; 51.9%) [35] and Oncidiuim sequences (22 810, or 44.8% of contigs; 23 591, or 19.6% of singleton reads) [34]. The remaining 77 883 sequences did not match any known sequences, and may be considered novel transcripts. Alternatively, these sequences may be too short to match known sequences in the databases, or they may be derived from untranslated or nonconserved regions with low homology to known protein sequences, as has been reported in several studies [51,52,53]. We observed that annotation success was positively correlated (R² = 0.41, p<0.001; Fig. S2) with sequence length, similar to data shown by Hoffman [54]. Therefore, long transcripts without BLAST hits are most likely to represent novel genes in Ophrys (Fig. S2).

Possible functions of transcripts with significant BLAST hits were classified for the three main Gene Ontology (GO) functional categories: biological process, molecular function, and cellular component (GO level 2, Fig. 4; levels 1–12, Table S4). The largest number of transcripts (21 138 transcripts) was annotated by
in 454 reads mapped back onto the reference transcriptome. (2527 transcripts, 10.4%).

are 'general function prediction only' (2893 transcripts, 11.9%).

gorous groups (Fig. 4D). The two major functional groups assigned transcripts (20.0%) were assigned to different eukaryotic orthologous groups of proteins (KOG) database [55]. In total, 24,412 proteins and to identify putative orthologues and paralogues, all three orchid species.

Figure 3. Overlap of transcriptome and proteome data among three orchid species. (A) Venn diagram showing the species overlap in 454 reads mapped back onto the reference transcriptome. (B) Bar graph indicating the extent of read sharing. (C) Venn diagram showing the overlap of Ophrys proteomes (HQ data of proteins with corresponding transcripts). (D) Bar graph indicating the extent of proteome overlap among species. (A,C) E, O. exaltata; G, O. garganica; S, O. sphegodes.

doi:10.1371/journal.pone.0064621.g003

Table 3. Summary of proteomics data.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Protein discovery (PD)</th>
<th>Highest quality (HQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O. exa.</td>
<td>O. gar.</td>
</tr>
<tr>
<td>N raw spectra</td>
<td>9676</td>
<td>8419</td>
</tr>
<tr>
<td>N clean spectra</td>
<td>9357</td>
<td>8181</td>
</tr>
<tr>
<td>N unique peptides</td>
<td>4072</td>
<td>3835</td>
</tr>
<tr>
<td>N proteins (transcriptome)</td>
<td>1748</td>
<td>1672</td>
</tr>
<tr>
<td>N proteins (non-transcriptome)</td>
<td>403</td>
<td>405</td>
</tr>
<tr>
<td>Total N proteins</td>
<td>2110</td>
<td>1943</td>
</tr>
</tbody>
</table>

molecular function, followed by biological process (19,960) and cellular component (19,272). In the molecular function category, most transcripts were assigned to ‘binding’ (46.1%) and ‘catalytic activity’ (42.1%). Within biological process, the most abundant categories were ‘metabolic process’ (42.0%) and ‘cellular process’ (35.9%). For cellular components, ‘cell’ and ‘organelle’ had the highest number of transcripts (55.0% and 37.1%, respectively).

To obtain additional annotation information for Ophrys transcripts and to identify putative orthologues and paralogues, all sequences were compared to the eukaryotic clusters of orthologous groups of proteins (KOG) database [55]. In total, 24,412 transcripts (20.0%) were assigned to different eukaryotic orthologous groups (Fig. 4D). The two major functional groups assigned are ‘general function prediction only’ (2893 transcripts, 11.9%) and ‘posttranslational modification, protein turnover, chaperones’ (2527 transcripts, 10.4%).

KEGG Pathways

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) classification system provides an alternative functional annotation for genes according to their association with biochemical pathways [56]. To evaluate the completeness of the Ophrys reference transcriptome, transcripts were assigned to KEGG pathways and to enzyme commission (EC) numbers. A total of 7,394 transcripts were assigned to KEGG pathways (Table 4) and the presence of Ophrys sequences for the majority of enzymes involved in essential biochemical pathways provides further evidence that the Ophrys transcriptome is relatively complete.

Protein Domains

A total of 20,858 unique Ophrys transcripts matched 4,274 protein domains in the Pfam conserved domain database (Tables 5 and S3), with Pkinase, RRM_1 and PPR being the most highly represented classes. Among these, Pkinase represents a conserved protein domain containing the catalytic function of protein kinases, which play roles in various cellular processes including division, proliferation, apoptosis, and differentiation [57]. RRM proteins are the largest group of single strand RNA-binding proteins in eukaryotes that play important roles in RNA processing and protein synthesis regulation [58]. PPR repeat proteins represent the biggest multigene family in plants, and are involved in almost all stages of gene expression, including mRNA transcription, processing, splicing, editing, translation and stability [59]. Among the protein domains of the enzymes putatively involved in hydrocarbon and anthocyanin biosynthesis (Table 5), the AMP-binding and 2OG-FeII_Oxy domains were most highly represented.

