The genome of an aneuploid cultivar facilitates elucidating the genetic bases of trait polymorphisms in *Phalaenopsis* orchids

Ruidong Jia[1‡]; Xinyi Zuo[2,3‡]; Liting Zou[2‡]; Li Wang[1]; Letian Lin[2]; Zhiqing Wang[2,9]; Yuexia Zhang[2]; Xingyu Chen[1]; Fanrui Meng[8]; Hantang Huang[10]; Lan Lan[11]; Zhimei Li[2]; Fei Wang[2,3]; Yuxuan Jin[2,12]; Hongyan Shan[4,5]; Rui Zhang[3]*; Hongzhi Kong[4,5,6]*; Peipei Wang[2,7]*

[1] State Key Laboratory of Vegetable Biobreeding, Key Laboratory of Biology and Genetic Improvement of Flower Crops (North China), Ministry of Agriculture and Rural Affairs, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[2] Kunpeng Institute of Modern Agriculture at Foshan, Foshan, Guangdong 528200, China

[3] College of Horticulture, Northwest A&F University, Yangling, Shaanxi 712100, China

[4] State Key Laboratory of Plant Diversity and Specialty Crops, Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China

[5] China National Botanical Garden, Beijing 100093, China.

[6] University of Chinese Academy of Sciences, Beijing 100049, China.

[7] Guangdong Huabo Ecological Industry Co., Ltd., Foshan, Guangdong 528241, China

[8] Bioinfo-Reach Bioinformatics Service, Guangzhou, Guangdong 510663, China

[9] College of Plant Protection, South China Agricultural University, Guangzhou, Guangdong 510642, China

[10] College of Horticulture, China Agricultural University, Beijing 100193, China

[11] State Agricultural Biotechnology Centre (SABC), College of Science, Health, Engineering and Education, Murdoch University, Perth, Western Australia, 6150, Australia

[12] Shanghai Jiaotong University, Shanghai, 200240, China

[‡] These authors contributed equally to this work

Authors for correspondence:

Peipei Wang

wangpeipei02@caas.cn

Hongzhi Kong

Email: hzkong@ibcas.ac.cn

Rui Zhang

ruizhang@nwafu.edu.cn

Abstract

As an important ornamental crop, *Phalaenopsis* orchids exhibit extraordinary diversity in floral traits with unclear genetic bases. To address this problem, we assemble a haplotype-resolved, chromosome-level genome for an aneuploid cultivar 'Santiago'. We reveal pervasive variation in the number and sequence among homoeologous genes (HGs) and demonstrate proportional expression dosage effects of both chromosome and HG counts. This genome facilitates deciphering the genetic variation that contribute to trait polymorphisms in a hybrid population, some of which are also observed among HGs in *P.* 'Santiago'. Specifically, nucleotide polymorphisms in a region that includes two tandemly arranged *PsAGL6-2* genes account for lip morphology variation. Presence/absence of *PsMYB12* genes determines presence/absence of venation-associated stripes. The *b*-type HG of *PsMYB2* is associated with the pink color of sepals/petals. *PsMYBx1* represses anthocyanin biosynthesis in multiple regions, in addition to inter-spot regions. Our study advances the genetic understanding of trait diversity in *Phalaenopsis*.

Main

Phalaenopsis orchids are a major ornamental crop prized for their beauty and long flowering. Thanks to the rapid development of tissue culture technology, the production and market scale of *Phalaenopsis* have expanded significantly. Cross-breeding has yielded numerous cultivars to meet the growing market demand, which display extraordinary diversity in horticultural traits, particularly in floral size, scent, organ morphology, and color patterning. However, the growing homogenization caused by cross-breeding necessitates the urgent integration of advanced biotechnologies—such as genome editing, molecular- or genomics-assisted breeding—to develop novel, high-value cultivars. This requires comprehensive understanding of the genetic mechanisms underlying key horticultural traits and their diversity.

Key genes regulating specific traits have been identified, such as floral scent[1,2], color[3], venation-associated stripes[3], spot[3], petal size[4], and cuticular wax[5]. The genetic bases of the presence/absence (P/A) variation of floral scent[1,6] and the variation in spot/patch patterns[3,7,8] have also been studied. However, except for these two traits, the genetic mechanisms underlying variation in many other important traits remain largely uncharacterized. For example, while the background pink color on sepals/petals and venation-associated stripes are known to be regulated by *PeMYB2* and *PeMYB12*, respectively[3], the specific genetic variation responsible for the P/A of these two traits across different cultivars are unclear. Similarly, the extensive polymorphisms in lip morphology raises the question of whether and what specific changes of its key identity regulator, *PeAGL6-2* (*AGAMOUS-LIKE 6-2*)[9], lead to the transformation into petal-like organs in 'Bigfoot' cultivars. To decipher the genetic bases of trait diversity, it is essential to associate allelic polymorphisms—including single nucleotide polymorphism (SNP), insertion and deletion (indel), copy number variation (CNV), and structural variant (SV)—with trait variation. A major bottleneck in *Phalaenopsis* is the lack of a chromosome-level reference genome; the existing genomes for *P. equestris* and *P. aphrodite* are only at the scaffold levels[10,11]. Consequently, the full scope of these allelic polymorphisms and their contributions to trait diversity is still poorly understood.

In this study, we assemble a haplotype-resolved, chromosome-level genome for a highly heterozygous aneuploid cultivar, *P.* 'Santiago' (ANTHURA BV, with US patent US20150047084P1, hereafter *P.* sa, **Fig. 1A**). Analysis of this genome reveals extensive variation in the number and sequence among homoeologous genes (HGs), particularly those

77 of genes known to regulate key floral traits. To understand the genetic mechanisms
78 underlying trait diversity in *Phalaenopsis*, we generate a hybrid population (*P*. '1071-13' × *P*.
79 'CH1804-01', named as H80) which exhibits extensive polymorphisms in lip morphology
80 and floral color patterning (**Fig. S1**). By analyzing the whole genome resequencing data of
81 this population using *P*. sa genome as the reference, we decipher genetic variation that
82 account for these trait polymorphisms. Some of these genetic variation are also observed
83 among HGs in the *P*. sa genome. In addition, we demonstrate multiple functions of *PsMYBx1*
84 in repressing anthocyanin biosynthesis in diverse regions, which exhibits variation in HG
85 count across cultivars as well. Our study highlights the usefulness of a single genome with
86 high heterogeneity in elucidating the genetic bases of key trait diversity. It also facilitates
87 marker-assisted breeding in *Phalaenopsis* by offering critical insights for molecular marker
88 development.

89 Results

90 The genome assembly and annotation of *P.* sa

91 The whole genome size of *P.* sa was estimated to be approximately 4.1 Gb by flow cytometry,
92 and 3.7 and 3.3 Gb by *k*-mer-based analysis using Genomic Character Estimator and
93 GenomeScope2, respectively (**Fig. S2A,B**). *K*-mer frequency analysis suggested an allo-
94 triploid genome (**Fig. S2C**), while karyotype analysis revealed 70 chromosomes per cell,
95 indicating likely aneuploidy (**Fig. S2D**) since a haplotype (1x) is supposed to contain 19
96 chromosomes[11]. The final assembly (**Methods**) comprised 70 pseudo-chromosomes, with 6 and
97 13 homoeologous chromosome groups (HChrGs) representing triploid (3x, set3) and
98 tetraploid (4x, set4) sets, respectively (**Fig. 1B,C**). The total assembly size was approximately
99 3.55 Gb, with much higher quality exceeding those of two other available *Phalaenopsis*
100 genome assemblies (**Table S1**). Predicted telomere sequences were identified at most
101 chromosome ends (**Fig. 1D**), further suggesting the relatively high assembly quality. Due to
102 the complex hybrid history, the *P.* sa genome did not display obvious phases of subgenomes
103 (**Fig. 1C, Fig. S3**). Synteny analysis revealed extensive structural variation between
104 homoeologous chromosomes (HChrs) (**Fig. 1D**), including approximately 140 inversions,
105 1079 translocations, and 513 segmental duplications (**Table S2**).

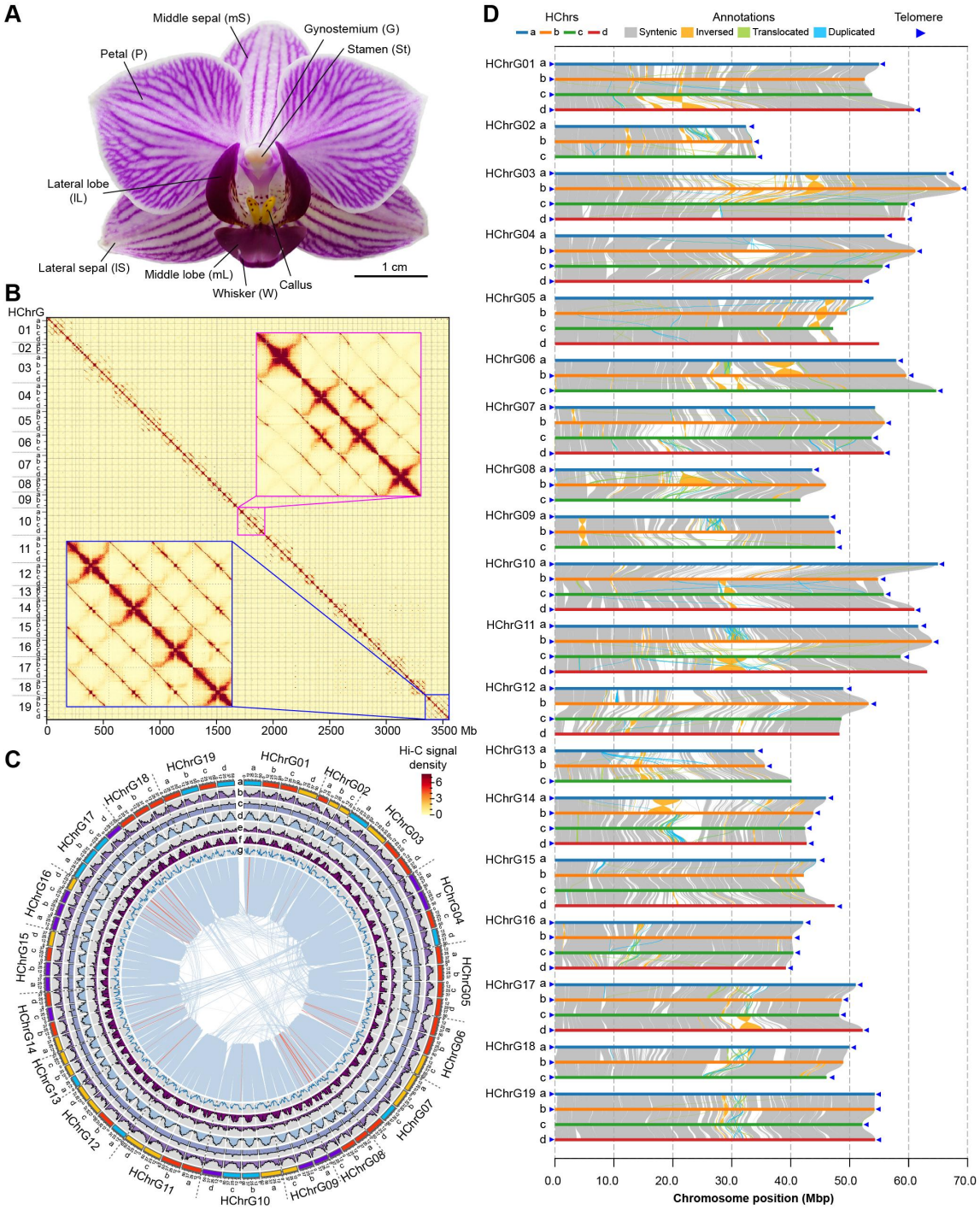

Fig. 1 Genome assembly of *Phalaenopsis* 'Santiago' (*P.* sa). (**A**) A *P.* sa flower has one middle sepal, two lateral sepals, two petals, one lip, and a gynostemium that connects the stamen with two pollinia and the ovary. (**B**) Hi-C heatmap showing chromosomal interactions, with bin size of 250 Kb. Two insets illustrate weak interaction signals between homoeologous chromosomes (HChrs) of HChr group 10 (HChrG10, pink) and HChrG19 (blue) as examples. (**C**) Circos plot displaying genomic features within 1.5-Mb bins in *P.* sa. a, chromosome sizes, with different colors indicating predicted subgenomes, which is inferred from **Fig. S3**. b, gene density, ranging from 0 to 99. c, GC content, 15.38%–53.56%. d, transposable element coverage, 37.86%–99.79%. e, Copia LTR retrotransposon coverage, 0%–17.73%. f, Gypsy LTR retrotransposon coverage, 0%–98.51%. g, tandem repeat density, 0–251. The center region shows syntenic blocks between (blue lines) or within (orange lines) chromosomes. (**D**) Collinearity among homoeologous chromosomes, with predicted telomeres indicated.

107 We predicted 93,708 protein-coding genes using *de novo*, homology-based, and
108 transcriptome-supported approaches (**Methods**, **Table S3**). Repetitive sequences constituted
109 67.97% (2.41 Gb) of *P. sa* genome (**Table S4**). Among these, long terminal repeats (LTRs)
110 represented 41.91% of the genome, consisting of Copia (4.18%), Gypsy (37.22%), and
111 unknown types (0.51%), while tandem repeats accounted for 2.48%. The peak LTR insertion
112 time was estimated at approximately 0.31 million years ago (**Fig. S2E**). Additionally, the
113 genome annotation included 308 microRNAs, 1,094 transfer RNAs, 8,135 ribosomal RNAs,
114 and 34,366 small nuclear RNAs (snRNAs) (**Table S5**).

