Haplotype-resolved genome assembly advances genetic understanding of trait diversity 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 trait diversity with unclear genetic bases. Here, we present a haplotype-resolved, chromosome-level genome for an aneuploid cultivar 'Santiago'. We reveal pervasive variation in the number and sequence among homoeologous genes. When serving as a reference, this genome facilitates the genetic understanding of trait variation in a hybrid population that is highly representative in trait polymorphisms. Specifically, nucleotide polymorphisms in a potential long-distance enhancer and the consequential expression variation of *PsAGL6-2*, presence/absence of *PsMYB12* and *b*-type homoeologous gene of *PsMYB2* determine lip morphology variation, presence/absence of venation-associated stripes and background pink color, respectively. Diverse functions of *PsMYBx1* in repressing anthocyanin accumulation across different cultivars further enhances the color patterning diversity in *Phalaenopsis*. Our study provides a practical framework for using a highly heterogeneous, haplotype-resolved genome to decode phenotypic diversity and has the potential to promote marker-assisted breeding in *Phalaenopsis*.

Main

Phalaenopsis orchids are a major ornamental crop prized for their beauty and long flowering. Flowers in *Phalaenopsis* exhibit extraordinary diversity in organ organization, size, morphology, color patterning, scent, longevity, and so on, making *Phalaenopsis* a valuable system for horticultural research. Key genes regulating specific traits have been identified, such as floral scent[1,2], floral organ identity[3–6], color patterning[7–10], petal size[11], and cuticular wax[12]. However, the genetic bases of these diversity remain largely uncharacterized, except for the presence/absence (P/A) variation of floral scent[1,13] and the variation in spot/patch patterns[7,9,10]. For example, the specific genetic variation responsible for the P/A of venation-associated stripes, diverse perianth colors and lip morphology across different cultivars are unclear.

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[14,15]. Consequently, the full scope of these allelic polymorphisms and their contributions to trait diversity is still poorly understood.

In this study, we present a haplotype-resolved, chromosome-level genome for a highly heterozygous aneuploid cultivar, *P.* 'Santiago' (hereafter *P.* sa, **Fig. 1A**). In addition, we generate a hybrid population (H80) which exhibits extensive and representative polymorphisms in lip morphology and floral color patterning (**Fig. S1**). By analyzing the whole genome resequencing data of this population using *P*. sa genome as the reference, we decipher genetic variation that account for these trait polymorphisms.

Results

The genome assembly and annotation of *P.* sa

We generated the assembly of *P.* sa using a stepwise strategy (**Methods**). The whole genome size of *P.* sa was estimated to range from 3.3 Gb to 4.1 Gb by *k*-mer-based analyses and flow cytometry (**Methods**, **Fig. S2A,B**). *K*-mer frequency analysis (**Fig. S2C**) and karyotype analysis (70 chromosomes per cell, **Fig. S2D**) indicated *P*. sa as an aneuploid. Nineteen representative pseudo-chromosomes (a haplotype [1x] is supposed to contain 19

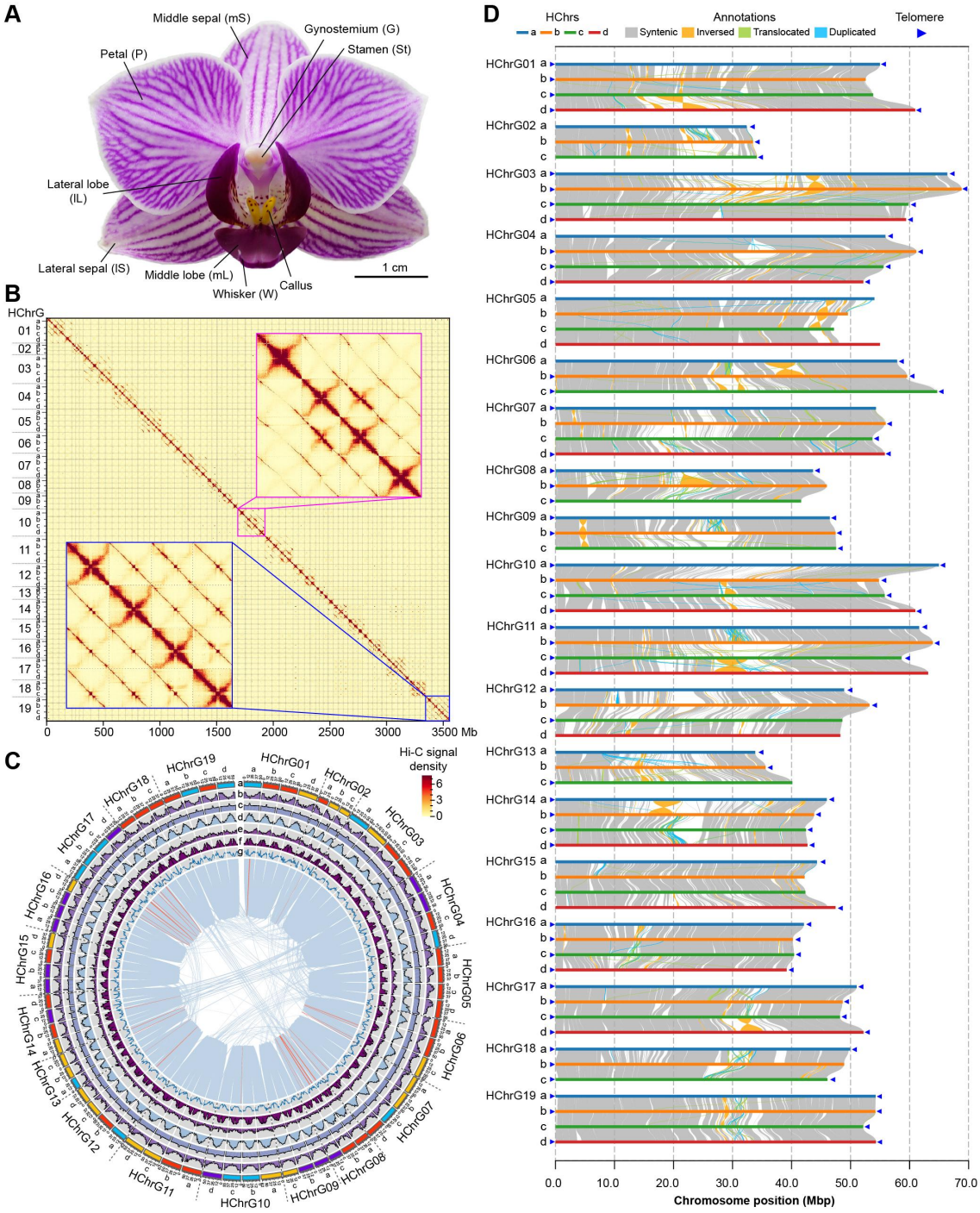

Fig. 1 Genome assembly of *Phalaenopsis* 'Santiago' (*P.* sa). (**A**) A *P.* sa flower has one middle sepal, two lateral sepals, two petals, one lip, and a gynostemium that connects the stamen with two pollinia and the ovary. (**B**) Hi-C heatmap showing chromosomal interactions, with bin size of 250 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.