Candidate Genes for Pollinator Attraction

Genes involved in VLCFA and hydrocarbon biosynthesis. We combined transcriptome and proteome data sets to create a comprehensive resource for gene discovery in Ophrys orchids, with the aim of identifying candidate genes for pollinator attraction and reproductive isolation among Ophrys species. An important group of such candidate genes is putatively involved in hydrocarbon (and thus VLCFA) biosynthesis [17] (Fig. 5A). While all biosynthetic enzymes in the pathway are potential candidate genes for species-specific pollinator attraction, stearoyl-ACP desaturase (SAD) and β-ketocyl-CoA synthase

Figure 3. Overlap of transcriptome and proteome data among three orchid species. (A) Venn diagram showing the species overlap in 454 reads mapped back onto the reference transcriptome. (B) Bar graph indicating the extent of read sharing. (C) Venn diagram showing the overlap of Ophrys proteomes (HQ data of proteins with corresponding transcripts). (D) Bar graph indicating the extent of proteome overlap among species. (A,C) E, O. exaltata; G, O. garganica; S, O. sphegodes.

doi:10.1371/journal.pone.0064621.g003

Table 3. Summary of proteomics data.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Protein discovery (PD)</th>
<th>Highest quality (HQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>O. exa.</td>
<td>O. gar.</td>
</tr>
<tr>
<td>N raw spectra</td>
<td>9676</td>
<td>8419</td>
</tr>
<tr>
<td>N clean spectra</td>
<td>9357</td>
<td>8181</td>
</tr>
<tr>
<td>N unique peptides</td>
<td>4072</td>
<td>3835</td>
</tr>
<tr>
<td>N proteins (transcriptome)</td>
<td>1748</td>
<td>1672</td>
</tr>
<tr>
<td>N proteins (non-transcriptome)</td>
<td>403</td>
<td>405</td>
</tr>
<tr>
<td>Total N proteins</td>
<td>2110</td>
<td>1943</td>
</tr>
</tbody>
</table>
(KCS) enzymes are obvious a priori candidates for changes in hydrocarbon double-bonds and chain lengths, respectively [17,60], which constitute the main odour differences among our study species. Homologues of all enzymes putatively involved in hydrocarbon biosynthesis were found in the Ophrys reference transcriptome, with 156 unique transcripts (132 putative genes; numbers include singleton reads) representing 20 candidate gene classes (Fig. 5A, Table S5). Seventy per cent (14/20) of these enzymes were confirmed by peptides (Table S6), including SAD and KCS. Proteins that were not confirmed by proteomics data were mostly membrane-associated (e.g., FAD or transporters) or very small (ferredoxin or ACP), and as such are less likely to be detected by shotgun proteomics.

Plant SAD proteins are a class of ubiquitous soluble enzymes that catalyse the insertion of a double-bond into acyl-ACP. Differential expression and/or activity of different orchid SADs is responsible for alkene double-bond differences among O. sphegodes and O. exaltata [20,24]. Homologous to seven known Arabidopsis SAD genes [61], we identified 13 different SAD transcripts in 7 gene models (isogroups) from transcriptome assembly that likely correspond to 6 known Ophrys SAD genes [20,24,33]. This suggests that our knowledge of the identity of Ophrys SAD genes is relatively complete. KCS, a critical component of the fatty acid (FA) elongase (FAE) complex, catalyses the condensation of malonyl-CoA and fatty acyl-CoA to extend the FA by two carbon units [62]. KCS is a good candidate for the differences in alkene chain length between O. sphegodes and O. garganica. Twenty-one KCS-like genes have been annotated in the Arabidopsis genome [63,64,65], and we identified 25 KCS transcripts (24 gene models) in Ophrys. This high number of transcripts indicates ample potential for evolutionary change in KCS-like genes. However, experimental studies are required to test the role of KCS in changing hydrocarbon chain length in Ophrys.

Genes involved in anthocyanin biosynthesis. Anthocyanidin pigments are formed as part of the flavonoid pathway, accumulate in the vacuole of epidermal cells, and can be responsible for red, orange, purple, and blue colours in flowers [66,67]. These pigments play important ecological functions, such as providing visual signals to attract pollinators from a distance. Because of their ubiquity in flowering plants, the biosynthesis and regulation of anthocyanins are well understood, and genes involved in these processes have been
characterised in several orchid species (e.g. [68,69]). While pollinator specificity in *Ophrys* is mostly due to hydrocarbon differences, floral coloration is involved in the mimicry of the pollinator female’s body colour [17]. For instance, *O. garganica* has darker flowers than *O. sphegodes* (see Fig. S1), corresponding to the darker body colour of its pollinator, the black *Andrena pilipes*. With one exception, homologues of all enzymes of the core anthocyanin biosynthesis pathway were found in the *Ophrys* reference transcriptome, with 61 unique transcripts (59 putative genes), representing 7 candidate enzyme classes (Fig. 5B, Table S6). Seventy-one per cent (5/7) of these enzymes were confirmed by peptides (Table S6). Interestingly, the one exception was flavonoid 3’,5’-hydroxylase (*F3’H*), for which no transcripts (or peptides) were found. It is possible that this absence of *F3’H* is not due to limited transcriptome coverage, but reflects the biology of (rather reddish) *Ophrys* flowers, because *F3’H* is required for the formation of (often bluish) delphinidin pigments [66,70]. Similar situations are known in several other plants without delphinidin pigments (e.g., *Ipomoea, Rosa, Dianthus*, and *Chrysanthemum*) [71], in which *F3’H* either is not expressed or was lost from the genome (e.g. in *Arabidopsis thaliana*) [72]. Given that delphinidin pigments may be present in the distantly related *O. speculum* [73], this loss of *F3’H* transcript and/or gene may only have occurred relatively recently.