115 Homoeologous gene variation in *P.* sa

116 During the genome annotation, we noticed substantial variation in HGs. According to
117 syntenic analysis (**Methods**), all the 93,708 genes were classified into 33,697 HG groups
118 (HGGs), which were further categorized into six types according to variation in the number
119 and sequence among HGs (**Fig. 2A**). Regarding variation in sequence, to facilitate the
120 analysis, we examined the Pfam domain constitutions of the encoded protein sequences
121 among HGs, with the assumption that proteins with different domain constitutions are likely
122 to perform divergent functions[12,13]. Type 1 contained 12,673 HGGs (37.6% of groups, 13.5%
123 of genes) with only single HG per gene; type 2 consisted of 7,839 HGGs (23.3%/21.8%) with
124 HChr-specific gene losses while ≥2 HGs were retained; type 3 included 1,964 HGGs
125 (5.8%/18.4%) with HChr-specific tandem duplicates; type 4 comprised 373 HGGs
126 (1.1%/3.3%) with same number of tandem duplicates per HG; type 5 contained 3,642 HGGs
127 (10.8%/12.8%) with conserved single HG per homoeologous locus (HL) but these HGs have
128 different protein domain constitutions; type 6 included 7,206 HGGs (21.4%/30.2%) with
129 single HG per HL which had conserved domain constitutions (**Fig. 2A**).

130 Genes in types 1-6 differed in terms of expression patterns, functional annotation (**Methods**)
131 and encoded protein length (**Fig. 2B,C**), and tended to belong to different transcription factor
132 (TF) families (**Fig. 2D**) and perform different biological functions (**Fig. 2E**). Specifically,
133 type 1 genes tended to be not expressed (Transcripts Per Kilobase [TPM]=0, 45.1%) or lowly
134 expressed (TPM≤2, 30.6%) in all 28 examined samples (**Table S6**), have no functional
135 annotations and have relatively short proteins (**Fig. 2B**). Among 50 type 1 genes that were
136 annotated as potential TFs, 13 were *FAR-RED IMPAIRED RESPONSE 1 (FAR1)* genes (**Fig.
137 2D**). These 13 *FAR1* genes had TPM≤2 in all examined samples (**Fig. S4**), indicating they
138 were undergoing pseudogenization and might be lost from the genome in the long term.

139 *FAR1* genes are essential for phytochrome signaling and balancing plant growth and defense
140 responses under shade conditions[15]. The deletion of certain *FAR1* genes from *Phalaenopsis*
141 genomes may be results of adaptation to shady, damp forest understory. In contrast, another

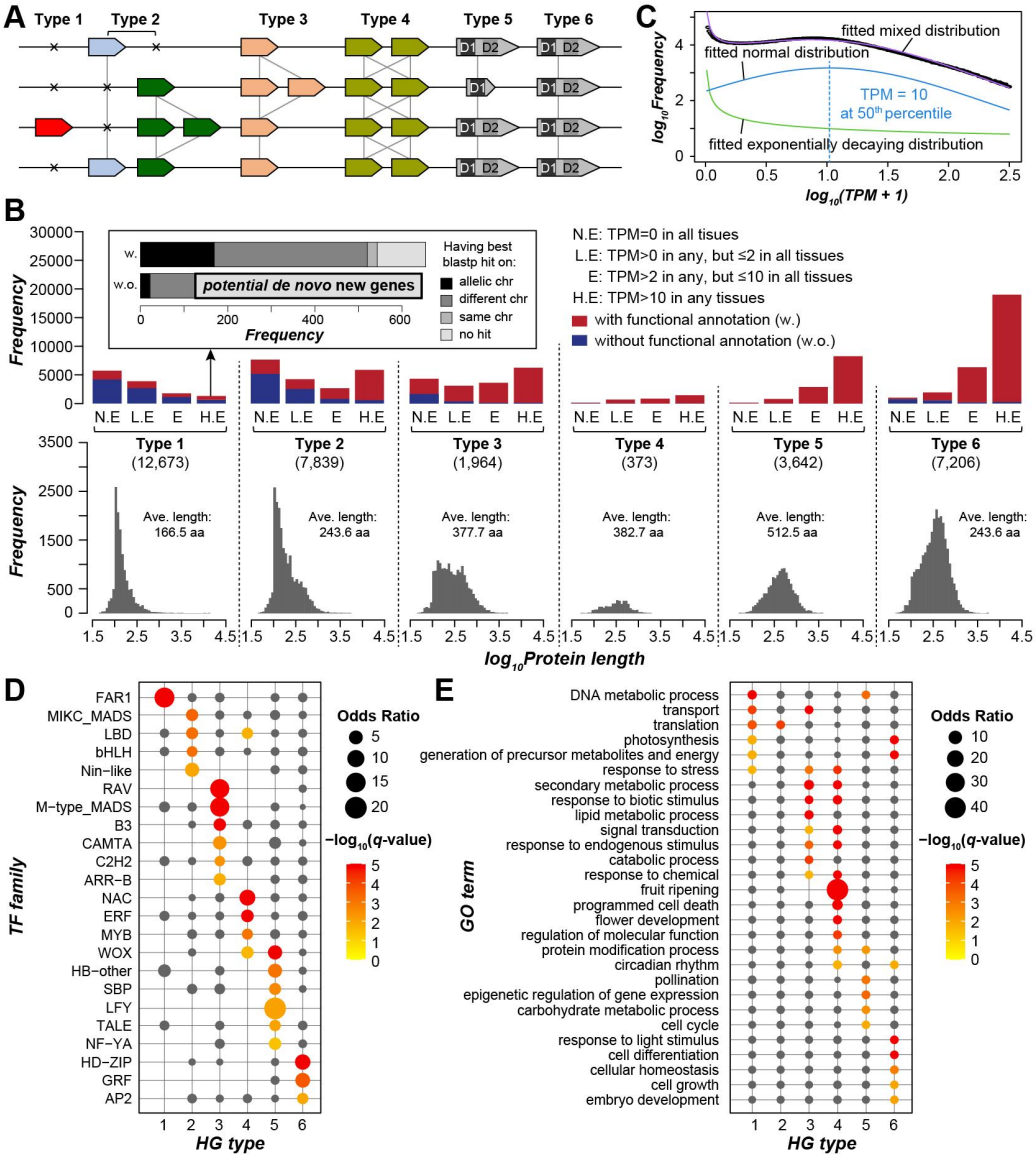

Fig. 2 Homoeologous gene variation in *P.* sa genome. (**A**) Diagram illustrating the six HG types. D1 and D2: protein domain 1 and domain 2. (**B**) Frequency of six HG types (upper panel) with the number of HGGs listed within the parentheses, and their protein length distribution (lower) with the average length per type indicated as well. Box colors indicate genes with or without functional annotations. TPM: Transcripts Per Kilobase; N.E.: not expressed; L.E.: lowly expressed; E.: expressed; H.E.: highly expressed. Inset: frequency of highly expressed (TPM>10) type 1 genes with (w.) or without (w.o.) functional annotation. Different gray scales indicate whether a gene has the best non-self blastp hit on different, homoeologous or same chromosomes, or has no blastp hit. (**C**) Mixed distribution of gene expression levels, illustrating how TPM=10 was selected as the cutoff threshold for high expression (see **Methods**). Fitted normal distribution indicates true expression, while fitted exponentially decaying distribution indicates noise expression. (**D,E**) Enriched transcription factors (**D**) and biological_process GO terms (**E**) in six HG types based on Fisher's exact test. Dot size: odds ratio; dot color: -$\log_{10}$ *q*-values which are the adjusted *p*-values using the "BH" method[14]; cases with *q*-values≥0.05 or odds ratio<1 were
142 colored gray.

143 519 type 1 genes were highly expressed (TPM>10, 10.3%) in certain tissues, but had no
144 functional annotation and no non-self blastp hit within *P.* sa genome. These genes might be
145 potential *de novo* new genes, as 46.8% (243) of them were preferentially expressed in stamen
146 (**Fig. S5A-C**), consistent with previous findings that new genes tended to arise from male
147 reproductive organs[16,17]. The remaining 276 genes were preferentially expressed in other
148 tissues (**Fig. S5D-G**), such as middle root (*Psa_17aG011940*), leaf (*Psa_16dG007910*), shoot
149 apical meristem and inflorescence stalk (*Psa_04bG014720*), suggestive of their potential
150 functions in the corresponding tissues.

151 Type 2 genes also tended to have short proteins (**Fig. 2B**), and be enriched for TFs that play
152 important roles in floral development (MIKC-type MADS-Box), cell proliferation,
153 determination and differentiation (Basic Helix-Loop-Helix, bHLH), lateral organ formation
154 and responses to environmental stimuli (Lateral Organ Boundaries Domain, LBD), and
155 nitrate responses (Nin-like) (**Fig. 2D**). Type 3 genes tended to have longer proteins, be
156 enriched for TFs involved in flowering, embryogenesis (such as M-type MADS), stress
157 responses and signaling (Related to ABI3 and VP1 [RAV], Calmodulin-binding transcription
158 activator [CAMTA]) (**Fig. 2D**). GO terms enriched in type 3 included those participating in
159 interaction with environments, such as secondary metabolic processes, response to biotic and
160 endogenous stimuli, and signal transduction (**Fig. 2E**). Type 4 genes tended to be enriched in
161 TFs and functions related to plant growth and stress responses, including NAM, ATAF1/2
162 and CUC2 (NAC), ETS2 repressor factor (ERF), MYB and WUSCHEL-related homeobox
163 (WOX) families (**Fig. 2D**), and biological functions in fruit ripening, flower development,
164 responses to biotic and endogenous stimuli (**Fig. 2E**). Genes in types 5 and 6 had even longer
165 proteins, were enriched for TFs responsible for and GO terms related to essential and
166 conserved functions (**Fig. 2B,D,E**), such as LEAFY (LFY) in type 5 and GROWTH-
167 REGULATING FACTOR (GRF) in type 6, and biological processes including pollination
168 (type 5) and cell differentiation and growth (type 6). These results indicate that the HChr-
169 specific gene duplicates/losses (types 2,3), together with those tandem duplicates retained in
170 all HLs (type 4), may have driven the high diversity of floral traits and the adaptation of
171 *Phalaenopsis* to diverse environments.

172 Dosage effects of homoeologous gene count

173 Considering the aneuploidy nature of and the substantial HG count variation in *P*. sa genome,
174 we asked whether the chromosome ploidy (3x vs 4x) and HG count variation affect the total

175 expression amount (summed expression levels of all HGs) per gene. To avoid the influences
176 of tandem duplicates and gene structural variation on the total expression amount per gene,
177 we focused on HGGs which had a single gene copy per HL if the HG is present [such as
178 0:0:1:1 (2/4), 1:1:1 (3/3)], with no differences in protein domain constitutions among HGs
179 (C1-C8, **Fig. 3A**). In addition, to ensure a valid comparison, our analysis was restricted to
180 reciprocal best-matched (RBM) gene pairs that differed in chromosome ploidy or HG count.
181 This strategy was adopted because the gene expression in the hybrid parents of *P.* sa were not
182 available.

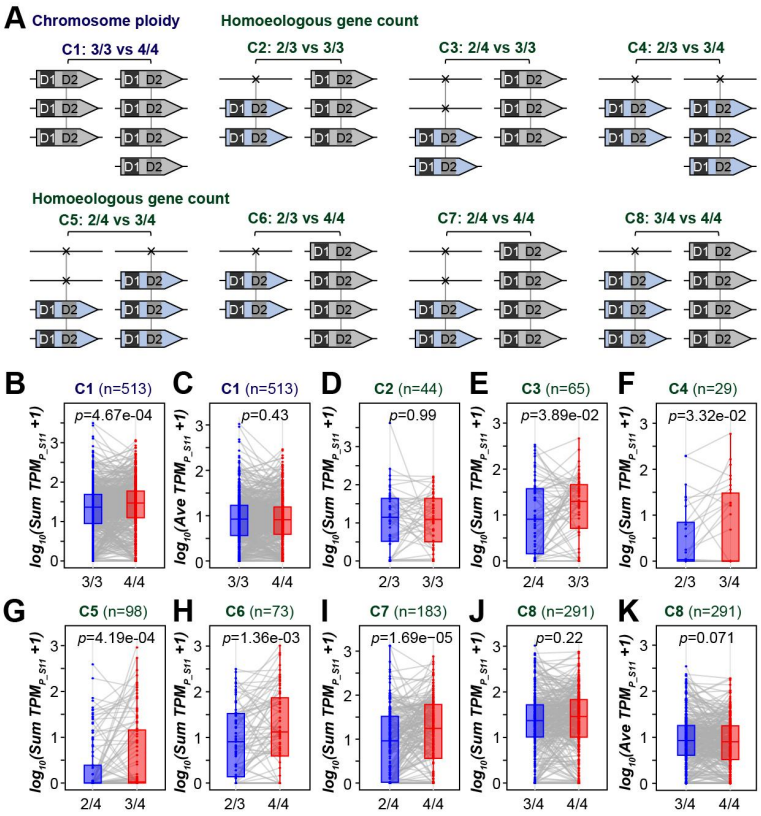

Fig. 3 Dosage effects of chromosome ploidy and homoeologous gene count on gene expression. (**A**) Diagrams showing eight comparisons, where C1 compares reciprocal best-matched (RBM) gene pairs located on HChrGs with different ploidies (3x vs 4x), while C2-C8 compare RBM genes with different HG counts. 2/3 and 2/4: only two HGs were retained for HChrGs with 3x and 4x, respectively; 3/4: three HGs were retained for HChrGs with 4x. (**B,C**) Summed (**B**) and averaged (**C**) HG expression levels in petals at S11 of RBM gene pairs between HChrGs with 3x or 4x. (**D-J**) Summed HG expression levels for RBM gene pairs in C2-C8. (**K**) Averaged HG expression levels for RBM gene pairs in C8. Regarding (**B-K**), see **Dataset 1** for expression results in other tissues.
183