92 chromosomes)[15] were first assembled using the PacBio HiFi reads and Hi-C short reads
93 (**Methods**), to guide the following haplotype-resolved assembly. The primary unitigs were
94 mapped to these 19 pseudo-chromosomes and grouped accordingly (**Methods**). Hi-C reads
95 were used to scaffold these grouped unitigs to a haplotype-resolved, chromosome-level
96 assembly per pseudo-chromosome. Finally, all these pseudo-chromosome-specific assemblies
97 and unassigned unitigs were jointly refined and corrected using Hi-C reads (**Methods**).

98 The final assembly comprised 70 pseudo-chromosomes, with 6 and 13 homoeologous
99 chromosome groups (HChrGs) having 3 and 4 haplotypes, respectively (**Fig. 1B,C**). The total
100 assembly size was approximately 3.55 Gb, with a structural assembly quality indicator
101 (AQI)[16] of 96.57% and an assembly consensus quality value (QV)[17] of 65.98. Benchmarking
102 Universal Single-Copy Orthologs (BUSCO) assessment[18] against the liliopsida database
103 indicated 98.8% completeness, exceeding those of two other available *Phalaenopsis* genome
104 assemblies (**Table S1**). Predicted telomere sequences were identified at most chromosome
105 ends (**Fig. 1D**), further suggesting the relatively high assembly quality. Due to the complex
106 hybrid history, the *P.* sa genome did not display obvious phases of subgenomes (**Fig. 1C, Fig.
107 S3**). We predicted 93,708 protein-coding genes using *de novo*, homology-based, and
108 transcriptome-supported approaches (**Methods**, **Table S2**), with a BUSCO (liliopsida)
109 completeness of 98.5%. Repetitive sequences constituted 67.97% (2.41 Gb) of *P. sa* genome
110 (**Table S3**). Additionally, the genome annotation included 34,366 non-coding RNAs (**Table
111 S4**).

112 Homoeologous gene variation in *P.* sa

113 Syntenic analysis (**Methods**) revealed extensive structural variation between HChrs (**Fig. 1D**,
114 **Table S5**). Accordingly, all the 93,708 protein-coding genes were classified into 33,697
115 homoeologous gene groups (HGGs) (**Methods**), which were further categorized into six
116 types according to variation in the number and sequence among HGs (**Fig. 2A**). Genes in
117 type 1 contained a single HG each; type 2 genes had ≥2 HGs while with HChr-specific HG
118 losses; type 3 genes had HChr-specific tandem duplicates; type 4 genes had the same number
119 of tandem duplicates per HChr; type 5 genes had single copy per HChr but with different
120 protein domain constitutions; type 6 genes had single copy per HChr with same domain
121 constitutions (**Fig. 2A**). Protein domain constitutions (**Methods**) were examined because
122 proteins with different domain constitutions are likely to perform divergent functions[19,20].

123

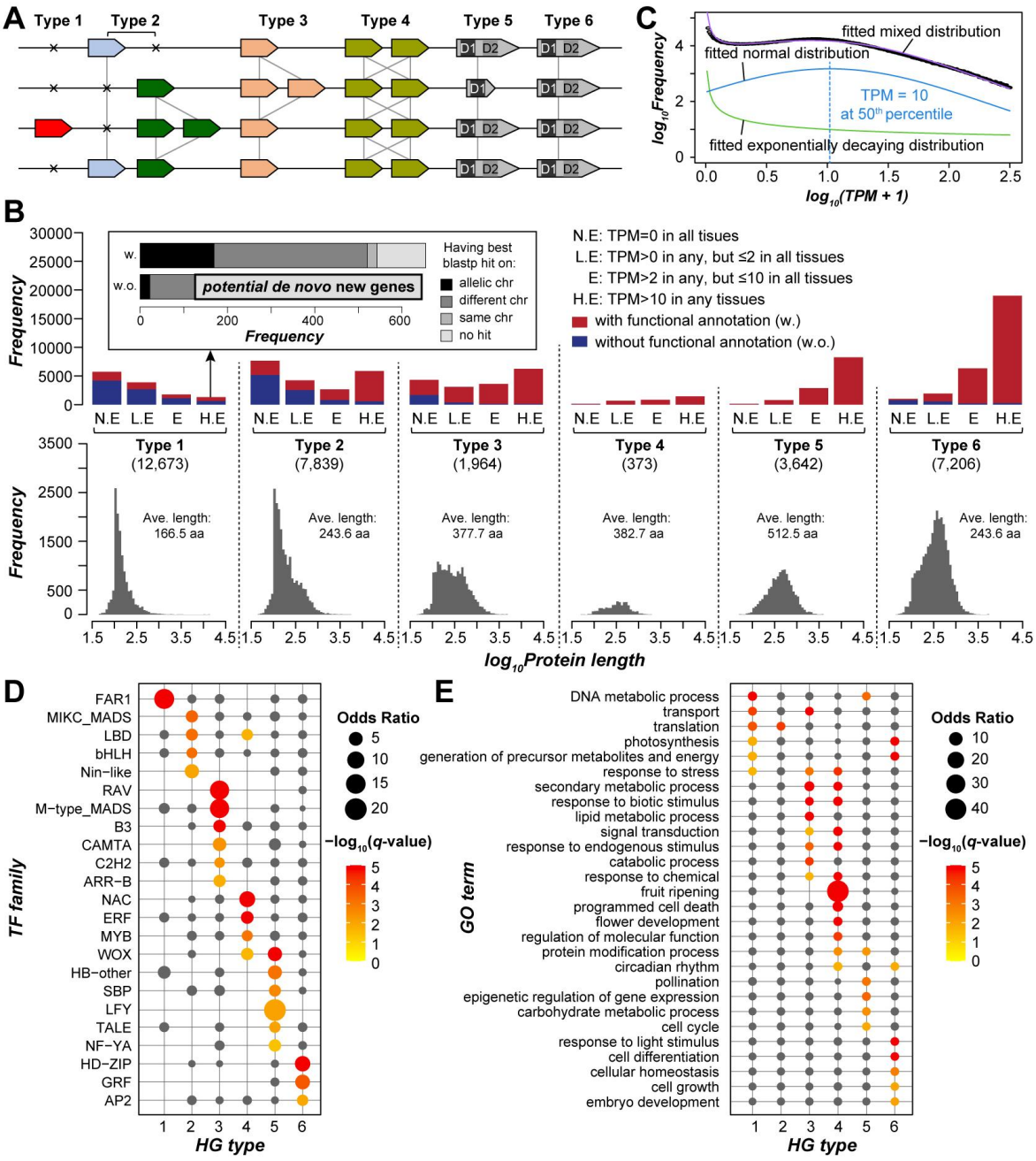

Fig. 2 Homoeologous gene variation in *P.* sa genome. (**A**) Diagram illustrating the six HG types. D1 and D2: protein domain 1 and domain 2. (**B**) Frequency of six HG types (upper panel) with the number of HGGs listed within the parentheses, and their protein length distribution (lower) with the 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; cases with *q*-values≥0.05 or odds ratio<1 were colored gray.