**Transcription factors.** Transcription factors (TFs) are important regulators of gene expression in response to plant developmental processes and environmental factors [74]. Since they are potential candidates for species differences in pollinator attraction, TFs in the *Ophrys* transcriptome were identified by a comparison to *Arabidopsis thaliana* and *Oryza sativa* transcription factor databases [75] and by KOG annotation. Overall, 3319 unique transcripts (2.7% of the *Ophrys* transcriptome), encoding members of 56 putative TF families, were identified (Fig. 6; Table S7), which is higher than the number of TFs identified in *Phalaenopsis* (786 transcripts, 1.83%). The most abundant TF families were WRKY, NF-YA and NAC factors (Fig. 6). Moreover, LFY, M-type, STAT, VOZ and WOX factors, which had not been found in the *Phalaenopsis* transcriptome [35], were detected in the *Ophrys* transcriptome. Other abundantly represented TF families in our dataset include Myb, bHLH and MADS factors.

MADS-domain (M-type and MIKC) TFs are involved in controlling all major aspects of plant development [76] and have been shown to regulate anthocyanin biosynthesis in pigmented tubers of sweet potato [77]. MADS-domain proteins are important regulators of floral development [78,79], and duplications of B-class MADS-box genes have been implicated in the evolution of complex orchid flowers [80]. For instance, the combinatorial expression of several paralogues is thought to specify flower labellum identity, which may also depend on the action of TCP TFs [80]. In the *Ophrys* transcriptome, MADS-domain factors were represented by 56 transcripts. These included 4 transcripts of GLOBOSA/PISTILLATA and 3 transcripts of DEFICIENS/APER-TALA3 (clade 2, 3 and 4 [80,81]; Table S8) B-class gene homologues. Moreover, 6 TCP TFs were found in the transcriptome. These genes represent candidate genes for labellum development.

The Myb and basic helix-loop-helix (bHLH) families are the largest and second largest classes of TFs in plants, respectively [82]. In rice and *Arabidopsis* there are 339 and 230 Myb factors, and 162 and 111 bHLH TFs, respectively [82]. Myb (Myb and MYB-related) factors account for 197 transcripts in the *Ophrys* transcriptome, whereas the bHLH family is represented by 111 transcripts. Myb TFs are especially interesting candidate genes, because they are implicated in a range of functions, including the regulation of secondary metabolism [83,84,85,86], cell morphogenesis [87,88,89], control of the cell cycle [90,91], floral and seed development [92,93,94,95], responses to biotic and abiotic stresses [96,97,98,99], and light and hormone signalling pathways [100,101,102]. Particularly, Myb factors controlling cell shape as well as VLCFA and anthocyanin biosynthesis may be important candidates for changes in floral traits like odour, colour and epidermal cell shape, all of which are involved in specific pollinator attraction in *Ophrys* [17].

One interesting Myb gene, *MIB30*, has been found to be involved in VLCFA biosynthesis by controlling the expression of

### Table 4. Summary of KEGG pathway analysis.

<table>
<thead>
<tr>
<th>KEGG pathways</th>
<th>KEGG sub-pathways</th>
<th>N transcripts</th>
<th>N enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>Amino Acid Metabolism</td>
<td>892</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>Biosynthesis of Other Secondary Metabolites</td>
<td>252</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Carbohydrate Metabolism</td>
<td>1300</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td>Energy Metabolism</td>
<td>707</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>Glycan Biosynthesis and Metabolism</td>
<td>173</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Lipid Metabolism</td>
<td>393</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Metabolism of Coenzymes and Vitamins</td>
<td>444</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Metabolism of Other Amino Acids</td>
<td>195</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>Metabolism of Terpenoids and Polymers</td>
<td>87</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Nucleotide Metabolism</td>
<td>657</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>Overview</td>
<td>1934</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>Xenobiotics Biodegradation and Metabolism</td>
<td>170</td>
<td>30</td>
</tr>
<tr>
<td>Genetic Information Processing</td>
<td>Translation</td>
<td>84</td>
<td>20</td>
</tr>
<tr>
<td>Environmental Information Processing</td>
<td>Signal Transduction</td>
<td>66</td>
<td>9</td>
</tr>
<tr>
<td>Organismal Systems</td>
<td>Immune System</td>
<td>40</td>
<td>1</td>
</tr>
</tbody>
</table>

The number of *Ophrys* transcripts in a given sub-pathway, as well as the corresponding number of distinct enzymes in the KEGG database are shown. doi:10.1371/journal.pone.0064621.t004
four enzymes forming the FAE complex in *Arabidopsis* [103]. Two transcripts homologous to *MYB30* were identified in *Ophrys* (Table S7). A specific group of Myb factors, R2R3 MYBs, interacts with bHLH factors and WD-repeat (WDR) proteins to form the MYB–bHLH–WDR (MBW) complexes that regulate anthocyanin biosynthesis [70]; R2R3 MYB subgroup 6 (SG6) TFs such as *AtMYB113*, *AtMYB114*, *AtMYB75/PAP1* and *AtMYB90/PAP2* activate anthocyanin biosynthesis in many different species [83,104,105]. Four transcripts represent SG6 genes in the *Ophrys* transcriptome (Table S7).