184 We found that gene expression scaled with chromosome ploidy: summed expression levels of
185 type 6 set4 genes (4/4) were generally higher than those of their 3/3 RBM counterparts
186 (p=4.67e-04, two-sided Wilcoxon rank sum test, **Fig. 3B**), but became comparable when
187 averaged per chromosome count (p=0.43, **Fig. 3C**). This dosage effect was similar to those

188 found in other aneuploid organisms, where gene expression in aneuploids were compared with
189 their corresponding diploids, such as maize[18], budding yeast[19] and *Candida albicans*[20]. HG
190 counts also had dosage effects on gene expression for all but two comparisons (C3-C7, **Fig.**
191 **3E-I**): summed expressions of HGs with HChr-specific gene losses were generally lower than
192 those of their RBM counterparts with more HGs. However, the expression tended to be
193 compensated when HG from only one HL was lost (C2, C8): no significant differences in
194 summed expression levels between 2/3 (HG from one out of the three HLs was lost) and 3/3
195 genes (p=0.99, **Fig. 3D**), and between 3/4 and 4/4 genes (p=0.22, **Fig. 3J**). This was further
196 supported by the significantly higher averaged levels of 3/4 genes than 4/4 genes (p=0.071,
197 corresponding to p=0.035 of one-sided Wilcoxon rank sum test, **Fig. 3K**). These results
198 suggest that buffering mechanisms counteract the disruption of homeostasis caused by
199 immediate HChr-specific gene losses (i.e., 2/3 and 3/4 genes), but not that caused by
200 chromosome inequivalence.

201 Genetic bases of key trait polymorphisms

202 The extensive HG variation in *P*. sa were mainly derived from allelic polymorphisms of its
203 hybrid parents, which represented part of the entire polymorphism set of *Phalaenopsis* that
204 contribute to the trait diversity. To evaluate the utility of *P*. sa genome in revealing the
205 genetic mechanisms underlying those diversity, we investigated a hybrid population (H80)
206 that was generated in our lab and displayed substantial variation in key floral traits,
207 particularly in lip morphology and color patterning (**Fig. 4A, Fig. S1**). Different trait
208 categories can combine randomly, resulting in a unique phenotype for each individual in this
209 population, making it a highly representative set of *Phalaenopsis* orchids. We resequenced
210 the genomes of 46 H80 representative individuals and performed genome-wide association
211 study (GWAS) using *P*. sa genome as the reference.

212 Key regulators of these traits have been previously identified, including *PeAGL6-2* in lip
213 organ identity determination[9], *PeMYB2*, *PeMYB12* and *PeMYB11* in the formation of floral
214 background color, stripes and spots, respectively[3], and *MYBx1* in repressing anthocyanin
215 accumulation in non-patched regions[7,8]. Expression profiles of these genes (see **Table S7** for
216 their gene IDs in the *P.* sa genome) reflected their relatively conserved functions in *P*. sa and
217 H80 (**Fig. 4B, Fig. 5A,B**). Nevertheless, these genes displayed different types of HG
218 variation in *P.* sa (**Fig. 4C, Fig. 5C, Fig. 6A**). Therefore, we focused on identifying genetic
219 variants within or near these genes that were significantly associated with corresponding trait

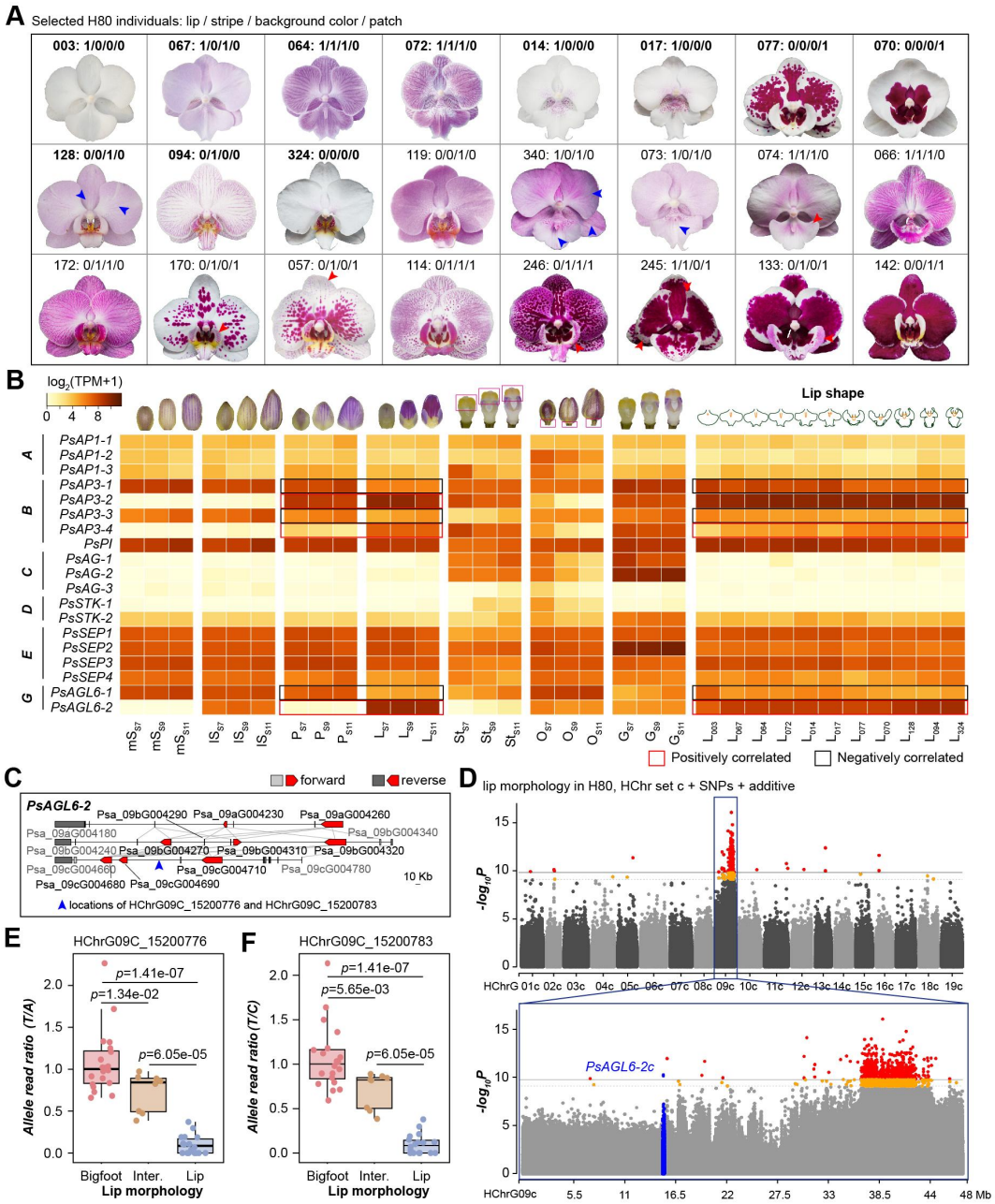

Fig. 4 Floral trait variation in H80 and the genetic basis of lip morphology variation. (**A**) Phenotypes of 24 H80 individuals (see **Fig. S1** for all 46 individuals). The category of each individual in lip morphology (0: lip, 1: petal-like organ), stripe P/A (0: A, 1: P), background color (0: white, 1: pink), and spot/patch types (0: spotted or non-spotted, 1: patched) were indicated. Bold font: individuals with transcriptome data of the lip. Red arrowhead indicates stripe, while blue arrowhead indicates stripe-like shadow. (**B**) Heatmap showing expression profile of A-G classes of MIKC-type MADS-box genes in different floral organs at three stages (S7, S9 and S11), and in the S9 lips of 11 representative H80 individuals. Color scale indicates log-transformed expression levels. mS: middle sepals; lS: lateral sepals; P: petals; L: lip; St: stamen; O: ovary; G: gynostemium. Red and black boxes: expression of genes positively or negatively correlated with petal-to-lip transitions, respectively. (**C**) HG variation of *PsAGL6-2*. Red boxes: *PsAGL6-2* genes with directions indicated; light or dark gray boxes: neighboring genes with forward or reverse directions, respectively; boxes linked using gray lines between HChrs: HGs. The blue arrowhead indicates locations of two significant SNPs. (**D**) Manhattan plot for GWAS of lip morphology in H80, using SNPs referred to HChr set c with the additive marker-effect model. Horizontal solid and dashed lines indicate -$\log_{10}$ (0.01/variant counts) and -$\log_{10}$ (0.05/variant counts) (**Methods**), respectively; red and orange dots denote SNPs significantly associated with lip morphology variation; blue dots represent SNPs located within or adjacent (±2 kb) to the *PsAGL6-2* gene clusters. (**E,F**) Allele read ratio of two SNPs (count of reads supporting "T" to count of reads supporting "A" in **E**, and count of reads supporting "T" to that of "C" in **F**) within H80 individuals with different lip morphologies. *P*-values are from two-sided Wilcoxon rank-sum tests.

221 polymorphisms, and decoding how they contributed to trait differentiation in H80. Since H80
222 individuals were shown to be aneuploid as well by karyotype analysis (**Fig. S2F**, 2n=64, >3x),
223 they were treated as tetraploids in the GWAS to facilitate the analysis. Moreover, given that
224 70 chromosomes in *P.* sa can not be readily phased into subgenomes (**Fig. 1C**), the original
225 HChr designations (such as that HChrG01a~HChrG19a were designated as HChr set a) were
226 retained when serving as the reference genome during the GWAS.

227 **Lip morphology variation and nucleotide polymorphisms nearby *PsAGL6-2* genes**

228 *PsAGL6-2* had two to four tandemly arranged duplicates on three HChrs (corresponding to
229 HG type 3, **Fig. 4C**), with only one copy encoding a full-length protein per HChr (**Fig. S6**).
230 GWAS analysis for the lip morphology (lips vs petal-like organs) identified a number of
231 significant variants (including SNPs, indels and SVs) on the right arms of HChrG09s,
232 regardless of which HChr set was used as references (**Fig. S7-9**). However, only when the
233 HChr set c was used as the reference (**Fig. 4D, S9A**), two significant SNPs were identified in
234 the intergenic region between two *PsAGL6-2* copies (**Fig. 4C**, ~65 Kb upstream of
235 *Psa_09cG004690* and ~86 Kb downstream of *Psa_09cG004710*), which were located on the
236 left arms. The allele read ratios (the number of reads supporting one HG divided by the
237 number of reads supporting the other HG) of these two SNPs were significantly associated
238 with lip morphology in H80 (**Fig. 4E,F**, **Table S8**). Furthermore, expression levels of
239 *PsAGL6-2* showed positive correlation with lip morphology. These results suggest that the
240 transformation of lips into petal-like organs in H80 did not result from loss-of-function
241 mutants in *PsAGL6-2*, as it remained expressed at substantial levels in these organs. Instead,
242 the morphological change is associated with reduced expression of *PsAGL6-2*, potentially
243 due to the introduction of an allele containing mutations in a long-distance enhancer (likely in
244 the intergenic region between *Psa_09cG004690* and *Psa_09cG004710*). Additionally,
245 expression of other proposed lip [*PsAPETALA3-4* (*PsAP3-4*)] and petal identity genes
246 (*PsAP3-1*, *PsAP3-3*, and *PsAGL6-1*; see **Fig. S10** for the gene tree of MADS-box genes)[21]
247 exhibited positive and negative correlations with petal-to-lip transitions, respectively (**Fig.
248 4B**). These results suggest that expression of these genes might be regulated by *PsAGL6-2*,
249 thereby coordinating lip morphogenesis.

250 **Presence/absence of stripes and presence/absence of *PsMYB12* genes**

251 As the regulator of stripe, expression of *PsMYB12* genes was highly associated with stripe
252 formation in *P.* sa and the presence of stripes in H80 population (**Fig. 5A,B**). It lost its entire
253 tandemly duplicated gene cluster from HChrG04b (HG type 2, **Fig. 5C**), indicating the P/A

of stripes in *Phalaenopsis* is potentially determined by *PeMYB12* P/A. CNV-based GWAS (**Fig. 5D, Methods**) using HChr set a as the reference identified significant variants located in or nearby *PsMYB12* genes while with the fewest false positives compared with others (**Fig. S11**). Consistent with this, stripes P/A in H80 was perfectly associated with P/A of resequencing reads mapped to *PsMYB12* genes (**Fig. 5F**, **Fig. S12, Table S9**), further confirming that stripe P/A are determined by *PeMYB12* P/A.