139 Genes in types 1-6 differed in terms of expression pattern, functional annotation (**Methods**)
140 and encoded protein length (**Fig. 2B,C**). They tended to belong to different transcription
141 factor (TF) families (**Fig. 2D**) and perform different biological functions (**Fig. 2E**).
142 Specifically, type 1 genes tended to be not or lowly expressed in all 28 examined samples
143 (**Table S6**), have no functional annotations and have relatively short proteins (**Fig. 2B**). All
144 the 13 *FAR-RED IMPAIRED RESPONSE 1 (FAR1*) genes in type 1 (**Fig. 2D**) were not or
145 lowly expressed (**Fig. S4**), indicating the ultimate destiny of being lost from the genome.
146 *FAR1* genes are essential for phytochrome signaling and balancing plant growth and defense
147 responses under shade conditions[21]. The deletion of certain *FAR1* genes from the
148 *Phalaenopsis* genome may be results of adaptation to shady, damp forest understory. In
149 contrast, some other type 1 genes might be *de novo* new genes endowing *Phalaenopsis* with
150 new phenotypic and adaptive characteristics: 519 genes were highly expressed (TPM>10) in
151 certain tissues while having no functional annotations and no non-self blastp hits within *P.* sa
152 genome (**Fig. 2B**). Among them, 243 (46.8%) were preferentially expressed in stamen (**Fig.
153 S5A-C**), consistent with previous findings that new genes tended to arise from male
154 reproductive organs[22,23]. The remaining 276 genes were preferentially expressed in other
155 tissues (**Fig. S5D-G**), suggestive of their potential functions in the corresponding tissues.

156 Type 2 genes were enriched for TFs that play important roles in floral development, cell
157 proliferation, determination and differentiation, lateral organ formation, responses to
158 environmental stimuli, and nitrate responses (**Fig. 2D**). Type 3 genes were enriched for TFs
159 and GO terms involved in flowering, embryogenesis, stress responses and signaling (**Fig.
160 2D,E**). Type 4 genes were enriched in TFs and functions related to plant growth and stress
161 responses (**Fig. 2D,E**). Genes in types 5 and 6 were enriched for TFs responsible for and GO
162 terms related to essential and conserved functions (**Fig. 2D,E**). These results indicate that the
163 HChr-specific gene duplicates/losses (types 2,3), together with those tandem duplicates
164 retained in all HChrs (type 4), may have driven the high diversity of floral traits and the
165 adaptation of *Phalaenopsis* to diverse environments.

166 Dosage effects of homoeologous gene count

167 Considering the aneuploidy nature of and the substantial HG count variation in *P*. sa genome,
168 we asked whether the chromosome ploidy (3x vs 4x) and HG count variation affect the total
169 expression amount per gene. To ensure a valid comparison, only reciprocal best-matched
170 (RBM) gene pairs with different HG counts—each gene had no HChr-specific tandem

duplicates and had the same protein domain constitutions among HGs—were examined (C1-C8, **Fig. 3A**). This strategy was adopted because the gene expression in the hybrid parents of *P.* sa were not available. Gene expression scaled with chromosome ploidy: summed expression levels of type 6 genes with 4 HGs (4/4) were generally higher than those of their 3/3 RBM counterparts (**Fig. 3B**), but became comparable when averaged per chromosome count (**Fig. 3C**). This dosage effect was similar to those found in other aneuploid organisms, where gene expression in aneuploids were compared with their corresponding diploids, such as maize[24], budding yeast[25] and *Candida albicans*[26].

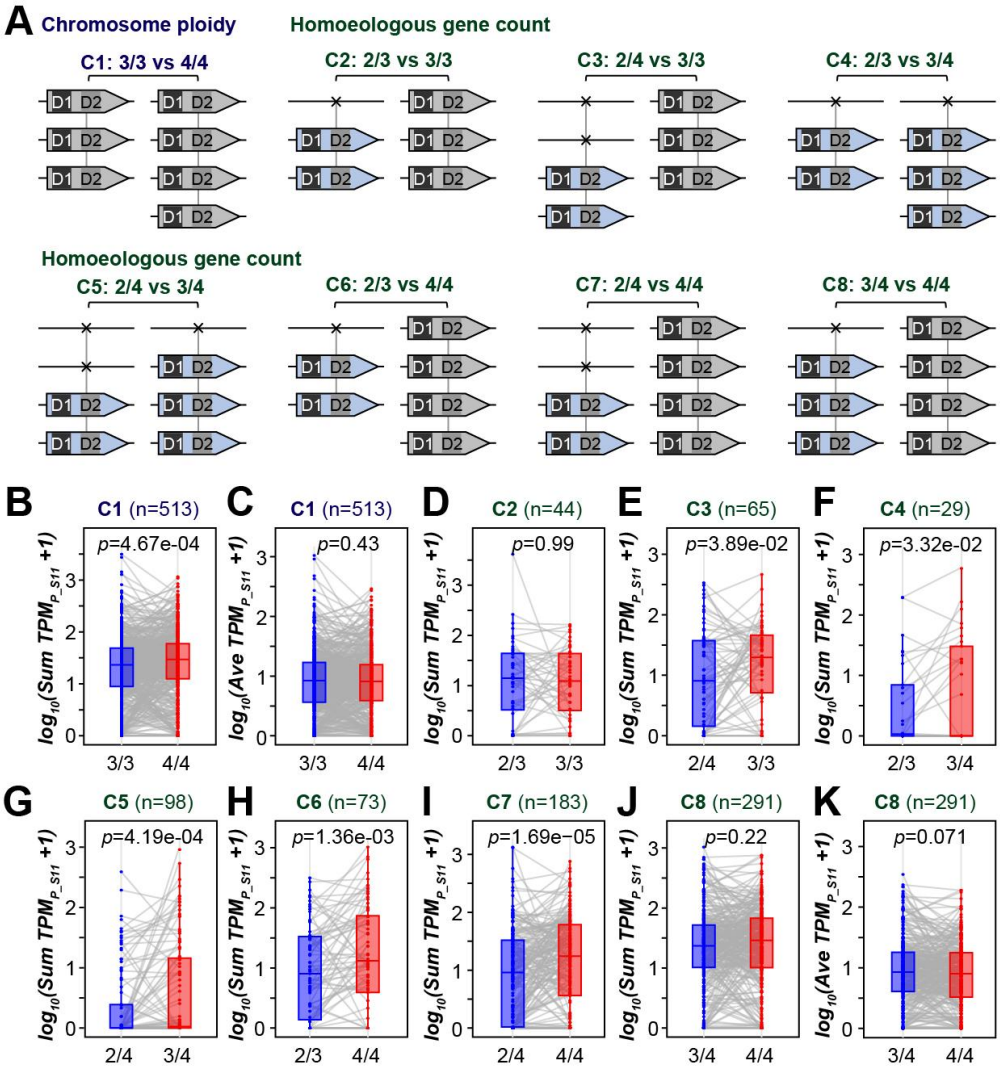

Fig. 3 Dosage effects of chromosome ploidy and homoeologous gene count on gene expression. (**A**) Diagrams showing eight comparisons, where C1 compares reciprocal best-matched (RBM) gene pairs located on HChrGs with different ploidies (3x vs 4x), while C2-C8 compare RBM genes with different HG counts. 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.