Conical cells may enhance the colour intensity and brightness of petal surface and, thereby, increase flower attractiveness to the pollinators [106,107,108,109]. MIXTA, a Myb-related TF, drives the formation of conical epidermal cells from the flat epidermal cells of the snapdragon, *Antirrhinum majus* [109,110], and *Amb1* encodes a similar R2R3 MYB TF that has a role in controlling trichome, conical cell and mesophyll cell morphogenesis in the ventral petal of *Antirrhinum majus* [111]. One *Amb1*-like and three *MIXTA*-like genes were identified in *Ophrys* (Table S7).

Taken together, our *Ophrys* reference transcriptome with its focus on flowers allowed us to identify several putative transcription factors that may be relevant to our understanding of the molecular basis of pollinator attraction. This includes MADS, TCP and Myb factors that may be involved in the orchid's mimicry of pollinator females by regulating odour, colour and floral morphological traits. Our data thus provide a starting point for investigating the molecular mechanisms of pollinator attraction and adaptation.

### Table 5. Summary of Pfam domains.

<table>
<thead>
<tr>
<th>Accession ID</th>
<th>Description</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF08097.19</td>
<td>Pkinase Protein kinase domain</td>
<td>507</td>
</tr>
<tr>
<td>PF08076.16</td>
<td>RRM_1 RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)</td>
<td>237</td>
</tr>
<tr>
<td>PF08035.14</td>
<td>PPR pentatricopeptide repeat</td>
<td>220</td>
</tr>
<tr>
<td>PF08040.26</td>
<td>WD40 WD domain, G-beta repeat</td>
<td>203</td>
</tr>
<tr>
<td>PF08071.14</td>
<td>Pkinase_Tyr Protein tyrosine kinase</td>
<td>199</td>
</tr>
<tr>
<td>PF08072.18</td>
<td>RVT_2 Reverse transcriptase (RNA-dependent DNA polymerase)</td>
<td>190</td>
</tr>
<tr>
<td>PF08050.12</td>
<td>p450 Cytochrome P450</td>
<td>148</td>
</tr>
<tr>
<td>PF08054.12</td>
<td>PPR_1 Pentatricopeptide repeat</td>
<td>142</td>
</tr>
<tr>
<td>PF08065.20</td>
<td>rve Integrase core domain</td>
<td>93</td>
</tr>
<tr>
<td>PF08004.23</td>
<td>AAA ATPase family associated with various cellular activities (AAA)</td>
<td>90</td>
</tr>
<tr>
<td>PF08020.23</td>
<td>DEAD DEAD/DEAH box helicase</td>
<td>90</td>
</tr>
<tr>
<td>PF08021.25</td>
<td>Helicase_C Helicase conserved C-terminal domain</td>
<td>80</td>
</tr>
<tr>
<td>PF08015.21</td>
<td>Mitocarr Mitochondrial carrier protein</td>
<td>77</td>
</tr>
<tr>
<td>PF08048.12</td>
<td>PP2C Protein phosphatase 2C</td>
<td>75</td>
</tr>
<tr>
<td>PF08036.21</td>
<td>efhand EF hand</td>
<td>74</td>
</tr>
<tr>
<td>Other domains</td>
<td></td>
<td>18,055</td>
</tr>
</tbody>
</table>

### Protein domains in candidate genes for hydrocarbon biosynthesis

<table>
<thead>
<tr>
<th>Accession ID</th>
<th>Description</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF08051</td>
<td>AMP-binding AMP-binding enzyme</td>
<td>62</td>
</tr>
<tr>
<td>PF08016</td>
<td>ABC2_membrane ABC-2 type transporter</td>
<td>43</td>
</tr>
<tr>
<td>PF08016</td>
<td>adh_short short chain dehydrogenase</td>
<td>35</td>
</tr>
<tr>
<td>PF08037</td>
<td>ECH Enoyl-CoA hydratase/isomerase family</td>
<td>23</td>
</tr>
<tr>
<td>PF08053</td>
<td>Acyltransferase</td>
<td>13</td>
</tr>
<tr>
<td>PF08040.8</td>
<td>FA_desaturase_2 Fatty acid desaturase</td>
<td>13</td>
</tr>
<tr>
<td>PF08039.21</td>
<td>FAE1_CUT1_RppA FAE1/Type III polyketide synthase-like protein</td>
<td>13</td>
</tr>
<tr>
<td>PF08093</td>
<td>NAD_binding_4 Male sterility protein</td>
<td>13</td>
</tr>
<tr>
<td>Other domains</td>
<td></td>
<td>60</td>
</tr>
</tbody>
</table>