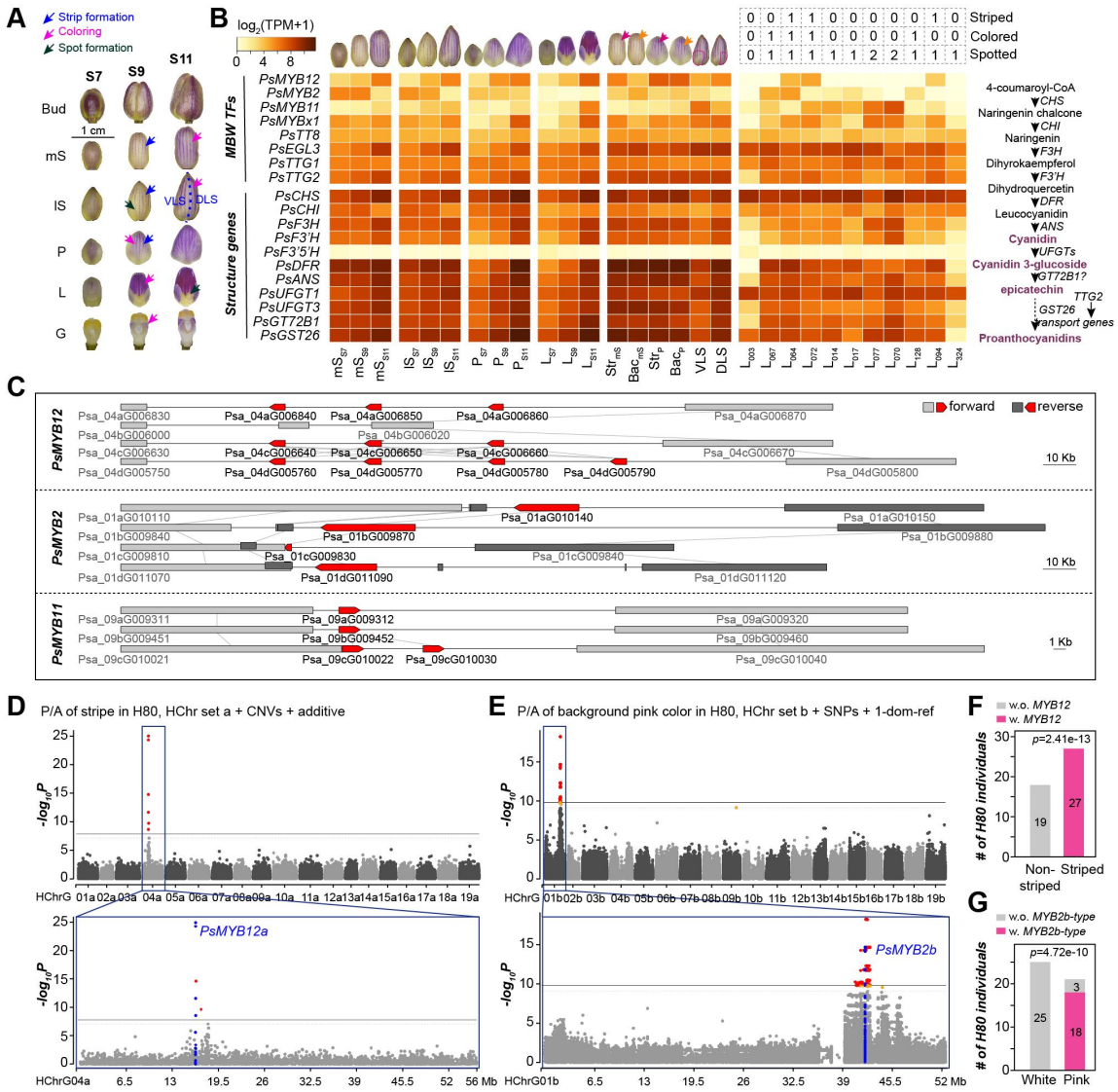

Fig. 5 Genetic bases of color patterning variation in H80. (**A**) Floral organs at three developmental stages showing important events of color patterning. VLS and DLS: ventral and dorsal half of the lateral sepals, respectively. (**B**) Heatmap showing expression profiles of genes participating in color patterning. The classifications of each H80 individual in the presence (0) or absence (1) of stripes, background pink colors and spots (2 indicates patches) are listed. MBW TFs: transcription factors that form the MYB-bHLH-WD40 complexes. The biosynthesis pathway for the prominent anthocyanin in *Phalaenopsis* orchids—cyanidin—is shown on the right. Str: stripe; Bac: inter-stripe background region. (**C**) HG variation of *PsMYB12*, *PsMYB2* and *PsMYB11*. (**D**) Manhattan plot for GWAS of stripe P/A in H80, using CNVs referred to HChr set a with the additive marker-effect model. (**E**) Manhattan plot for GWAS of background color in H80, using SNPs referred to HChr set b with the 1-dom-ref marker-effect model. (**F,G**) Association of P/A of *PsMYB12* genes (**F**) and P/A of *PsMYB2b*-type sequences (**G**) with stripe P/A and background color, respectively. *P*-values are from Fisher's exact test.

261 Since we didn't identify any genes orthologous to *PsMYB12* in the other eight orchid species
262 (including the non-striped *P. aphrodite*) except for *P. equestris* (**Fig. S13A**), we asked when
263 and how the *PsMYB12* genes and stripes originated in *Phalaenopsis*. Interestingly, stripes
264 were only observed for the branch leading to *P. equestris* and *P. lindenii* in wild
265 *Phalaenopsis* species (**Fig. S13B**). To test if *PsMYB12* genes originated before or after the
266 most recent common ancestor (MRCA) of *Phalaenopsis*, we resequenced the *Vanda*
267 'Pakchong Blue', an orchid cultivar outside *Phalaenopsis* which displays venation-associated
268 stripes on the perianth as well. We detected reads mapped to *PsMYB12* genes, albeit with a
269 high degree of divergence within and around these genes (**Fig. S13C**). Given that other
270 orchids like *Paphiopedilum* also have stripes, these results suggest that *PsMYB12* genes
271 predate the MRCA of *Phalaenopsis* and have undergone repeated, independent losses in
272 *Phalaenopsis* species. Alternatively, these genes may be absent in the MRCA of
273 *Phalaenopsis* and later reintroduced into the MRCA of *P. equestris* and *P. lindenii*,
274 potentially via gene introgression.

275 **Background pink color and the *b*-type HG of *PsMYB2***

276 *PsMYB2* had a single HG per HL (**Fig. 5C**), but the one on HChrG01c (*Psa_01cG009830*)
277 had lost its R2 domain (HG type 5, **Fig. S14A**) and was expressed at much lower levels than
278 other HGs (**Fig. S14B**). Instead, *PsMYB2b* (*Psa_01bG009870*) was expressed at the highest
279 levels among its HGs (**Fig. S14B**). GWAS using 1-dom-ref model (**Methods**) with HChr set
280 b as the reference identified significant SNPs and indels located in or nearby *PsMYB2* genes
281 while with the fewest false positives among others (**Fig. 5E**, **Fig. S15-17**). In this 1-dom-ref
282 model, the *PsMYB2b*-type HG was considered as dominant in determining the pink color.
283 Consistently, the P/A of *PsMYB2b*-type sequences can nearly perfectly predict the P/A of
284 background pink color for all but three individuals—H80-066, H80-072 and H80-114 (**Fig.
285 5G**, **Table S10**). The background color in these exceptions may be governed by other non-
286 *PsMYB2* genes, since *PsMYB2* expression was almost undetectable in the lip of H80-072 (the
287 summed TPM of four HGs was 1.32, **Fig. 5B**). Given that the protein sequence of PsMYB2b
288 differed from PsMYB2a by only two amino acids (**Fig. S14A**), these results suggest that the
289 pink color in *P.* sa and most H80 individuals with pink flowers was likely determined by the
290 HChr set b-specific expression of *PsMYB2*, which was driven by its distinct promoter (**Table
291 S10**). It requires further investigation of which region of the *PsMYB2b* promoter drives its
292 HG-specific expression.

293 **Spot/patch pattern variation**

294 *PsMYB11* had an extra tandem copy (*Psa_09cG010030*) on HChrG09c (HG type 3, **Fig. 5C**),
295 which harbored a few mutations in protein sequences and the upstream sequences (**Fig.**
296 **S14C,D**), compared with other HGs. *Psa_09cG010030* was highly expressed in the lip of *P.*
297 sa at S11 where the spots formed, and in lips of H80-077 and H80-070 which had patches
298 instead of spots (**Fig. 4A**, **Fig. S14E**). GWAS for the P/A of patches identified numerous
299 variants distributed across the entire genome, but no variants were near the *PsMYB11* genes,
300 regardless of variant types used and HChr sets referred to (**Fig. S18-20**). Given that the patch
301 formation was reported to be determined by the *HORT1* insertion in the promoter sequences
302 of *PeMYB11* genes[7], and that *P.* sa flowers harbored only small spots, the failure to identify
303 *PsMYB11*-related variants is potentially due to the lack of *HORT1* insertion in *PsMYB11*
304 promoters.

305 **Functions of *PsMYBx1* in repressing anthocyanin accumulation**

306 *PsMYBx1* had only one HG (HG type 1, **Fig. 6A**). Its expression was associated with both
307 stripe and spot/patch formation in *P.* sa and H80 (**Fig. 5A,B**), indicating potentially dual
308 function of *PsMYBx1* in repressing anthocyanin accumulation in both inter-stripe and inter-
309 spot/patch regions. This dual function was supported by the disruption of stripes in regions
310 surrounding patches, such as in H80-057, H80-114, and H80-245 (**Fig. 4A**). Given this
311 complexity of color patterning imposed by *PsMYBx1* genes, no single color patterning trait
312 variation can be readily determined to be related to *PsMYBx1*. Thus, instead of performing
313 GWAS, we silenced *MYBx1* genes via virus-induced gene silencing (**Methods**) in another
314 *Phalaenopsis* cultivar, *P.* 'Jinbianlinglong' (*P.* ji) (**Fig. 6B,C**), for which substantial seedlings
315 are available and the flowering regulation is relatively readily. *MYBx1* had four HGs in *P.* ji
316 (**Fig. S21B**), compared with a single HG in *P.* sa. *PjPDS* (*Phytoene desaturase*) was co-
317 silenced to indicate regions with gene silencing phenotypes where the yellow-green color was
318 faded (**Fig. 6B**). In *PjMYBx1-PjPDS*-silenced plants, ectopic anthocyanin pigmentation was
319 observed on the inflorescence stalk and abaxial surface of perianth, and in the inter-stripe
320 regions (**Fig. 6C**). However, the inter-spot regions remained unchanged (blue arrowheads in
321 **Fig. 6C**), contrasting with the proposed role of *MYBx1* in repressing anthocyanin
322 accumulation in inter-spot regions[8]. Although expression of color patterning-related genes
323 examined were generally decreased in *PjMYBxl-PjPDS*-silenced sepals (**Fig. 6D**) and petals
324 (**Fig. 6E**) compared with those in mock-treated plants, a slight increase in *PjMYB12*
325 expression was detected in petals. Taken together, these results demonstrate the multiple

326 functions of *MYBx1* in repressing anthocyanin accumulation in diverse regions, and suggest
327 variation in its functions across different *Phalaenopsis* orchids.

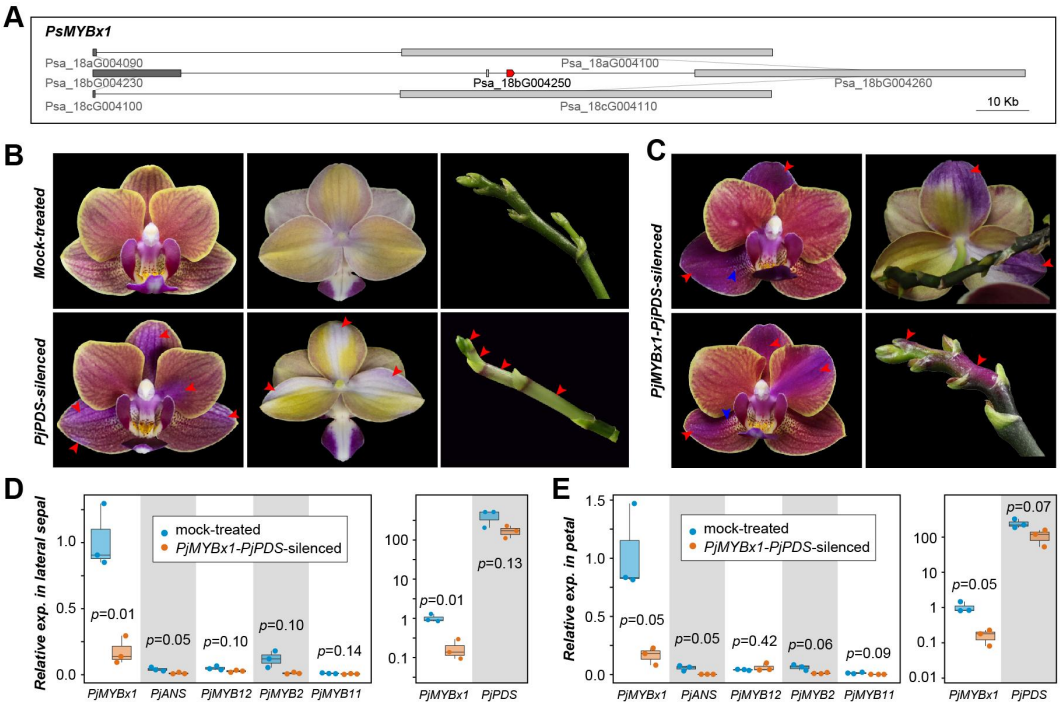

Fig. 6 Functions of *PjMYBx1* in repressing anthocyanin accumulation in diverse regions. (**A**) HG variation of *PsMYBx1*. (**B**) Flower and inflorescence images showing phenotypes of mock-treated (upper panel) and *PjPDS*-silenced (lower panel) plants. (**C**) Flower and inflorescence images of *PjMYBx1-PjPDS*-silenced plants. Red arrowheads indicate regions with phenotypic changes when compared with mock-treated plants, while the blue ones indicate the unchanged inter-spot regions. (**D,E**) Expression of color patterning-related genes in mock-treated and *PjMYBx1-PjPDS*-silenced sepals (**D**) and petals (**E**). *P*-values are from t-test.