188 In addition to chromosome ploidy, HG counts also had dosage effects on gene expression as
189 demonstrated by all but two comparisons (C3-C7, **Fig. 3E-I**): genes with more HGs tended to
190 have higher expression levels than their RBM counterparts with fewer HGs. The other two
191 comparisons (C2, C8) indicated expression compensation when only one HG was lost: no
192 significant differences in expression between 2/3 and 3/3 genes, and between 3/4 and 4/4
193 genes (**Fig. 3D,J)**, but significantly higher averaged levels of 3/4 genes than 4/4 genes
194 (p=0.071, corresponding to p=0.035 of one-sided Wilcoxon rank sum test, **Fig. 3K**). These
195 results suggest that buffering mechanisms counteract the disruption of homeostasis caused by
196 immediate HChr-specific gene losses (i.e., 2/3 and 3/4 genes), but not that caused by
197 chromosome inequivalence.

198 Genetic bases of key trait polymorphisms

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

208 Key regulators of these traits have been previously identified (see **Table S7** for their gene IDs
209 in the *P.* sa genome), including *PeAGL6-2* (*AGAMOUS-LIKE 6-2*) in lip organ identity
210 determination[27], *PeMYB2*, *PeMYB12* and *PeMYB11* in the formation of floral background
211 color, venation-associated stripes and spots, respectively[7], and *MYBx1* in repressing
212 anthocyanin accumulation in non-patched regions[8–10]. Their functions were relatively
213 conserved in *P*. sa and H80 according to their expression profiles (**Fig. 4B, Fig. 5A,B**). Thus,
214 we focused on identifying genetic variants within or near these genes that were significantly
215 associated with corresponding trait polymorphisms, and assessing how they contributed to
216 trait differentiation in H80. Since H80 individuals were shown to be aneuploids as well by
217 karyotype analysis (**Fig. S2F**, 2n=64, >3x), they were treated as tetraploids in the GWAS to
218 facilitate the analysis. Moreover, given that 70 chromosomes in *P.* sa can not be readily
219 phased into subgenomes (**Fig. 1C**), the original HChr designations (such as that

HChrG01a~HChrG19a were designated as HChr set a) were retained when serving as the reference genome during the GWAS.

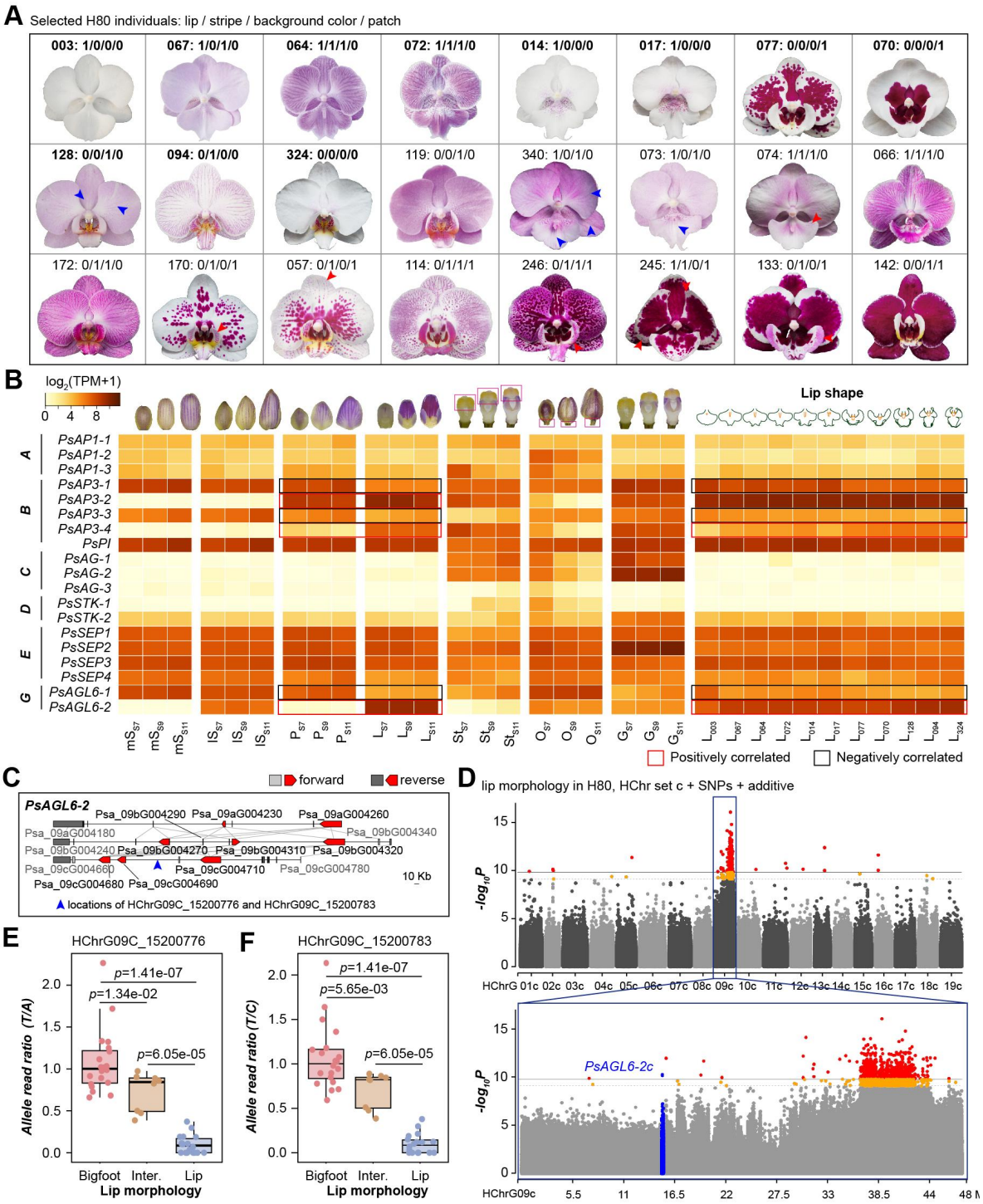

Fig. 4 Floral trait variation in H80 and the genetic basis of lip morphology variation. (**A**) Phenotypes of 24 H80 individuals (see **Fig. S1** for all 46 individuals). The category of each individual in lip morphology (0: lip, 1: petal-like organ), stripe P/A (0: A, 1: P), background color (0: white, 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 $-\log_{10}(0.01/\text{variant counts})$ and $-\log_{10}(0.05/\text{variant counts})$ (**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.

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

PsAGL6-2 had 2-4 tandemly arranged duplicates on three HChrs (corresponding to HGG type 3, **Fig. 4C**), with only one copy encoding a full-length protein per HChr (**Fig. S6**). GWAS analysis for the lip morphology (lips vs petal-like organs) identified two significant SNPs in the intergenic region between two *PsAGL6-2* copies (**Fig. 4C**) only when HChr set c was used as the reference (**Fig. 4D, Fig. S7-9**). The allele read ratios (the number of reads supporting one HG divided by the number of reads supporting the other HG) of these two SNPs were significantly associated with lip morphology in H80 (**Fig. 4E,F**, **Table S8**). Expression levels of *PsAGL6-2* showed positive correlation with lip morphology (**Fig. 4B**). These results suggest that the transformation of lips into petal-like organs in H80 is associated with reduced expression of *PsAGL6-2*, potentially due to the introduction of an allele containing mutations in a long-distance enhancer (~65 Kb upstream of *Psa_09cG004690* and ~86 Kb downstream of *Psa_09cG004710*, **Fig. 4C**). How these mutations reduced *PsAGL6-2* expression remained to be determined, potentially through genome editing of these loci in cultivars bearing lips or petal-like organs. Additionally, expression of other proposed lip (*PsAPETALA3-4* [*PsAP3-4*]) and petal identity genes (*PsAP3-1*, *PsAP3-3*, and *PsAGL6-1*; see **Fig. S10** for the gene tree of MADS-box genes)[28] exhibited positive and negative correlations with petal-to-lip transitions, respectively (**Fig. 4B**). These results suggest that expression of these genes might be regulated by *PsAGL6-2*, thereby coordinating lip morphogenesis.