### Protein domains in candidate genes for anthocyanin biosynthesis

<table>
<thead>
<tr>
<th>Accession ID</th>
<th>Description</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF08017</td>
<td>2OG_Fell_Oxy 2OG-Fell(II) oxygenase superfamily</td>
<td>63</td>
</tr>
<tr>
<td>PF08021</td>
<td>UDPGT UDP-glucoronosyl and UDP-glucosyl transferase</td>
<td>19</td>
</tr>
<tr>
<td>PF08027</td>
<td>Chal_sti_synt_C Chalcone and stilbene synthases, C-terminal domain</td>
<td>10</td>
</tr>
<tr>
<td>PF08019</td>
<td>Chal_sti_synt_C Chalcone and stilbene synthases, N-terminal domain</td>
<td>7</td>
</tr>
<tr>
<td>PF08043</td>
<td>Chalcone Chalcone-flavanone isomerase</td>
<td>4</td>
</tr>
</tbody>
</table>

Highly abundant protein domains in the *Ophrys* reference transcriptome and among candidate genes for hydrocarbon and anthocyanin biosynthesis. 'Occurrence' lists the number of transcripts matching a given domain.

doi:10.1371/journal.pone.0064621.t005
Evolutionary Implications

Evolution of sexual deception. Sexual deception in *Ophrys* may have evolved from food deception [112], and alkene production (and possibly labellum size) may have served as pre-adaptations for this mode of pollination [112,113]. Floral morphological development, hydrocarbon and anthocyanin biosynthesis play important roles in facilitating sexually deceptive pollination [17]. These pathways, however, are likely to be similar in *Ophrys* and related genera, which share many of its floral features and (pre-)adaptive traits, although with quantitative

Figure 5. Candidate genes in biosynthetic pathways. Schematic diagrams of (A) hydrocarbon and (B) core anthocyanin biosynthesis, indicating candidate protein classes. The number of transcripts for a given candidate gene class is indicated in blue numbers, where an asterisk (*) indicates that a given candidate gene was confirmed at the protein level. Fatty acids in (A) are shown using C:D notation, where C is the number of carbon atoms and D is the number of double-bonds. Greyed, dotted arrows in panel (B) indicate metabolic reactions hypothesised to be absent. Protein abbreviations and further details on the listed candidates are provided in Table S6.

doi:10.1371/journal.pone.0064621.g005
differences (e.g. in alkene levels [113]). This implies that sexual deception likely arose by modification of existing pathways, possibly by regulatory changes or changes associated with gene duplication (and subsequent neo- and subfunctionalisation). Therefore, comparative genomic studies incorporating other orchid species may lead to the identification of the genetic changes that allowed the evolution of sexual deception.

**Molecular mechanisms of reproductive isolation.** All biosynthetic and regulatory candidate genes uncovered in the transcriptome were present in all species, with the exception of \( F^\beta H \) and some transcription factors, all of which were only detected in \( O. sphegodes \) (Tables S6 and S7). However, this may be due to lower sequence coverage in the other two species, which might have prevented detection of rare transcripts. The widespread presence of candidate gene transcripts in all species therefore suggests that species differences in pollinator attraction, and consequently reproductive isolation, are unlikely to be caused by the simple presence or absence of expression in any candidate gene class. Rather, it may be expected that subtler changes, such as paralogue-, isoform-, or allele-specific sequence and/or expression changes, or even epigenetic phenomena [114,115], underlie reproductive isolation. This holds true at least for SAD genes specifying alkene double-bond positions [20,24] and appears similarly plausible for \( KCS \) and hydrocarbon chain length, where the presence of 25 transcript sequences presages complexity. More detailed studies that tease apart the effects of different gene copies, alleles and quantitative expression changes are required to understand the exact molecular architecture of pollinator-mediated reproductive isolation among \( Ophrys \) species.

**Conclusions**

Here, we employed high throughput next-generation sequencing technologies combined with shotgun proteomics to provide the first floral reference transcriptome and proteome data for sexually deceptive \( Ophrys \) from the Orchidoideae subfamily of orchids, for which proteome data were previously absent and sequence information scarce. We thereby provide significant resources for gene discovery and systems biology in orchids in general, by enabling sequence comparisons among disparate lineages of orchids. Likewise, our data set considerably expands the resources available for sexually deceptive plants, and provides an opportunity to advance our understanding of the molecular basis of plant-pollinator interactions, as well as pollinator-mediated selection and speciation. Our data are relevant for the characterisation of candidate genes for these and other processes, as demonstrated here by the identification of genes potentially involved in plant pseudo-pheromone biosynthesis and regulation. Based upon a priori knowledge and the abundance of these gene classes in the \( Ophrys \) transcriptome, especially \( KCS \) biosynthetic genes and \( Myb \) transcription factors warrant further attention as candidate genes for differences in specific pollinator attraction. Subtle changes in such genes may be responsible for reproductive isolation among sexually deceptive orchids.