328

329 # Discussion

330 The extraordinary phenotypic diversity and the high economic importance of *Phalaenopsis*
331 orchids necessitate a reference genome assembly with high quality to facilitate the
332 fundamental research and the breeding. The haplotype-resolved, chromosome-level genome
333 assembly of *P*. sa provided in this study served as a suitable reference in revealing the genetic
334 bases of trait diversity, particularly those of lip morphology, venation-associated stripes and
335 pink color on sepals/petals. This is exemplified by analysis in the H80 population. Given the
336 high heterogeneity of *P*. sa genome and the diverse genetic mechanisms underlying different
337 trait variation, we demonstrated that the choice of HChr sets as the reference genome, variant
338 types, and marker-effect models is critical for revealing the genetic bases of target trait
339 variation. The ratio of true to false positives of the identified significant variants can be the
340 primary metric for optimization. Specifically, for lip morphology variation, stripe P/A
341 variation, and background color variation, the optimal models were HChr set c + SNPs +

342 additive model, HChr set a + CNVs + additive model, HChr set b + SNPs/indels + 1-dom-ref
343 model, respectively. However, for spot/patch variation, genomes of patch-bearing
344 *Phalaenopsis* cultivars are supposed to serve as superior references compared with *P*. sa
345 genome.

346 *Phalaenopsis* flowers display extreme complexity in color patterning, which is much more
347 complicated than simply superposition of background color, spots and stripes. This
348 complexity is well demonstrated in the H80 population, such as that spots can be distributed
349 along the venation (H80-057 and H80-246), and spots/patches and stripes can be mutually
350 exclusive (H80-245) where the stripes are disrupted by spots/patches (**Fig. 4A**). This
351 complexity results from the substantial allelic polymorphisms of the regulators—particularly
352 *MYB* genes, including both R2R3-type (positive regulators) and R3-type (negative regulators).
353 Even in the single *P*. sa genome, these *MYB* genes exhibit HG variation in copy number
354 (*PsMYB11*), presence and absence (*PsMBY12* and *PsMYBx1*), gene structure, sequence and
355 expression (*PsMYB2*). Furthermore, *PsMYBx1* exhibits functional differentiation across
356 cultivars by repressing anthocyanin accumulation in distinct regions. Further studies are
357 needed to fully decipher the genetic bases of the extensive color patterning diversity and the
358 interplays among these regulators in *Phalaenopsis*.

359 Although genes regulating some of the key traits in *Phalaenopsis* have been well
360 established[1–3,5,9,22,23], what and how allelic polymorphisms of these genes have contributed to
361 the extraordinary trait diversity remain underexplored, restricting their utilization in breeding
362 practices. The genetic bases of three key trait variation revealed in this study—P/A of
363 *PsMYB12* genes, P/A of *b*-type HG of *PsMYB2*, and nucleotide polymorphisms nearby
364 *PsAGL6-2* genes—enable design of a multiplex polymerase chain reaction (MPCR) kit for
365 early-stage prediction of these three traits simultaneously. More traits can be predicted using
366 such a single MPCR experiment when the genetic bases of their diversity are deciphered, to
367 further improve the cost effectiveness. Given the unusually long juvenile phases of
368 *Phalaenopsis* orchids (three to five years), such a MPCR kit has great potential in
369 substantially shortening breeding cycles and reducing breeding costs.

370 Methods

371 Flow cytometry analysis

372 Young leaves of *P.* sa seedlings were used for the flow cytometry analysis, with young leaves
373 of cultivated tomato one month after seed germination being used as the internal reference.

374 Suspensions of *P.* sa and tomato leaves were strained using propidium iodide and mixed, and
375 a BD FACScalibur flow cytometer was used to detect the stained cell nuclei.

376 Karyotype analysis

377 Root tips of *P.* sa were harvested between 9:00 and 10:00 a.m., and approximately 0.5 cm
378 segments were excised using a clean razor blade. These segments were incubated in 0.1%
379 colchicine solution at room temperature in the dark for 5 h to induce chromosome
380 condensation, and were then fixed in Carnoy's solution (absolute ethanol:glacial acetic acid =
381 3:1, v/v) overnight. Fixed samples were rinsed three times with distilled water, dehydrated in
382 75%, 85%, 95%, and 100% ethanol for 10 minutes each, and then were stored in 75% ethanol
383 at 4 °C. To facilitate cell wall separation, pretreated samples were incubated in 75 mM KCl
384 solution at 37 °C for 30–60 minutes, and treated with 1 M HCl at 60 °C in a water bath for 10
385 minutes. Samples were stained with carbol fuchsin solution (Coolaber, Beijing, China) for 20
386 minutes, squashed on microscope slides, observed and photographed using a 60× objective
387 lens on an OLYMPUS BX53 optical microscope.

388 Genome sequencing and Hi-C sequencing

389 For HiFi sequences, the total DNA was extracted from *P.* sa young leaves, and used for the
390 library construction and genome sequencing with the PacBio Sequel II/Sequel IIe platform,
391 following the manufacturer's instructions at Haorui Genomics (Xi'an, China). Approximately
392 72.1 Gb of PacBio HiFi reads were generated using the CCS mode.

393 For Hi-C sequences, the frozen young leaves were ground with liquid nitrogen and then fixed
394 with 1% formaldehyde in vacuum at room temperature for 0.5-1h to preserve the three-
395 dimensional chromatin architecture, and 0.15 M glycine was added in the solution for 5 min
396 to quench the reaction. The mixed solution was then filtered with a 40 nm cell filter.
397 Chromatin was digested with 200 U DPNII restriction enzyme (Qiagen, 59971) at 37 ℃ for
398 16 h. DNA ends were labeled with biotin and incubated at 37 ℃ for 45 min. DNA ligation
399 was performed by the addition of T4 DNA ligase and incubation for 1-2 h. Then proteinase K
400 was added to reverse cross-linking during incubation at 65 ℃ for 1-2 h. DNA fragments were
401 purified and dissolved in 86 μL of water. Purified DNA was fragmented to a size of 350-500
402 bp, and DNA ends were then repaired. DNA fragments labeled with biotin were finally
403 separated on Dynabeads® M-280 Streptavidin (Life Technologies, 60210). The sequencing
404 was performed with the Illumina NovaSeq 6000 for paired-end sequences with 150 bp read
405 length. A total of ~130.6 Gb raw data was produced. After filtering adapter sequences and

read pairs with > 10% N or with low quality (Q ≤ 5) for >50% bases in at least one read, ~123.3 Gb clean data was retained.

Genome survey

The genome size was estimated via *k*-mer frequency analysis using the PacBio HiFi reads. The Jellyfish software (v2.2.10)[24] was used to generate the counts of 31-mers, and then the GenomeScope2 software (v2.0)[25] and Genomic Character Estimator (GCE) software (v1.0.2)[26] were applied to estimate the genome size and heterozygosity. SmudgePlot software (v0.2.2)[25] was used to estimate the ploidy level of *P.* sa genome.

De novo genome assembly

Since it lacks a chromosome-level reference assembly for *Phalaenopsis* species, at the first place, we attempted to assemble a set of draft representative pseudo-chromosomes to guide the subsequent refined assembling. Three strategies were conducted to assemble the draft genome: (**i**) all the PacBio HiFi reads were assembled using HiFiasm program (v0.19.6-r597)[27] with the parameter 'l=0' (to disable the purge of duplication); (**ii**) the PacBio HiFi reads were assembled using HiFiasm with the parameters 'l=3, hom-cov=57, hg-size=1.3g, n-hap=3, s=0.3, D=8, n-weight=6, n-perturb=80,000, f-perturb=0.3', with the assumption that the *P.* sa was a triploid, and the Hi-C reads were integrated to achieve haplotype-resolved assemblies; (**iii**) the PacBio HiFi reads were assembled with parameters 'l=3, hom-cov=39, hg-size=1.5g, n-hap=4, s=0.3, D=8, n-weight=6, n-perturb=80,000, f-perturb=0.3', with the assumption that the *P.* sa was a tetraploid, and the Hi-C reads were also integrated.

The Hi-C reads were mapped to the resulting primary unitigs from assembling **i**, and haplotigs of each haplotype from assembling **ii** and **iii**, separately, using bwa-mem2 (v2.2.1)[28]. The resulting bam files were further filtered using the 'filter_bam.py' utility in the HapHic (v1.0.2)[29] with the parameters 'mapq=1, NM=3'. The filtered bam files were then fed into the HapHic pipeline with parameters 'correct_nrounds=2, quick_view=T', and the output files .hic and .assembly were imported into the Juicebox Assembly Tools (JBAT, v1.08)[30]. Primary unitigs from strategy (i) and haplotigs from (ii) and (iii) were then ordered, joined, and manually adjusted based on Hi-C interaction patterns to obtain scaffolds, with the prior knowledge that nearby regions within the same chromosome having strong Hi-C signals whereas allelic regions having subdominant signals. The resulting scaffolds from all three strategies were compared, and 19 were selected as representative pseudo-chromosomes based on Hi-C signal continuity and Benchmarking Universal Single-Copy Orthologs (BUSCO)[31]

438 completeness. These 19 pseudo-chromosomes—seven from assembly **i**, ten from **ii**, and two
439 from **iii**—constitute the draft reference assembly (assembly v1.0).

440 Next, using the v1 draft assembly as a reference, we mapped all raw unitigs from assembly **ii**
441 and grouped them into 19 chromosomal groups. Unitigs within each group were ordered,
442 joined, and manually adjusted based on Hi-C interaction patterns, with the parameters
443 'mapq=0, NM=3' for filter_bam.py. This produced 70 preliminary pseudo-chromosomes.
444 Remaining unassigned unitigs were then incorporated and the entire set was reviewed and
445 corrected using Hi-C signals to yield the final assembly (assembly v2.0). SubPhase (v1.2.6)[32]
446 was used to phase subgenomes based on potential subgenome-specific *k*-mers, with
447 parameters "-k 15 -q 50 -f 12".

448 Genome assembly quality assessment

449 The *P.* sa genome assembly contiguity was accessed using Quality Assessment Tool for
450 Genome Assemblies (QUAST, v5.0.2)[33]. The BUSCO completeness was conducted using
451 BUSCO (v5.6.1)[31], against the eukaryota_odb12, viridiplantae_odb12, embrophyta_odb12,
452 and liliopsida_odb12 databases. The *K*-mer (21-mers) completeness and assembly consensus
453 quality value (QV) were estimated using Merqury (v1.4.1)[34]. The QV is a log-scaled
454 probability of errors for consensus base calls; the higher the QV values, the more accurate the
455 consensus base calls, with a QV value at 30 indicating 99.9% accuracy and 40 representing
456 99.99% accuracy[34]. The regional and structural assembly quality indicator (R-AQI and S-AQI)
457 were assessed using the Clipping information for Revealing Assembly Quality (CRAQ,
458 v1.0.9) algorithm[35]; the higher the AQI scores, the more accurate the genome assembly[35]. The
459 long terminal repeat retrotransposons (LTR-RTs) were first annotated using LTRharvest
460 (v1.6.1)[36] with the parameters '-maxlenltr 6000 -mindistltr 1500 -maxdistltr 25000 -mintsd 5
461 -maxtsd 5 -vic 10 -similar 85' and LTR_FINDER_parallel (v1.2)[37] with the parameters '-size
462 1000000 -time 300'. Then the LTR Assembly Index (LAI)[38] was calculated using
463 LTR_retriever (v2.9.5)[39] with the parameter '-u 0.5e-8'. The telomere and centromeric repeat
464 identification were conducted using quartet (v1.2.5) using default parameter settings[40].

465 Noncoding RNA and repetitive sequence identification

466 Noncoding RNAs were predicted using tRNAscan-SE (v2.0.9)[41] and Infernal software
467 (v1.1.4)[42] by searching the database Rfam (v14.10). The extensive *de-novo* TE annotator
468 (EDTA) (v2.1.0)[43]—that encompasses LTR_FINDER, LTRharvest, LTR_retriver, Generic
469 Repeat Finder, TIR-Learner, HelitronScanner, RepeatModeler and RepeatMasker— was used

470 to identify, annotate and classify repetitive elements of the genome, with parameters '--anno
471 1 --sensitive 1 --evaluate 1 --u 0.5e-8 --curatedlib ${SINE_LINE_lib}'. The library of short
472 interspersed elements (SINEs) and long interspersed nuclear elements (LINEs) was
473 downloaded from SINEBase (https://sines.eimb.ru/) as a compensation for the reference
474 sequences. The repetitive elements identified by EDTA included long terminal repeat (LTR)
475 retrotransposons (copia, gypsy and unknown ones) and non-LTR retrotransposons
476 (LINE_element and unknown ones), terminal inverted repeat (TIR) transposons (CACTA,
477 Mutator, PIF_Harbinger, Tc1_mariner and hAT), *Helitron* and other interspersed repeats.
478 Subsequently, the DeepTE (version updated on Oct 22th, 2022) script[44] was used to reclassify
479 the TEs that were initially annotated as 'LTR/unknown' by EDTA into copia, gypsy and
480 unknown classes, then the output file was fed back into EDTA to enhance the
481 comprehensiveness of TE annotation. To calculate LTR insertion time, we set the mutation
482 rate per generation (μ) to 0.5×10^{-8}, based on the assumption that the generation times of
483 orchids are four years[45]. The resulting hard-masked genome was converted into a soft-masked
484 genome, which was used in the subsequent analysis. The tandem repeat sequences in the *P.* sa
485 genome were identified using Tandem Repeats Finder (TRF, v4.09)[46] with the parameters '2
486 7 7 80 10 50 2000 -d -h'. Then the proportion of tandem repeats in the genome was
487 calculated using the script buildRepeatSummary.pl (v1.36) with default parameters[47].