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

Expression of *PsMYB12* genes was highly associated with stripe formation in *P.* sa and the

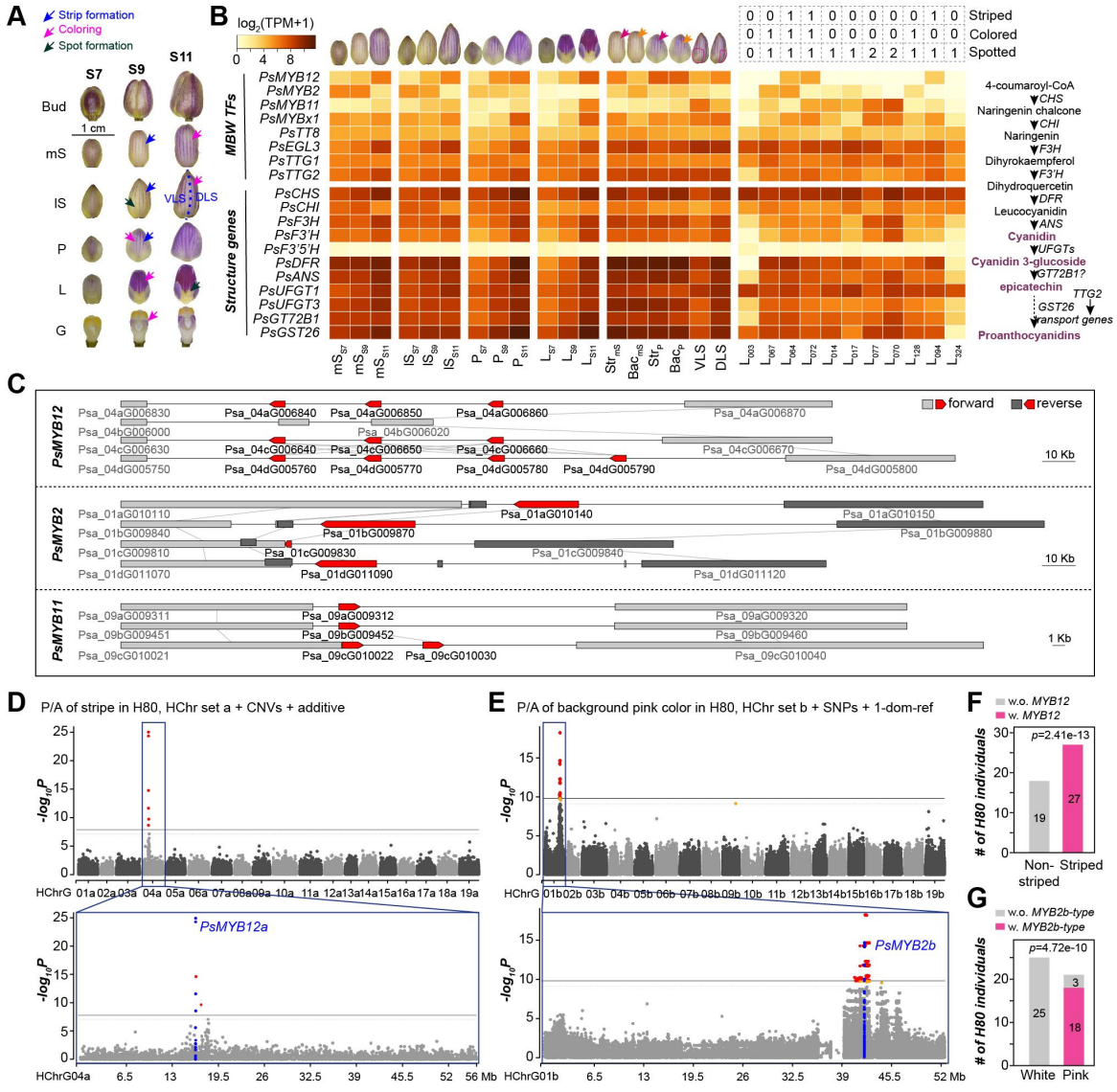

Fig. 5 Genetic bases of color patterning variation in H80. (**A**) Floral organs at three developmental stages showing important events of color patterning. VLS and DLS: ventral and dorsal half of the lateral sepals, respectively. (**B**) Heatmap showing expression profiles of genes participating in 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.

presence of stripes in H80 population (**Fig. 5A,B**). It lost its entire tandemly duplicated gene cluster from HChrG04b (HGG type 2, **Fig. 5C**). CNV-based GWAS for P/A of stripes (**Fig. 5D, Methods**) using HChr set a as the reference identified significant variants located in and

281 nearby *PsMYB12* genes while with the fewest false positives compared with others (**Fig. S11**).
282 Consistent with this, stripes P/A in H80 was perfectly associated with P/A of resequencing
283 reads mapped to *PsMYB12* genes (**Fig. 5F**, **Fig. S12, Table S9**), suggesting that stripe P/A
284 were determined by *PeMYB12* P/A.

285 Since we didn't identify any genes orthologous to *PsMYB12* in the other eight orchid species
286 (including the non-striped *P. aphrodite*) except for *P. equestris* (**Fig. S13A**), we asked when
287 and how the *PsMYB12* genes and stripes originated in *Phalaenopsis*. Interestingly, stripes
288 were only observed for the branch leading to *P. equestris* and *P. lindenii* in wild
289 *Phalaenopsis* species (**Fig. S13B**). To test if *PsMYB12* genes originated before or after the
290 most recent common ancestor (MRCA) of *Phalaenopsis*, we resequenced the *Vanda*
291 'Pakchong Blue', a stripe-bearing orchid cultivar outside *Phalaenopsis*. We detected reads
292 mapped to *PsMYB12* genes, albeit with a high degree of divergence within and near these
293 genes (**Fig. S13C**). Given that other orchids like *Paphiopedilum* also have stripes, these
294 results suggest that *MYB12* genes originated in other orchids but were absent in the MRCA of
295 *Phalaenopsis* and were introduced into the MRCA of *P. equestris* and *P. lindenii*, potentially
296 via gene introgression.