**Materials and Methods**

**Plant Materials**

Plant material of \( O. exaltata \) subsp. \( archipelagi \) (GOLZ & REINHARD) DEL PRETE, \( O. garganica \) NELSON, and \( O. sphegodes \) MILLER for 454 sequencing was grown in a greenhouse at the Botanic Garden of the University of Zurich, Switzerland. Additional flower labellum samples of these three species were used for proteome analysis and two further \( O. sphegodes \) flower labella for EST library and Illumina Solexa sequencing, all of these samples being from different plants. All flowers used in this study were unpollinated. Plant tissues were collected, flash-frozen in liquid nitrogen and stored at \(-80^\circ C\) until RNA or protein extraction. For 454 sequencing, the following tissues were used: leaves, bracts, labella, sepals, petals and columns, both from open flowers and flower buds from all developmental stages available on the sampled plants, the smallest...
available bud being five positions from the latest open flower on the inflorescence.

cDNA Normalisation and 454 Sequencing
Total RNA was extracted separately from different tissues (see above) collected from three different individual plants per species (21 RNA samples/species) using TRIzol® reagent (Invitrogen) and the supplier’s protocol. Extracts were further purified by using RNeasy MinElute Cleanup columns (Qiagen) according to the manufacturer’s protocol. The extracted RNAs were analysed for potential degradation by gel electrophoresis and on a Bioanalyzer 2100 (Agilent), and RNA concentration was quantified using the fluorometric Qubit Quantitation Platform (Invitrogen). Equal amounts of RNA from each biological individual were pooled to yield one RNA sample for each Ophrys species. To avoid genomic DNA contamination, RNA was treated with RNase-free DNase I (Qiagen). Full-length double-stranded cDNA was synthesised from 0.5 μg pooled total RNA using SMARTer PCR cDNA Synthesis Kit (Clontech, Palo Alto, CA, USA) according to the manufacturer’s instructions. To enhance gene discovery, the contribution of highly abundant transcripts was reduced before sequencing. To do so, 1 μg of each cDNA was normalised using the Trimmer cDNA Normalization Kit (Eurogentec, Moscow, Russia), according to the manufacturer’s instructions.

Approximately 500 ng of normalised cDNA of each sample were used to generate a single strand cDNA transcriptome library for the Roche/454 Life Sciences GS-FLX Titanium platform (Roche, Basel, Switzerland) following the Rapid Library Preparation Method Manual. Briefly, cDNA of each sample was sheared by nebulisation to produce fragments of approximately 300–400 bp, and oligonucleotide adapters were ligated to the fragmented cDNA. One adapter contained a barcode sequence that was used to discriminate the samples (i.e. species) from each other after sequencing, as all libraries were combined in a single pool. Transcriptome library sequencing was then performed according to the Roche GS-FLX XLR70 Titanium emPCR and sequencing manuals. The pooled sample was sequenced on a full picotitre plate on a Genome Sequencer FLX Instrument, according to the manufacturer’s instructions, at the Functional Genomics Centre Zurich, Switzerland.

EST Library Preparation and Sequencing
Poly (A)⁺ RNA was purified from 2 μg O. sphegodes flower labellum total RNA using the Oligotex mRNA isolation kit (Qiagen AG, Hombrechtkon, Switzerland) according to the manufacturer’s instructions. One standard cDNA library was prepared using the Creator SMART cDNA Library Construction Kit and the protocol provided, except that insert size selection was performed using Zymoclean Gel DNA Recovery Kit (Zymoclean, Orange, CA, USA). Oligo-dT-primed cDNA inserts larger than 1 kb were directionally cloned in pDNR-LIB vector (Clontech, Palo Alto, CA, USA) and transformed into XL-10 Gold (Stratagene, LaJolla, CA, USA). Colony PCR reactions were performed to test the library efficiency and insert size range. The library was stored by the addition of glycerol (20% v/v final concentration) and sent for sequencing (Applied Biosystems) at the Purdue University sequencing platform, West Lafayette, Indiana (USA).

Illumina Solexa Sequencing
Total RNA of one O. sphegodes flower labellum was used for an RNA-Seq experiment by Illumina Solexa sequencing at BGI Shenzhen (China), following Illumina’s sample preparation guidelines. Briefly, poly(A)⁺ was purified from total RNA and fragmented, cDNA synthesised and adapters ligated. After size selection, cDNA was subjected to Solexa paired-end sequencing on an Illumina Genome Analyzer II generating 75 nt long reads.

Processing of Sequence Reads
For Sanger reads, base calling, masking of vector sequences and low quality ends were done with phred (version 071220) via the phredPhrap script (version 080818) in consed (version 20.0) [116], and trimming of poly-A/T tails with seqclean (http://compbio.dfci.harvard.edu/tgi/software/). Raw Solexa reads were filtered by quality using the manufacturer’s software and default parameters. Raw 454 sequencing data was obtained with the GS Run Processor 2.5.3 (Roche) using shotgun quality filtering and trimming as defined by the default settings. Raw 454 reads were first trimmed of adapter sequences used in cDNA library preparation and normalisation with newbler in the Roche 454 Software Suite 2.5.3 (454 Life Science, Branford, CT). MEGA-BLAST 2.2.21 [117] was run to ascertain that trimmed reads were clean of any adapter and primer sequences. Trimmed 454 reads were further processed with seqclean to remove low complexity and low quality reads, and to remove any left-over poly-A/T tails.