488 Transcriptome sequencing

489 Total RNA was extracted for roots, leaves, stems and flower buds with mixed stages (one
490 duplicate) and other tissues at different development stages (three replicates, **Table S6**). The
491 library preparation and PCR amplification were performed according to the manufacturer's
492 instructions, and the RNA libraries were sequenced on the Illumina Novaseq 6000 platform
493 and 150 bp paired-end reads were generated. Clean data were further obtained by removing
494 reads containing adapters, reads containing poly-N and low quality reads from raw data,
495 using fastp (v0.24.1)[48] with parameters: '-g -q 5 -u 50 -n 15 -l 150 --overlap_diff_limit 1 --
496 overlap_diff_percent_limit 10'.

497 Protein-coding gene prediction

498 Protein-coding genes were predicted by integrating predictions based on sequence homology
499 and transcriptome, and the *ab initio* predictions using the soft-masked genome.

500 1. Homology-based prediction

501 Genome assembly and General Feature Format (GFF) files of six plant species were used for
502 homology-based prediction with Gene Model Mapper (GeMoMa, v1.9)[49]. The following
503 species were included: *Arabidopsis thaliana* (TAIR10.1, downloaded from NCBI), *Oryza*
504 *sativa* Japonica Group (IRGSP_1.0, from Ensembl Plant), *Asparagus officinalis* (Aspof.V1,
505 from Ensembl Plant), *Apostasia shenzhenica* (GCA_002786265.1, from NCBI)[50],
506 *Dendrobium catenatum* (GCF_001605985.2, from NCBI)[51], and *Phalaenopsis equestris*
507 (GCF_001263595.1, from NCBI)[10]. The GeMoMa parameters used were
508 'AnnotationFinalizer.r=NO p=false o=true tblastn=false'.

509 2. Transcriptome-based prediction

510 Clean RNA-seq reads from roots, leaves, stems, and mixed-stage flower buds were aligned to
511 *P. sa* genome using HISAT2 (hierarchical indexing for spliced alignment of transcripts,
512 v2.2.1)[52]. The alignments were processed by StringTie (v2.2.1)[53] to reconstruct genome-
513 guided transcripts. Potential open reading frames (ORFs, ≥100 amino acids) were identified
514 from these transcripts using TransDecoder (v5.7.1) (https://github.com/TransDecoder). A
515 genome-guided transcriptome was assembled using Trinity (v2.15.1)[54]. Both the StringTie
516 output and the Trinity assembly were then analyzed by the Program to Assemble Spliced
517 Alignments (PASA, v2.5.3)[55] with the parameters '--ALIGNERS blat,minimap2 --
518 TRANSDECODER' to generate transcriptome-supported gene models.

519 3. *Ab initio* prediction

520 The BRAKER3 pipeline (v3.0.3, incorporating Augustus and Genemark)[56], GeneID (2021
521 version)[57] and GlimmerHMM (2014 version)[58] were used for the *ab initio* gene prediction,
522 leveraging the transcriptome evidence as a guide.

523 4. Gene model integration and filtering

524 The EvidenceModeler (v2.1.0)[59,60] was used to integrate predictions from the above methods
525 into consensus gene structures. The following weights were assigned: GeneID (1),
526 GlimmerHMM (1), GeneMark (2), Augustus (4), GeMoMa (7), StringTie (5), and
527 PASA_transdecoder (10).

528 To refine the gene models, predicted protein sequences were searched against the Transposon
529 Protein Database (https://github.com/NBISweden/TransposonPSI/blob/master/

530 transposon_ORF_lib/transposon_db.pep) using TransposonPSI (v1.0.0), which employs PSI-
531 BLAST to identify TE-related sequences[61] with default parameters. Proteins with > 30%
532 similarity and > 60% coverage to known TE sequences were removed. In addition, gene
533 models with incomplete ORFs or ORFs ≤ 100 amino acids were excluded, if they met all the
534 three following conditions: lacking RNA-seq expression support, having < 50% coverage
535 overlap with GeMoMa models, and lacking functional annotation by InterProScan (v5.67-
536 99.0)[62].
537 5. Model optimization with extended RNA-seq data

538 The resulting gene models (v1.0) from above steps were further supplemented using
539 additional RNA-seq data from other 28 samples (**Table S6**). First, to avoid potential mapping
540 errors between tandem duplicates with high sequence similarities, RNA-seq reads were first
541 mapped to the v1.0 transcriptome, and unmapped reads were further mapped to the genome,
542 using TopHat2 (v2.1.1)[63] with the parameter '--max-multihits 1'. Next, individual bam files
543 were used to assemble transcripts (including alternative splicing isoforms) using StringTie.
544 The resulting GTF files from all samples were merged using StringTie to create a
545 comprehensive transcript annotation. TransDecoder was used to infer the ORFs, producing a
546 GFF file (v2.0) that included alternative splicing and five prime untranslated regions (5'UTR)
547 and 3'UTR information. Finally, the v2.0 annotations were compared with and used to
548 correct the v1.0 annotations, prioritizing v2.0 evidence in overlapping regions. This process
549 resulted in a final set of 93,708 protein-coding genes.

550 Protein-coding gene functional annotation

551 *P.* sa protein sequences were first searched against the eggNOG database (v5.0.2)[64] using
552 eggNOG-mapper (v2.1.12)[65], with parameters '-m diamond --seed_ortholog_evalue 1e-5 --
553 tax_scope Viridiplantae', for annotations including Gene Ontology (GO), Kyoto
554 Encyclopedia of Genes and Genomes (KEGG) Orthology (KO), and Clusters of Orthologous
555 Groups (COG). In order to obtain the functional description of related protein-coding genes
556 in model plant species, protein sequences of *P.* sa genome were searched against the UniProt
557 database (uniprot_sprot_plants.fasta that contains manually annotated protein sequences from
558 sequenced plant genomes, updated on March 27th, 2024, https://www.uniprot.org/) and
559 protein sequences from Arabidopsis (Araport11_pep_20240409, downloaded from The
560 Arabidopsis Information Resource) using BLASTP (v2.12.0+)[66] with parameters '-max_hsps
561 1 -max_target_seqs 1 -evalue 1e-5'. Furthermore, *P.* sa protein sequences were searched

562 against databases such as Pfam, SMART, CDD, Superfamily and others, using InterProScan
563 (v5.67-99.0) with parameters '-goterms -iprlookup'. Gene Functional Annotation for Plants
564 (GFAP, v3.1)[67] was used to annotate GO, KEGG and Pfam items by searching *P.* sa protein
565 sequences against a plant-specific database with the parameter '-awd psd -go -kegg -pfam',
566 and to obtain homology-based annotations by searching against the plant-specific NCBI non-
567 redundant (nr) database with the parameter '-awd nr'. Lastly, protein sequences were
568 searched against the Reference Protein Sequence Dataset (RPSD, release_2024_07_31) using
569 Ensemble Enzyme Prediction Pipeline (E2P2, v5)[68] with default parameters for the enzyme
570 function annotation. Transcription factors in *P.* sa genome were annotated in the Plant
571 Transcription Factor Database (PlantTFDB)[69].

572 **Homoeologous gene identification**

573 The Synteny and Rearrangement Identifier (SyRI, v1.7.0)[70] and MCScanX v1.0.0[71] were used
574 to conduct synteny analysis among different HChrs with default parameter settings, based on
575 genomic and protein sequences, respectively. Later on, the homoeologous genes were
576 identified with in-house scripts (see **Code availability**) using the outputs of SyRI
577 preferentially and MCScanX supplementarily, to avoid improper assignment of multiple
578 MCScanX short syntenic blocks between HChrs.

579 **Transcriptome analysis**

580 RNA-seq reads were mapped to the transcriptome and genome using Tophat2 (v2.1.1)[63], with
581 parameter '--max-multihits 1', using indexes from both the genome and the final
582 transcriptome. Read counts for protein-coding genes were obtained using htseq-count
583 (v2.0.5)[72] and TPM was calculated using the "calculateTPM" function in R package scater
584 (v1.0.4)[73]. The TPM values were normalized across samples using quantile normalization,
585 followed by averaging across the three biological replicates to obtain a mean TPM value for
586 each tissue.

587 To categorize genes according to their expression levels, we defined four thresholds based on
588 TPM values: 1) TPM = 0 indicating no expression; 2) 0 < TPM ≤ 2 indicating low expression;
589 3) 2 < TPM ≤ 10 indicating moderate expression; 4) TPM > 10 indicating high expression.
590 The threshold of TPM = 10 distinguishing high expression was determined by modeling the
591 observed gene expression data as a mixed distribution. This model comprises a normal
592 distribution, representing true biological expression, and an exponentially decaying
593 distribution, representing technical noise or background. These two components were fitted

594 using the nls function from the R "stats" package[74]. The value of TPM = 10 corresponds
595 approximately to the 50th percentile of the fitted normal distribution, suggesting that roughly
596 half of the genuine gene expression occurs below this level.

597 Weighted gene co-expression network analysis (WGCNA)

598 Co-expression analysis for HG type 1 genes that were highly expressed (TPM > 10) in ≥ 1
599 tissue and had no non-self blast hist in *P.* sa genome was performed using WGCNA
600 (v1.72.1)[75], under a signed network model, with a minimum module size of 10.

601 Gene tree building

602 To build a robust phylogeny tree for *MYB2*, *MYB11* and *MYB12* genes, coding sequences
603 (CDS) were obtained from multiple orchid species—*P. aphrodite* (*P. ap*), *P. equestris* (*P. eq*),
604 *Apostasia shenzhenica* (*A. sh*), *Cremastra appendiculata* (*C. ap*), *Cymbidium sinense* (*C. si*),
605 *Dendrobium chrysotoxum* (*D. ch*), *Gastrodia elata* (*G. el*), *Platanthera guangdongensis* (*P.*
606 *gu*), *Vanilla planifolia* (*V. pl*)—via the Orchid Multiomics Database
607 (http://orchidcomics.com/, access on June 11th, 2025)[76]. For each query gene, the top 20
608 BLAST hits per species (top 80 hits for *P.* sa) were extracted, translated into amino acid
609 sequences, and aligned using MEGA v7.0[77]. The resulting amino acid alignment was
610 manually curated to retain only reliably alignable regions, which were then converted back to
611 a corresponding CDS alignment. The best-fit nucleotide substitution model (GTR+I+G4) was
612 selected using ModelTest-NG (v0.1.7)[78]. Finally, a maximum likelihood phylogenetic tree
613 was constructed with RAxML-NG (v1.2.2) [79], employing 1000 bootstrap replicates.

614 The subclass of floral organ identity genes was determined by phylogenetic analysis. CDS of
615 MIKC-type MADS-box genes from *P.* sa and other orchids (**Fig. S10**) were translated into
616 amino acid sequences. These sequences were aligned using MEGA with manual adjustment.
617 A phylogeny was inferred using the 'JTT+I+G4' model with 1000 bootstrap iterations.

618 Genetic variant calling

619 Total DNA was extracted for flower stalks of 46 H80 individuals. Sequencing services of
620 partial H80 individuals were provided by BGI Genomics (Shenzhen, China), and the others
621 were provided by Novogene Co., Ltd (Beijing, China). All whole-genome resequencing in
622 this study was performed on the DNBSEQ-T7 platform using a 150-bp paired-end strategy,
623 generating ~60 Gb of data per sample. Raw reads were processed using fastp (v0.24.1)[48] with
624 parameters '-g -q 5 -u 50 -n 15 -l 150 --overlap_diff_limit 1 --overlap_diff_percent_limit 10'.

625 Clean reads were aligned to the *P*. sa genome using BWA-MEM algorithm in Burrow-
626 Wheeler Aligner (BWA, 0.7.18-r1243-dirty)[80] with default settings, followed by sorting with
627 SAMtools (v1.21)[81]. Duplicate reads were marked using the MarkDuplicates module in the
628 Genome Analysis Toolkit (GATK, v4.6.1.0)[82].

629 Population-level SNP and indel calling was conducted using GATK. First, variants were
630 called per sample with HaplotypeCaller using '-ERC GVCF -ploidy 4'. Individual gVCF files
631 were merged with CombineGVCFs and jointly genotyped with GenotypeGVCFs. SNPs and
632 indels were separated using SelectVariants and filtered with VariantFiltration using the
633 following criteria: for SNPs, 'QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum
634 < -12.5 || ReadPosRankSum < -8.0'; for indels, 'QD < 2.0 || FS > 200.0 || MQ < 40.0 || SOR >
635 10.0 || ReadPosRankSum < -20.0'. Multiallelic sites were split into biallelic records using
636 BCFtools norm (v1.21)[81] with '--multiallelics -both'.

637 SVs were called and genotyped using smoove (v0.2.8)[83] with default parameters. CNVs were
638 detected with CNVpytor (v1.3.2)[84] using a bin size of 100. Finally, all variants (SNPs, indels,
639 SVs and CNVs) were annotated using ANNOVAR (v2020-06-07)[85].