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

298 *PsMYB2* had a single copy per HChr (**Fig. 5C**), but the one on HChrG01c (*Psa_01cG009830*)
299 had lost its R2 domain (HGG type 5, **Fig. S14A**). GWAS for the P/A of background pink
300 color using 1-dom-ref model (**Methods**) with HChr set b as the reference identified
301 significant SNPs and indels located in and nearby *PsMYB2* genes while with the fewest false
302 positives among others (**Fig. 5E**, **Fig. S15-17**). In this 1-dom-ref model, the *PsMYB2b*-type
303 HG was considered as dominant in determining the pink color, consistent with its higher
304 expression levels than other HGs (**Fig. S14B**). The P/A of *PsMYB2b*-type sequences can
305 nearly perfectly predict the P/A of background pink color on sepals/petals for all but three
306 individuals—H80-066, H80-072 and H80-114 (**Fig. 5G**, **Table S10**). The background color
307 in these exceptions may be governed by other non-*PsMYB2* genes, since *PsMYB2* expression
308 was almost undetectable in the lip of H80-072 (the summed TPM of four HGs was 1.32, **Fig.
309 5B**, **Fig. S14B**). The protein sequence of PsMYB2b differed from PsMYB2a by only two
310 amino acids in the C-terminal region (**Fig. S14A**), suggesting that the pink color in *P.* sa and
311 most H80 individuals with pink flowers was likely determined by the *PsMYB2b*-specific
312 expression. It requires further investigation of which region of the *PsMYB2b* promoter drives
313 its specific expression.

314 **Spot/patch pattern variation**

315 *PsMYB11* had an extra tandem copy (*Psa_09cG010030*) on HChrG09c (HGG type 3, **Fig.**
316 **5C**), which harbored a few mutations in protein sequences and the upstream sequences (**Fig.**
317 **S14C,D**) compared with other HGs. Expression of *PsMYB11* genes, especially
318 *Psa_09cG010030*, was highly associated with the spot formation in *P*. sa and the spot/patch
319 pattern variation in H80 (**Fig. 4A**, **Fig. S14E**). GWAS for the P/A of patches identified
320 numerous variants distributed across the entire genome, but no variants were within or near
321 the *PsMYB11* genes (>96.8 Kb), regardless of variant types used and HChr sets referred to
322 (**Fig. S18-20**). Given that the patch formation was reported to be determined by the *HORT1*
323 insertion in the promoter sequences of *PeMYB11* genes[9], and that *P.* sa flowers harbored only
324 small spots, the failure to identify *PsMYB11*-related variants is potentially due to the lack of
325 *HORT1* insertion in *PsMYB11* promoters.

326 **Diverse functions of *PsMYBx1* in repressing anthocyanin accumulation**

327 *PsMYBx1* had only one HG (HGG type 1, **Fig. 6A**). Its expression was associated with both
328 stripe and spot/patch formation in *P.* sa and H80 (**Fig. 5A,B**), indicating potentially dual
329 function of *PsMYBx1* in repressing anthocyanin accumulation in both inter-stripe and inter-
330 spot/patch regions. This dual function was supported by the disruption of stripes in regions
331 surrounding patches, such as in H80-057, H80-114, and H80-245 (**Fig. 4A**). Given this
332 complexity of color patterning imposed by *PsMYBx1* genes, no single color patterning
333 variation can be readily determined to be related to *PsMYBx1*. Thus, instead of performing
334 GWAS, we silenced *MYBx1* genes via virus-induced gene silencing (**Methods**) in another
335 cultivar, *P.* 'Jinbianlinglong' (*P.* ji) (**Fig. 6B,C**), for which substantial seedlings are available
336 and the flowering regulation is relatively readily. *MYBx1* had four HGs in *P*. ji (**Fig. S21B**),
337 compared with a single HG in *P*. sa. *PjPDS* (*Phytoene desaturase*) was co-silenced to
338 indicate regions with gene silencing phenotypes where the yellow-green color was faded (**Fig.**
339 **6B**).

340 In *PjMYBx1-PjPDS*-silenced plants, ectopic anthocyanin pigmentation was observed on the
341 inflorescence stalk, abaxial surface of perianth, and in the inter-stripe regions (**Fig. 6C**).
342 However, the inter-spot regions remained unchanged (blue arrowheads in **Fig. 6C** where the
343 yellow-green color was faded), contrasting with the proposed repressing role of *MYBx1* in
344 these regions[10]. Although expression of color patterning-related genes examined were
345 generally decreased in *PjMYBxl-PjPDS*-silenced sepals (**Fig. 6D**) and petals (**Fig. 6E**)
346 compared with those in mock-treated plants, a slight increase in *PjMYB12* expression was

detected in petals. Taken together, these results demonstrate the multiple functions of *MYBx1* in repressing anthocyanin accumulation in diverse regions, and suggest variation in its functions across different *Phalaenopsis* orchids.

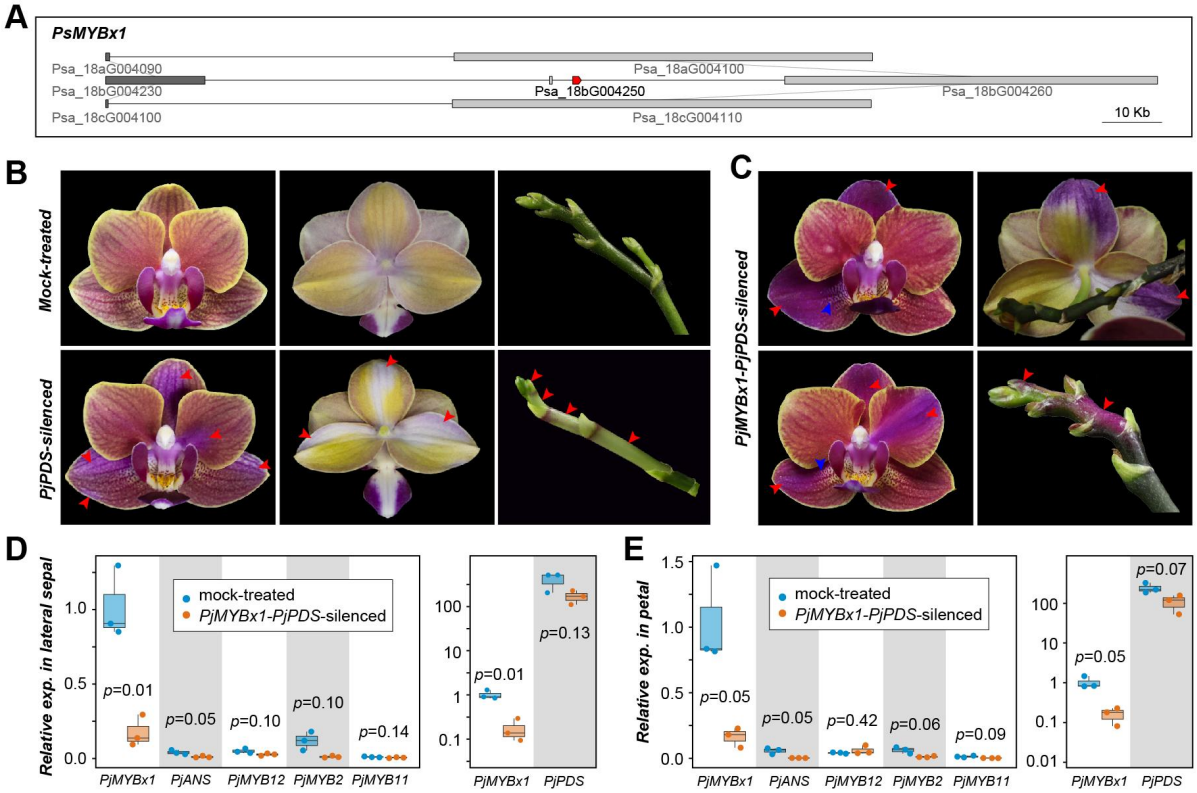

Fig. 6 Functions of *PjMYBx1* in repressing anthocyanin accumulation in diverse regions. (**A**) HG variation of *PsMYBx1*. (**B**) Flower and inflorescence images showing phenotypes of mock-treated (upper panel) and *PjPDS*-silenced (lower panel) plants. (**C**) Flower and inflorescence images of *PjMYBx1-PjPDS*-silenced 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.