Transcriptome Assembly
O. sphegodes Solexa reads were first assembled into contigs using SOAPdenovo [118], using paired-end information to link contigs into scaffolds, and where possible, to fill gaps with reads. Scaffolds were clustered using TGICL [119], each cluster producing one or more consensus sequences; unclustered scaffolds were termed singleton contigs. O. sphegodes Sanger reads and 454 reads from all three species were both (1) assembled separately for each species and (2) pooled and assembled with 454 newbler 2.5.3 (454 Life Science, Branford, CT). The pooled assembly was further merged with the O. sphegodes Solexa assembly using minimus2 under the criteria of minimum 40 bases overlap with at least 94% identity [120]. The merged assembly from Sanger, 454 and Solexa data is referred to as the Ophrys reference transcriptome.

Transcriptome and Proteome of Ophrys Orchids
Identification of Candidate Genes

A priori candidate genes were identified by homology to genes of known or putative function in model organisms. To identify hydrocarbon and anthocyanin biosynthetic genes, a BLASTN search of the Ophrys reference transcriptome against Arabidopsis thaliana TAIR9 coding sequences was performed (at e-value of <e-03), retaining information of the best BLAST hit for each orchid transcript. Homologues of Arabidopsis genes for selected gene classes were considered candidate genes. In addition, text/term searches of BLAST/NR best hits, GO annotations and EC numbers were used to identify candidate orchid genes. For transcription factor identification, KOG searches for transcription terms were performed alongside BLAST searches against A. thaliana and Oryza sativa TF databases publicly available from PlantTFDB 2.0 (which currently does not contain sequences from Asparagales) [75]. TFs of special interest, putatively involved in the regulation of VLCFA, anthocyanin biosynthesis and flower morphology, were identified from the literature.

Shotgun Proteomics

Frozen labellum tissue of an unpollinated flower from each study species was ground to powder without allowing it to thaw, and resuspended in 150 μL urea protein extraction buffer (65 mM Tris-HCl, 8 M urea, 10% glycerol, 5% β-mercaptoethanol, 2% SDS, 0.05% bromophenol blue), denatured and separated by SDS-PAGE, sliced (13 gel slices per sample) and trypsinised as described previously [129]. Proteins were subjected to electrospray ionisation-based LC-MS/MS analysis with a 2D linear ion trap Finnigan LTQ (Thermo Electron Corporation), equipped with an Ultimate Nano HPLC System (Dionex Corporation) exactly as described by Grobei et al. [129].

Proteome Analysis

MS/MS-derived spectra were searched against different peptide databases using Mascot Search Engine version 2.3 (Matrix Science Ltd., UK). The databases used were SwissProt and TAIR9 [130,131], including decoys and known contaminants (as in Grobei et al. 2009 [129]), and (1) orchid peptides predicted by ESTScan and (2) a 6-frame translation of the orchid transcriptome. Mascot searches, with a peptide mass tolerance of 3 Da, allowed for one trypsin miscleavage, for Met oxidation and Cys carboxymethylation, and were further analysed and validated in R 2.14.2 [132] and Scaffold 3.3 (Proteome Software Inc., USA), which uses the Peptide- Prophet and Protein-Prophet algorithms [133,134]. Spectra from known contaminants were removed from the data set for final analysis.

Supporting Information

Figure S1 Comparison of Ophrys flowers. Five flowers each from different individuals of (A) O. exaltata subsp. archipelago, (B) O. sphagodes and (C) O. garganica, showing inter-species and intra-species variation (plants all from Gargano, Southern Italy). Images were scaled for comparison, the white bar indicating 1 cm. O. sphagodes tends to have comparatively small flowers with a brown labellum and a greenish perigon, whereas O. garganica flowers are usually larger, with a darker labellum and sometimes coloured petals. O. exaltata tends to have comparatively large flowers with a slightly elongated brown labellum, typically with a protrusion at its apex, and usually a white perigon. The speculum (brighter, more reflective part of the labellum) can be quite variable in all species, and they all have longer trichomes at the sides of the labellum (‘hairy margin’) as compared to its centre. Micromorphological features of Ophrys flowers are described elsewhere [13].

Figure S2 Transcript length/annotation relationship. Plot showing the percentage of hits with annotation information from NCBI nr and UniProt databases versus sequence length. (PDF)

Table S1 Summary of sequencing data. This table lists details of the 454, Sanger and Solexa sequence data sets obtained in this study. (DOCX)

Table S2 List of orchid proteins. This table lists all proteins matching sequences in the Ophrys transcriptome by their transcriptome sequence ID. In the column HQ, an asterisk (*) denotes proteins in the HQ data set (no asterisk indicates PD data set). The ‘Species’ column indicates in which orchid species a given protein was found, where: E, O. exaltata; G, O. garganica; S, O. sphagodes. In the column ‘Orchid-only’, y/n (yes/no) indicate if a given protein was found only in the Ophrys transcriptome (y) or if it also matched proteins in the TAIR and/or SwissProt databases (n). The ‘Description’ column refers to the term provided in the transcript ID’s best BLASTX/nr database hit. (XLSX)