640 Genome-wide association study (GWAS) analysis

641 GWAS was performed using the R package GWASpoly (v2.13)[86] based on the Q + K mixed
642 model. First, the VCF files were transformed into dosage files using VCF2dosage function
643 with parameters '-ploidy 4 -min.DP 8 -max.missing 0.5 -min.minor 3'. To control for
644 population structure, the covariance matrix for polygenic effect was calculated using the
645 set.K function with the parameter 'LOCO=TRUE', where a different covariance matrix was
646 calculated for each chromosome according to markers from all other chromosomes[87]. During
647 this step, missing data were imputed using the k-Nearest Neighbors approach implemented
648 by the KNNImputer function from the sklearn.impute module[88]. The imputation was
649 performed five times with different k values (2, 5, 10, 15, and 20), and the average of the five
650 imputed values was used. The set.params function was used to set the threshold of minor
651 allele frequency with the parameter 'geno.freq = 1 - 5/N, MAF=0.05'. The GWASpoly
652 function was then used to assess the significance of each marker associated with target trait
653 variation under the additive model and two simplex dominance models (1-dom-ref and 1-
654 dom-alt) hypothesis. In the additive model, the marker effect scales linearly with the alternate
655 allele dosage. For the simplex dominance models, a single copy of the dominant allele
656 (reference or alternate allele) is sufficient for the complete dominance phenotype, thereby all

657 heterozygotes possessing this allele exhibit equivalent marker effects. Two significance
658 thresholds were adopted: p-value<0.05/variant counts and p-value<0.01/variant counts,
659 according to the Bonferroni adjustment approach. Lastly, the CMplot (v4.5.1)[89] package was
660 used to generate the Manhattan plots. For better visualization, $-\log_{10}(p)$ values greater than 30
661 were truncated to 30 in the Manhattan plots.

662 Virus-induced gene silencing

663 Seedlings of *P.* 'Jinbianlinglong' (*P.* ji) were cultivated in a growth room under a 14-h
664 light/10-h dark photoperiod, with a light intensity of 220 $\mu mol \cdot m^{-2} \cdot s^{-1}$ and a temperature of
665 25–27 °C. Partial sequences of *PjPDS* and *PjMYBx1* were introduced into the TRV2-based
666 pYL156 vector, as illustrated in **Fig. S21**. The resulting TRV2 plasmids were transformed
667 into *Agrobacterium tumefaciens* strain GV3101. Bacterial cultures containing either TRV2-
668 *PDS* or TRV2-*PDS-MYBx1*, and cultures containing the TRV1 plasmids were incubated
669 separately. When the OD_{600} reached approximately 2.0, the cells were harvested and
670 resuspended in an induction buffer (10 mM morpholineethanesulfonic acid, 10 mM $MgCl_2$,
671 and 200 uM acetosyringone) to a final OD_{600} of 2.0. Each TRV2 suspension was mixed with
672 an equal volume of the TRV1 suspension, and Silwet L-77 was added to a final concentration
673 of 0.005%. The mixture was incubated statically at room temperature for 3 h.

674 The resulting *Agrobacterium* mixture was infiltrated into the newly developed inflorescence
675 stalk and its nearest leaf. After infiltration, plants were kept in darkness for 48 h before being
676 returned to the growth room. Plants were monitored until phenotypic changes in the flowers
677 became evident.

678 To confirm the silencing of target genes in plants showing phenotypic changes, total RNA
679 was extracted from the corresponding floral tissues using the FastPure Universal Plant Total
680 RNA Isolation Kit (Vazyme, Nanjing, China). First-strand cDNA was synthesized using
681 MightyScript Plus First Strand cDNA Synthesis Master Mix (gDNA digester) (Sangon
682 Biotech, Shanghai, China). Quantitative reverse-transcription PCR (qRT-PCR) was
683 performed using SupRealQ Ultra Hunter SYBR qPCR Master Mix (U+) (Vazyme). The
684 $2^{-\Delta\Delta Ct}$ method was employed to calculate the relative expression of target and related genes,
685 with *PeActin* serving as the internal control. Primer sequences for VIGS and qRT-PCR were
686 listed in **Table S11.**

Data availability

The raw sequencing data used in this study have been submitted to the NCBI database under BioProject PRJNA1143085 (will be released upon publication), including PacBio HiFi, Hi-C and transcriptome sequences for *P.* sa, and whole genome sequences for 46 H80 individuals. The *P.* sa genome assembly, annotation, homoeologous gene table, and expression matrix for all protein-coding genes will be released in our database (http://www.phaldb.cn/) upon publication.

Code availability

Custom scripts and codes used in this study are provided at GitHub (https://github.com/peipeiwang6/Manuscript/tree/main/2025_Psa_genome).

Acknowledgement

We thank Dr. Huiqi Zhao for sharing the protocol for karyotype analysis, Prof. Zhiqiang Wu, Prof. Xingtan Zhang and Dr. Caiyao Zhao for their helpful suggestions on genome sequencing and assembling, Dr. Xukun Li for advice on anatomical photography and other experiments, Fu Chen for providing seedlings of *P.* 'Jinbianlinglong'. The English language of this manuscript was partially polished using DeepSeek, while all original content was written by the authors and the revised text was adopted following author review.

This work was supported by the National Natural Science Foundation of China (32370241), the Scientific Research Foundation of Kunpeng Institute of Modern Agriculture at Foshan (KIMAQD2022003), and the collaborative project (HXQS2024009 from Prof. Jue Ruan) to P.W.; National Key R&D Program of China (2023YFD1600505) to R.J.; Project supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (32221001) to H.K.; Chinese Universities Scientific Fund (2452024391) to R.Z.

Author contributions

PW, RJ, HK and RZ conceived and designed the study. PW, XZ, LZ, ZL, ZW, FW, YJ, YZ, and LW collected the plant materials of *P.* sa for genome and transcriptome sequencing; RJ conducted the hybridization and provided the plant materials; RJ, LW, YZ and XC conducted the karyotype analysis; PW, LZ, HH and LLan (Lan Lan) assembled the genome; LZ and PW conducted the genome annotation; LZ conducted the functional annotation for protein-coding

717 genes; LZ, FM and PW conducted the genome assembly quality assessment and syntenic
718 analysis; PW and FW conducted the transcriptome analysis; PW, XZ and FW conducted the
719 gene tree building; XZ conducted the genetic variant calling and GWAS analyses; ZW, LLin
720 (Letian Lin) and ZL conducted the VIGS experiments; LLin performed the qRT-PCR; PW
721 conducted all the other analyses; PW, RJ, XZ, LZ, LW, ZW and LLin wrote the draft with
722 contributions from all the other authors. All authors read and approved the final manuscript.

723 Supplementary figure legends

724 **Fig. S1 Phenotypes of H80 individuals.** Each panel shows an individual plant, with its
725 classifications for the P/A of stripe, background pink color and spot/patch, and lip
726 morphology provided on the top-left, top-right, bottom-left and bottom-right corners per
727 panel, respectively. Red arrowheads indicate stripes, while blue ones indicate venation-
728 associated shadow.

729 **Fig. S2 Genome survey of *P.* sa and karyotype analysis for H80.** (**A,B**) *k*-mer coverage
730 distribution of *P.* sa sequencing reads, under the assumptions that *P.* sa is a triploid (**A**) or
731 tetraploid (**B**). (**C**) *k*-mer coverage analysis indicates *P.* sa is an allo-triploid. (**D**) Karyotype
732 analysis reveals 70 chromosomes per cell in *P.* sa. (**E**) Inferred long terminal repeats (LTRs)
733 insertion time. (**F**) Karyotype analysis reveals 64 chromosomes per cell in H80 individuals.

734 **Fig. S3 Subgenome assignments of 70 *P.* sa chromosomes according to subgenome-**
735 **specific *k*-mers.** From outer to inner circles (1-9): (1) subgenome assignments; (2) significant
736 enrichment of subgenome-specific *k*-mers; (3) normalized proportion of subgenome-specific
737 *k*-mers; (4-7) count of subgenome-specific *k*-mers for four sets, respectively; (8) density of
738 long terminal repeat retrotransposons (LTR-RTs), with gray indicating nonspecific LTR-RTs;
739 (9) homoeologous blocks.

740 **Fig. S4 Expression profiles of *FAR-RED IMPAIRED RESPONSE 1 (FAR1*) genes in *P.***
741 **sa.** *FAR1* genes are grouped into different HG types, as illustrated in **Fig. 2A**. Color scale in
742 the heatmap indicates log-transformed Transcripts Per Kilobase (TPM) values, where blue
743 and light blue indicate original TPM=0 and TPM<2, respectively.

744 **Fig. S5 Expression profiles of potential *de novo* new genes.** (**A-G**) Expression profiles of
745 seven co-expression modules inferred from WGCNA analysis (**Methods**) for type 1 genes
746 that were highly expressed (TPM>10) in ≥ 1 tissue and had no non-self blastp hists in *P.* sa
747 genome. *Z*-scores of TPM values among all the tissues are shown. Example genes that were

748 preferentially expressed in different tissues are indicated.

749 **Fig. S6 Protein sequence alignment (A) and expression (B) of *PsAGL6-2* HGs.** HGs with
750 truncated sequences were indicated with green shadow.

751 **Fig. S7 Manhattan plot for GWAS of lip morphology in the H80 population using HChr**
752 **set a as the reference.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used,
753 respectively, with the additive marker-effect model. Horizontal solid line and dashed line
754 indicate -$\log_{10}$ (0.01/variant counts) and -$\log_{10}$ (0.05/variant counts) (**Methods**), respectively;
755 red and orange dots denote variants significantly associated with the lip morphology variation;
756 blue dots represent variants located within or adjacent (±2 kb) to the tandem duplicate cluster
757 of *PsAGL6-2a* genes.

758 **Fig. S8 Manhattan plot for GWAS of lip morphology in H80 using HChr set b as the**
759 **reference.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively,
760 with the additive marker-effect model. Red and orange dots denote variants significantly
761 associated with the lip morphology variation; blue dots represent variants located within or
762 adjacent (±2 kb) to the tandem duplicate cluster of *PsAGL6-2b* genes.

763 **Fig. S9 Manhattan plot for GWAS of lip morphology in H80 using HChr set c as the**
764 **reference.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively,
765 with the additive marker-effect model. Red and orange dots denote variants significantly
766 associated with the lip morphology variation; blue dots represent variants located within or
767 adjacent (±2 kb) to the tandem duplicate cluster of *PsAGL6-2c* genes.

768 **Fig. S10 Gene tree for MIKC-type MADS-box genes in orchids.** NCBI accession ID (such
769 as JL017169) was listed between gene name (such as *CeMADS32*) and the reference (such as
770 2021 Ai).

771 **Fig. S11 Manhattan plot for GWAS of stripe P/A in H80 using the additive marker-**
772 **effect model.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs that were called by
773 referring to HChr set a were used, respectively. (**E-H**) GWAS results when SNPs, indels, SVs
774 and CNVs that were called by referring to HChr set b were used. Red and orange dots denote
775 variants significantly associated with the P/A variation of stripe; blue dots represent variants
776 located within or adjacent (±2 kb) to the tandem duplicate cluster of *PsMYB12a* genes.

777 **Fig. S12 IGV images showing reads of seven H80 individuals with or without stripes**

that were mapped onto the *PsMYB12* gene-containing regions. Red arrowheads indicate stripes, while blue ones indicate venation-associated shadow in individuals which had no intact reads mapped onto *PsMYB12* genes.

Fig. S13 Evolution of color patterning-related *MYB* genes. (**A**) Gene tree of *MYB* genes from orchids with genome assemblies. (**B**) Evolutionary history of stripes on perianths in the genus *Phalaenopsis*. The phylogeny of *Phalaenopsis* species were modified from our previous review paper (Wang et al., 2025). The red star indicates the origination of stripes in the most recent common ancestor of *P. equestris* and *P. lindenii*. Three flower images show the venation-associated distribution of spots in other *Phalaenopsis* species. (**C**) Integrative Genomics Viewer images showing reads of *Vanda* 'Pakchong Blue' that were mapped onto genic regions of two *PsMYB12* genes as examples.

Fig. S14 Homoeologous gene variation in sequence and expression of *PsMYB2* and *PsMYB11*. (**A**) Protein sequence alignment of *PsMYB2* HGs. (**B**) HG-specific expression of *PsMYB2* in *P*. sa floral organs (upper) and the lips of selected H80 individuals (lower), with values in the table indicating TPM values. (**C,D**) Protein sequence and promoter DNA sequence (1 Kb upstream) alignments of *PsMYB11* HGs, respectively. (**E**) HG-specific expression of *PsMYB11* in *P*. sa floral organs (upper) and the lips of selected H80 individuals (lower). The HG showing high sequence divergence with others was indicated using green shadow.

Fig. S15 Manhattan plot for GWAS of background color in H80 using the additive marker-effect model. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs that were called by referring to HChr set a were used, respectively. Red and orange dots denote variants significantly associated with the P/A variation of pink color; blue dots represent variants located within or adjacent (±2 kb) to *PsMYB2a*.

Fig. S16 Manhattan plot for GWAS of background color in H80 using the 1-dom-ref marker-effect model. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs that were called by referring to HChr set b were used, respectively. Red dots denote variants significantly associated with the P/A variation of pink color; blue dots represent variants located within or adjacent (±2 kb) to *PsMYB2b*.

Fig. S17 Manhattan plot for GWAS of background color in H80 using the 1-dom-alt marker-effect model. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs that were

called by referring to HChr set b were used, respectively. Red and orange dots denote variants significantly associated with the P/A variation of pink color; blue dots represent variants located within or adjacent (±2 kb) to *PsMYB2b*.