Discussion

The haplotype-resolved, chromosome-level genome of *P*. sa served as a suitable reference in revealing the genetic bases of trait diversity in *Phalaenopsis*. Given the high heterogeneity of *P*. sa genome and the diverse genetic mechanisms underlying different trait variation, we demonstrated that the choice of HChr sets as the reference genome, variant types, and marker-effect models is critical for revealing the genetic bases of target trait variation. The ratio of true to false positives of the identified significant variants can be the primary metric for optimization. Specifically, for lip morphology variation, stripe P/A variation, and pink color P/A variation, the optimal models were HChr set c + SNPs + additive model, HChr set a + CNVs + additive model, HChr set b + SNPs/indels + 1-dom-ref model, respectively.

However, for spot/patch variation, genomes of patch-bearing *Phalaenopsis* cultivars are supposed to serve as superior references compared with *P.* sa genome.

Although genes regulating some of the key traits in *Phalaenopsis* have been well established[1,2,7,12,27,29,30], what and how allelic polymorphisms of these genes have contributed to the extraordinary trait diversity remain underexplored, restricting their utilization in breeding practices. The genetic bases of three key trait variation revealed in this study—nucleotide polymorphisms nearby *PsAGL6-2* genes, P/A of *PsMYB12* genes, and P/A of *b*-type HG of *PsMYB2*—enable design of a multiplex polymerase chain reaction (MPCR) kit for early-stage prediction of these three traits simultaneously. Given the unusually long juvenile phases of *Phalaenopsis* orchids (3-5 years), such a MPCR kit has great potential in substantially shortening breeding cycles and reducing breeding costs.

Methods

Flow cytometry analysis

Young leaves of *P.* sa (ANTHURA BV, with US patent US20150047084P1) seedlings were used for the flow cytometry analysis, with young leaves of cultivated tomato one month after seed germination being used as the internal reference. Suspensions of *P.* sa and tomato leaves were strained using propidium iodide and mixed, and a BD FACScalibur flow cytometer was used to detect the stained cell nuclei.

Karyotype analysis

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

398 ## Genome sequencing and Hi-C sequencing

399 For HiFi sequences, the total DNA was extracted from *P.* sa young leaves, and used for the
400 library construction and genome sequencing with the PacBio Sequel II/Sequel IIe platform,
401 following the manufacturer's instructions at Haorui Genomics (Xi'an, China). Approximately
402 72.1 Gb of PacBio HiFi reads were generated using the CCS mode.

403 For Hi-C sequences, the frozen young leaves were ground with liquid nitrogen and then fixed
404 with 1% formaldehyde in vacuum at room temperature for 0.5-1h to preserve the three-
405 dimensional chromatin architecture, and 0.15 M glycine was added in the solution for 5 min
406 to quench the reaction. The mixed solution was then filtered with a 40 nm cell filter.
407 Chromatin was digested with 200 U DPNII restriction enzyme (Qiagen, 59971) at 37 ℃ for
408 16 h. DNA ends were labeled with biotin and incubated at 37 ℃ for 45 min. DNA ligation
409 was performed by the addition of T4 DNA ligase and incubation for 1-2 h. Then proteinase K
410 was added to reverse cross-linking during incubation at 65 ℃ for 1-2 h. DNA fragments were
411 purified and dissolved in 86 μL of water. Purified DNA was fragmented to a size of 350-500
412 bp, and DNA ends were then repaired. DNA fragments labeled with biotin were finally
413 separated on Dynabeads® M-280 Streptavidin (Life Technologies, 60210). The sequencing
414 was performed with the Illumina NovaSeq 6000 for paired-end sequences with 150 bp read
415 length. A total of ~130.6 Gb raw data was produced. After filtering adapter sequences and
416 read pairs with > 10% N or with low quality (Q ≤ 5) for >50% bases in at least one read,
417 ~123.3 Gb clean data was retained.

418 ## Genome survey

419 The genome size was estimated via *k*-mer frequency analysis using the PacBio HiFi reads.
420 The Jellyfish software (v2.2.10)[31] was used to generate the counts of 31-mers, and then the
421 GenomeScope2 software (v2.0)[32] and Genomic Character Estimator (GCE) software
422 (v1.0.2)[33] were applied to estimate the genome size and heterozygosity. SmudgePlot software
423 (v0.2.2)[32] was used to estimate the ploidy level of *P.* sa genome.

424 ## *De novo* genome assembly

425 Since it lacks a chromosome-level reference assembly for *Phalaenopsis* species, at the first
426 place, we attempted to assemble a set of draft representative pseudo-chromosomes to guide
427 the subsequent refined assembling. Three strategies were conducted to assemble the draft
428 genome: (**i**) all the PacBio HiFi reads were assembled using HiFiasm program (v0.19.6-
429 r597)[34] with the parameter 'l=0' (to disable the purge of duplication); (**ii**) the PacBio HiFi

430 reads were assembled using HiFiasm with the parameters 'l=3, hom-cov=57, hg-size=1.3g, n-
431 hap=3, s=0.3, D=8, n-weight=6, n-perturb=80,000, f-perturb=0.3', with the assumption that
432 the *P.* sa was a triploid, and the Hi-C reads were integrated to achieve haplotype-resolved
433 assemblies; (**iii**) the PacBio HiFi reads were assembled with parameters 'l=3, hom-cov=39,
434 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
435 assumption that the *P.* sa was a tetraploid, and the Hi-C reads were also integrated.

436 The Hi-C reads were mapped to the resulting primary unitigs from assembling **i**, and
437 haplotigs of each haplotype from assembling **ii** and **iii**, separately, using bwa-mem2 (v2.2.1)[35].
438 The resulting bam files were further filtered using the 'filter_bam.py' utility in the HapHic
439 (v1.0.2)[36] with the parameters 'mapq=1, NM=3'. The filtered bam files were then fed into the
440 HapHic pipeline with parameters 'correct_nrounds=2, quick_view=T', and the output
441 files .hic and .assembly were imported into the Juicebox Assembly Tools (JBAT, v1.08)[37].
442 Primary unitigs from strategy (i) and haplotigs from (ii) and (iii) were then ordered, joined,
443 and manually adjusted based on Hi-C interaction patterns to obtain scaffolds, with the prior
444 knowledge that nearby regions within the same chromosome having strong Hi-C signals
445 whereas allelic regions having subdominant signals. The resulting scaffolds from all three
446 strategies were compared, and 19 were selected as representative pseudo-chromosomes based
447 on Hi-C signal continuity and Benchmarking Universal Single-Copy Orthologs (BUSCO) [18]
448 completeness. These 19 pseudo-chromosomes—seven from assembly **i**, ten from **ii**, and two
449 from **iii**—constitute the draft reference assembly (assembly v1.0).