Table S3 List of proteins without transcript. This table lists all proteins that were identified in the SwissProt/TAIR9 databases, but did not match any sequence in the Ophrys transcriptome, listing the protein description, source database and associated accession numbers. In the column HQ, an asterisk (*) denotes proteins in the HQ data set (no asterisk indicates PD data set). The ‘Species’ column indicates in which orchid species a given protein was found, where: E, O. exaltata; G, O. garganica; S, O. sphagodes. (XLSX)

Table S4 GO classification. This table contains the GO classification of the assembled sequences into three main categories (biological function, cellular component and molecular function) at all levels (1–12). ‘All species’ lists the number of transcripts in a given category in the Ophrys reference transcriptome, and additional columns list the corresponding number of transcripts for the individual species separately. (XLS)

Table S5 Pfam protein domains. This table lists the Pfam annotations in the Ophrys reference transcriptome by transcript ID. (XLSX)

Table S6 List of candidate genes. Details of transcripts encoding candidate biosynthetic proteins putatively involved in hydrocarbon and anthocyanin biosynthesis, sorted alphabetically per category. TAIR ID lists the Arabidopsis gene for which homologues were found; EC is the enzyme commission number searched for a given candidate protein; ‘N transcripts’ lists the number of unique transcripts matching a candidate protein category in the Ophrys reference transcriptome (counting singleton reads as transcripts); ‘N gene models’ is composed of the number of isogroups from sequence assembly plus transcripts that were not assigned to isogroups (so that for instance, singleton reads would be counted as a new gene model); ‘Unique transcripts’ lists the unique transcriptome sequence IDs. Species with transcripts and peptides list in which Ophrys species transcripts or peptides were found, respectively, where: E, O. exaltata; G, O. garganica; S, O. sphagodes. In the column ‘Species with peptides’, an asterisk indicates that a protein was part of the HQ data set (no asterisk, PD data set). The ordering of elements in the two rightmost
columns corresponds to the ordering in the ‘Unique transcripts’ column.

(DO CX)

**Table S7** List of transcription factors. This table lists the identification method (TFDB, transcription factor database; KOG, KOG annotation) for TF identification, along with the unique transcript ID and annotation details (accession number or KOG term). The column ‘Transcription Factor’ lists the TF class identified from the database or the KOG description term, as appropriate. Species with transcripts or peptides for a given TF are identified from the database or the KOG description term. The column ‘Lineage’ lists the gene lineage (clade 1, 2, 3, and 4; Orchidoideae GLO1 and GLO2) assigned to the accession number of a BLAST hit, as defined by Mondragón-Palomino & Theißen (2008; [81]) for DEF, and Kim & al. (2007; [135]) and Cantone & al. (2011; [136]) for GLO. The best lineage assignment is highlighted in bold for each transcript. Accession numbers with defined lineage are as follows: DEF clade 1: AY196350, AY378148, FJ804111; clade 2: AY196350, AY378149, FJ804097, FJ804106, FJ804115; clade 3: AY378150, AB232663, DQ198389, FJ804099, FJ804107, FJ804110, FJ804117; clade 4: AY378147, FJ804108, FJ804112, FJ804116; GLO1: AB2922665, AB450305, AB450307, AB450310, AB450302, AB450308, AB450309, AB450299, AB450303, AB450306; GLO2: AB232664, AB375712, AB375707, AB375711, AB375713, AB375709, AB375706, AB375704, AB375710. (XLSX)

**File S1** This zip-compressed file contains (1) the assembled Ophrys reference transcriptome (FASTA), along with (2) an MD5 check-sum, (3) an FAI index file, (4) 454 isogroup (transcript to gene) mapping (tab-delimited text), and (5) a text file describing the naming scheme for the transcript identifiers used. (ZIP)

**File S2** This spreadsheet document contains the annotation information for the Ophrys reference transcriptome in several tables. (XLSX)

**References**

1. Darwin CR (1877) The various contrivances by which orchids are fertilised by insects. London: John Murray, 2nd ed.

**Acknowledgments**

The authors thank Natalia Dudareva and the Purdue Sanger sequencing platform, Bernd Roschitzki and Simon Barlow for help with protoen bioinformatic, Wen-Clu Tsi for providing their Phalaenopsis transcriptome data, Salvatore Cozzolino for discussion, and Stephen Schauer and two reviewers for discussion and comments.

**Author Contributions**

Conceived and designed the experiments: SX ZJL UG FPS PM. Performed the experiments: KEMS MAS ACG LP SX. Analyzed the data: KEMS WQ MAS PJMS. Wrote the paper: KEMS PJMS. 454 data set: Conceived and designed the experiments: SX ZJL UG FPS PM.Performed the experiments: KEMS MAS ACG LP SX. Analyzed the data: KEMS WQ MAS PJMS. The authors thank Natalia Dudareva and the Purdue Sanger sequencing platform, Bernd Roschitzki and Simon Barlow for help with protoen bioinformatic, Wen-Clu Tsi for providing their Phalaenopsis transcriptome data, Salvatore Cozzolino for discussion, and Stephen Schauer and two reviewers for discussion and comments.

Transcriptome and Proteome of Ophrys Orchids

PLOS ONE | www.plosone.org 13 May 2013 | Volume 8 | Issue 5 | e64621