Fig. S18 Manhattan plot for GWAS of patch P/A in H80 using HChr set a as the reference. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively, with the additive marker-effect model. Red and orange dots denote variants significantly associated with the P/A variation of patch; blue dots represent variants located within or adjacent (±2 kb) to *PsMYB11a*.

Fig. S19 Manhattan plot for GWAS of patch P/A in H80 using HChr set b as the reference. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively, with the additive marker-effect model. Red and orange dots denote variants significantly associated with the P/A variation of patch; blue dots represent variants located within or adjacent (±2 kb) to *PsMYB11b*.

Fig. S20 Manhattan plot for GWAS of patch P/A in H80 using HChr set c as the reference. (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively, with the additive marker-effect model. Red and orange dots denote variants significantly associated with the P/A variation of patch; blue dots represent variants located within or adjacent (±2 kb) to the tandem duplicate cluster of *PsMYB11c* genes.

Fig. S21 *PjPDS* and *PjMYBx1* sequences and the VIGS constructions. (**A**) CDS and 3'UTR sequence alignment for four *PjPDS* HGs. (**B**) CDS sequence alignment for four *PjMYBx1* HGs. (**C**) Diagrams showing the structure of TRV2 plasmid PYL156, and the insertion locations of *PjPDS* and *PjMYBx1* sequences. *PjPDS_1* and *PjPDS_2* are sequences from two *PjPDS* HGs marked with green shadow.

Supplementary tables

Table S1 Comparison of basic statistics for genome assembly and annotation of three *Phalaenopsis* species

Table S2 Structural variation summary

Table S3 Statistics of repetitive sequences in *P.* sa genome

Table S4 Statistics of *P.* sa genes and the functional annotations

838 **Table S5** Statistics of noncoding RNAs in *P*. sa genome

839 **Table S6** Information for RNA-seq samples

840 **Table S7** IDs for genes mentioned in the manuscript

841 **Table S8** Allele read ratios of two SNPs in H80 individuals with different lip morphology

842 **Table S9** Association between the presence/absence of stripes and *PsMYB12* genes in 46
843 H80 individuals

844 **Table S10** Association between *PsMYB2b*-type sequences with background colors

845 **Table S11** Primer sequences used for VIGS and qRT-PCR

846 Reference

847 1. Hsiao, Y. *et al.* A novel homodimeric geranyl diphosphate synthase from the orchid
848 *Phalaenopsis bellina* lacking a DD(X)$_{2–4}$D motif. *Plant J* **55**, 719–733 (2008).
849 2. Huang, H. *et al. Terpene synthase-b* and *terpene synthase-e/f* genes produce
850 monoterpenes for *Phalaenopsis bellina* floral scent. *Front Plant Sci* **12**, 700958 (2021).
851 3. Hsu, C.-C., Chen, Y.-Y., Tsai, W.-C., Chen, W.-H. & Chen, H.-H. Three R2R3-MYB
852 transcription factors regulate distinct floral pigmentation patterning in *Phalaenopsis* spp.
853 *Plant Physiol* **168**, 175–191 (2015).
854 4. Chen, W., Hsu, W., Hsu, H. & Yang, C. A tetraspanin gene regulating auxin response
855 and affecting orchid perianth size and various plant developmental processes. *Plant*
856 *Direct* **3**, e00157 (2019).
857 5. Lu, H.-C. *et al. R2R3-MYB* genes coordinate conical cell development and cuticular wax
858 biosynthesis in *Phalaenopsis aphrodite*. *Plant Physiol* **188**, 318–331 (2022).
859 6. Chuang, Y.-C. *et al.* A dual repeat cis-element determines expression of *GERANYL*
860 *DIPHOSPHATE SYNTHASE* for monoterpene production in *Phalaenopsis* orchids. *Front*
861 *Plant Sci* **9**, 765 (2018).
862 7. Hsu, C.-C., Su, C.-J., Jeng, M.-F., Chen, W.-H. & Chen, H.-H. A *HORT1* retrotransposon

863 insertion in the *PeMYB11* promoter causes harlequin/black flowers in *Phalaenopsis*
864 orchids. *Plant Physiol* **180**, 1535–1548 (2019).
865 8. Lu, T.-W., Chen, W.-H., Chen, P.-Y., Shu, Y.-C. & Chen, H.-H. Correction: Perturbation
866 of periodic spot-generation balance leads to diversified pigmentation patterning of
867 harlequin *Phalaenopsis* orchids: in silico prediction. *BMC Plant Biol* **24**, 782 (2024).
868 9. Hsu, H.-F. *et al.* Multifunctional evolution of B and *AGL6* MADS box genes in orchids.
869 *Nat Commun* **12**, 902 (2021).
870 10. Cai, J. *et al.* The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* **47**,
871 65–72 (2015).
872 11. Chao, Y. *et al.* Chromosome-level assembly, genetic and physical mapping of
873 *Phalaenopsis aphrodite* genome provides new insights into species adaptation and
874 resources for orchid breeding. *Plant Biotechnol J* **16**, 2027–2041 (2018).
875 12. Bashton, M. & Chothia, C. The generation of new protein functions by the combination
876 of domains. *Structure* **15**, 85–99 (2007).
877 13. Vishwanath, S., de Brevern, A. G. & Srinivasan, N. Same but not alike: Structure,
878 flexibility and energetics of domains in multi-domain proteins are influenced by the
879 presence of other domains. *PLoS Comput Biol* **14**, e1006008 (2018).
880 14. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and
881 powerful approach to multiple testing. *J R Stat Soc Series B* **57**, 289–300 (1995).
882 15. Liu, Y. *et al. Arabidopsis FHY3* and *FAR1* regulate the balance between growth and
883 defense responses under shade conditions. *Plant Cell* **31**, 2089–2106 (2019).
884 16. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of *de novo* genes in
885 *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
886 17. Dion-Côté, A.-M. Spermatogenesis: A hotspot for new genes. *eLife* **8**, e50136 (2019).
887 18. Yang, H. *et al.* Global modulation of gene expression and transcriptome size in aneuploid

888 combinations of maize. *Proc Natl Acad Sci U S A* **122**, e2426749122 (2025).
889 19. Pavelka, N. *et al.* Aneuploidy confers quantitative proteome changes and phenotypic
890 variation in budding yeast. *Nature* **468**, 321–325 (2010).
891 20. Selmecki, A., Forche, A. & Berman, J. Aneuploidy and isochromosome formation in
892 drug-resistant *Candida albicans*. *Science* **313**, 367–370 (2006).
893 21. Wang, F. *et al.* Molecular mechanisms underlying floral trait formation in *Phalaenopsis*
894 orchids. *Hortic Res* uhaf340 (2025) doi:10.1093/hr/uhaf340.
895 22. Chen, W.-H., Jiang, Z.-Y., Hsu, H.-F. & Yang, C.-H. Silencing of *FOREVER YOUNG*
896 *FLOWER* -like genes from *Phalaenopsis* orchids promotes flower senescence and
897 abscission. *Plant Cell Physiol* **62**, 111–124 (2021).
898 23. Wang, R., Mao, C. & Ming, F. PeMYB4L interacts with PeMYC4 to regulate
899 anthocyanin biosynthesis in *Phalaenopsis* orchid. *Plant Sci* **324**, 111423 (2022).
900 24. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
901 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
902 25. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and
903 Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432
904 (2020).
905 26. Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de
906 novo genome projects. Preprint at https://doi.org/10.48550/ARXIV.1308.2012 (2013).
907 27. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
908 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
909 28. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration
910 of BWA-MEM for multicore systems. in *2019 IEEE International Parallel and*
911 *Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil,
912 2019). doi:10.1109/IPDPS.2019.00041.

913 29. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using
914 Hi-C data without reference genomes. *Nat Plants* **10**, 1184–1200 (2024).
915 30. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with
916 unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
917 31. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and
918 annotation completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
919 32. Jia, K. *et al.* SubPhase: a robust allopolyploid subgenome phasing method based on
920 subgenome-specific *k*-mers. *New Phytol* **235**, 801–809 (2022).
921 33. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
922 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
923 34. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality,
924 completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245
925 (2020).
926 35. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome
927 assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat*
928 *Commun* **14**, 1–12 (2023).
929 36. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software
930 for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
931 37. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling
932 rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48 (2019).
933 38. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR
934 Assembly Index (LAI). *Nucleic Acids Res* **46**, e126 (2018).
935 39. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for
936 identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422
937 (2018).

938 40. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and
939 centromeric repeat identification. *Hortic Res* **10**, uhad127 (2023).
940 41. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection
941 and functional classification of transfer RNA genes. *Mol Biol Evol* **49**, 9077–9096 (2021).
942 42. Nawrocki, E. P. Annotating functional RNAs in genomes using Infernal. *Methods Mol*
943 *Biol* **1097**, 163–197 (2014).
944 43. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a
945 streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
946 44. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo
947 classification of transposons with convolutional neural network. *Bioinformatics* **36**,
948 4269–4275 (2020).
949 45. Zhang, W. *et al.* Genome sequence of *Apostasia ramifera* provides insights into the
950 adaptive evolution in orchids. *BMC Genomics* **22**, 536 (2021).
951 46. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
952 *Res* **27**, 573–580 (1999).
953 47. Smit, A. & Hubley, R. RepeatModeler Open-1.0. 2008-2015
954 <http://www.repeatmasker.org>.
955 48. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and
956 deduplication using fastp. *iMeta* **2**, e107 (2023).
957 49. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction
958 utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* **1962**, 161–
959 177 (2019).
960 50. Zhang, G.-Q. *et al.* The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–
961 383 (2017).
962 51. Zhang, G.-Q. *et al.* The *Dendrobium catenatum* Lindl. genome sequence provides

963 insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep*
964 **6**, 19029 (2016).
965 52. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
966 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–
967 915 (2019).
968 53. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
969 RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
970 54. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a
971 genome from RNA-Seq data. *Nat Biotechnol* **29**, 644–652 (2011).
972 55. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript
973 alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
974 56. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and
975 protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* **34**,
976 769–777 (2024).
977 57. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr Protoc*
978 *Bioinformatics* **18**, (2007).
979 58. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open
980 source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
981 59. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
982 EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7
983 (2008).
984 60. Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A. & Wortman, J. R. Approaches to
985 fungal genome annotation. *Mycology* **2**, 118–141 (2011).
986 61. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
987 database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

988 62. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
989 *Bioinformatics* **30**, 1236–1240 (2014).
990 63. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with
991 RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
992 64. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
993 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
994 *Res* **47**, D309–D314 (2019).
995 65. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.
996 eggNOG-mapper v2: functional annotation, orthology assignments, and domain
997 prediction at the metagenomic scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
998 66. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
999 (2009).
1000 67. Xu, D. *et al.* GFAP: ultrafast and accurate gene functional annotation software for plants.
1001 *Plant Physiol* **193**, 1745–1748 (2023).
1002 68. Schläpfer, P. *et al.* Genome-wide prediction of metabolic enzymes, pathways, and gene
1003 clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
1004 69. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory
1005 interactions in plants. *Nucleic Acids Res* **45**, D1040–D1045 (2017).
1006 70. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic
1007 rearrangements and local sequence differences from whole-genome assemblies. *Genome*
1008 *Biol* **20**, 277 (2019).
1009 71. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene
1010 synteny and collinearity. *Nucleic Acids Res* **40**, e49–e49 (2012).
1011 72. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-
1012 throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945

1013 (2022).
1014 73. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing,
1015 quality control, normalization and visualization of single-cell RNA-seq data in R.
1016 *Bioinformatics* **33**, 1179–1186 (2017).
1017 74. R Core Team. R: a language and environment for statistical computing. R foundation for
1018 statistical computing, Vienna. (2023).
1019 75. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
1020 analysis. *BMC Bioinformatics* **9**, 559 (2008).
1021 76. Wei, Y. *et al.* OrchidMD: an integrated and user-interactive orchid multi-omics database
1022 for mining genes and biological research. *Plant Biotechnol J* pbi.70445 (2025)
1023 doi:10.1111/pbi.70445.
1024 77. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis
1025 version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870–1874 (2016).
1026 78. Darriba, D. *et al.* ModelTest-NG: a new and scalable tool for the selection of DNA and
1027 protein evolutionary models. *Mol Biol Evol* **37**, 291–294 (2020).
1028 79. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast,
1029 scalable and user-friendly tool for maximum likelihood phylogenetic inference.
1030 *Bioinformatics* **35**, 4453–4455 (2019).
1031 80. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1032 Preprint at https://doi.org/10.48550/ARXIV.1303.3997 (2013).
1033 81. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008
1034 (2021).
1035 82. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of
1036 samples. Preprint at https://doi.org/10.1101/201178 (2017).
1037 83. Pedersen, B. S. smoove: structural-variant calling and genotyping with existing tools.

1038 (2020).
1039 84. Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: A tool for copy
1040 number variation detection and analysis from read depth and allele imbalance in whole-
1041 genome sequencing. *GigaScience* **10**, giab074 (2021).
1042 85. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
1043 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
1044 86. Rosyara, U. R., Jong, W. S. D., Douches, D. S. & Endelman, J. B. Software for genome-
1045 wide association studies in autopolyploids and its application to potato. *Plant Genome* **9**,
1046 plantgenome2015.8.73 (2016).
1047 87. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and
1048 pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100–106
1049 (2014).
1050 88. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**,
1051 2825–2830 (2011).
1052 89. Yin, L. CMplot: Circle Manhattan Plot. (2024).