450 Next, using the v1 draft assembly as a reference, we mapped all raw unitigs from assembly **ii**
451 and grouped them into 19 chromosomal groups. Unitigs within each group were ordered,
452 joined, and manually adjusted based on Hi-C interaction patterns, with the parameters
453 'mapq=0, NM=3' for filter_bam.py. This produced 70 preliminary pseudo-chromosomes.
454 Remaining unassigned unitigs were then incorporated and the entire set was reviewed and
455 corrected using Hi-C signals to yield the final assembly (assembly v2.0). SubPhase (v1.2.6)[38]
456 was used to phase subgenomes based on potential subgenome-specific *k*-mers, with
457 parameters "-k 15 -q 50 -f 12".

458 Genome assembly quality assessment

459 The *P.* sa genome assembly contiguity was accessed using Quality Assessment Tool for
460 Genome Assemblies (QUAST, v5.0.2)[39]. The BUSCO completeness was conducted using
461 BUSCO (v5.6.1)[18], against the eukaryota_odb12, viridiplantae_odb12, embrophyta_odb12,

462 and liliopsida_odb12 databases. The *K*-mer (21-mers) completeness and assembly consensus
463 quality value (QV) were estimated using Merqury (v1.4.1)[17]. The QV is a log-scaled
464 probability of errors for consensus base calls; the higher the QV values, the more accurate the
465 consensus base calls, with a QV value at 30 indicating 99.9% accuracy and 40 representing
466 99.99% accuracy[17]. The regional and structural assembly quality indicator (R-AQI and S-AQI)
467 were assessed using the Clipping information for Revealing Assembly Quality (CRAQ,
468 v1.0.9) algorithm [16]; the higher the AQI scores, the more accurate the genome assembly[16].
469 The long terminal repeat retrotransposons (LTR-RTs) were first annotated using LTRharvest
470 (v1.6.1)[40] with the parameters '-maxlenltr 6000 -mindistltr 1500 -maxdistltr 25000 -mintsd 5
471 -maxtsd 5 -vic 10 -similar 85' and LTR_FINDER_parallel (v1.2)[41] with the parameters '-size
472 1000000 -time 300'. Then the LTR Assembly Index (LAI)[42] was calculated using
473 LTR_retriever (v2.9.5)[43] with the parameter '-u 0.5e-8'. The telomere and centromeric repeat
474 identification were conducted using quartet (v1.2.5) using default parameter settings[44].

475 Noncoding RNA and repetitive sequence identification

476 Noncoding RNAs were predicted using tRNAscan-SE (v2.0.9)[45] and Infernal software
477 (v1.1.4)[46] by searching the database Rfam (v14.10). The extensive *de-novo* TE annotator
478 (EDTA) (v2.1.0)[47]—that encompasses LTR_FINDER, LTRharvest, LTR_retriver, Generic
479 Repeat Finder, TIR-Learner, HelitronScanner, RepeatModeler and RepeatMasker— was used
480 to identify, annotate and classify repetitive elements of the genome, with parameters '--anno
481 1 --sensitive 1 --evaluate 1 --u 0.5e-8 --curatedlib ${SINE_LINE_lib}'. The library of short
482 interspersed elements (SINEs) and long interspersed nuclear elements (LINEs) was
483 downloaded from SINEBase (https://sines.eimb.ru/) as a compensation for the reference
484 sequences. The repetitive elements identified by EDTA included long terminal repeat (LTR)
485 retrotransposons (copia, gypsy and unknown ones) and non-LTR retrotransposons
486 (LINE_element and unknown ones), terminal inverted repeat (TIR) transposons (CACTA,
487 Mutator, PIF_Harbinger, Tc1_mariner and hAT), *Helitron* and other interspersed repeats.
488 Subsequently, the DeepTE (version updated on Oct 22th, 2022) script[48] was used to reclassify
489 the TEs that were initially annotated as 'LTR/unknown' by EDTA into copia, gypsy and
490 unknown classes, then the output file was fed back into EDTA to enhance the
491 comprehensiveness of TE annotation. To calculate LTR insertion time, we set the mutation
492 rate per generation (μ) to 0.5×10^{-8}, based on the assumption that the generation times of
493 orchids are four years[49]. The resulting hard-masked genome was converted into a soft-masked
494 genome, which was used in the subsequent analysis. The tandem repeat sequences in the *P.* sa

genome were identified using Tandem Repeats Finder (TRF, v4.09)[50] with the parameters '2 7 7 80 10 50 2000 -d -h'. Then the proportion of tandem repeats in the genome was calculated using the script buildRepeatSummary.pl (v1.36) with default parameters[51].

Transcriptome sequencing

Total RNA was extracted for roots, leaves, stems and flower buds with mixed stages (one duplicate) and other tissues at different development stages (three replicates, **Table S6**). The library preparation and PCR amplification were performed according to the manufacturer's instructions, and the RNA libraries were sequenced on the Illumina Novaseq 6000 platform and 150 bp paired-end reads were generated. Clean data were further obtained by removing reads containing adapters, reads containing poly-N and low quality reads from raw data, using fastp (v0.24.1)[52] with parameters: '-g -q 5 -u 50 -n 15 -l 150 --overlap_diff_limit 1 --overlap_diff_percent_limit 10'.

Protein-coding gene prediction

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

1. Homology-based prediction

Genome assembly and General Feature Format (GFF) files of six plant species were used for homology-based prediction with Gene Model Mapper (GeMoMa, v1.9)[53]. The following species were included: *Arabidopsis thaliana* (TAIR10.1, downloaded from NCBI), *Oryza sativa* Japonica Group (IRGSP_1.0, from Ensembl Plant), *Asparagus officinalis* (Aspof.V1, from Ensembl Plant), *Apostasia shenzhenica* (GCA_002786265.1, from NCBI)[54], *Dendrobium catenatum* (GCF_001605985.2, from NCBI)[55], and *Phalaenopsis equestris* (GCF_001263595.1, from NCBI)[14]. The GeMoMa parameters used were 'AnnotationFinalizer.r=NO p=false o=true tblastn=false'.

2. Transcriptome-based prediction

Clean RNA-seq reads from roots, leaves, stems, and mixed-stage flower buds were aligned to *P. sa* genome using HISAT2 (hierarchical indexing for spliced alignment of transcripts, v2.2.1)[56]. The alignments were processed by StringTie (v2.2.1)[57] to reconstruct genome-guided transcripts. Potential open reading frames (ORFs, ≥100 amino acids) were identified from these transcripts using TransDecoder (v5.7.1) (https://github.com/TransDecoder). A

525 genome-guided transcriptome was assembled using Trinity (v2.15.1)[58]. Both the StringTie
526 output and the Trinity assembly were then analyzed by the Program to Assemble Spliced
527 Alignments (PASA, v2.5.3)[59] with the parameters '--ALIGNERS blat,minimap2 --
528 TRANSDECODER' to generate transcriptome-supported gene models.

529 3. *Ab initio* prediction

530 The BRAKER3 pipeline (v3.0.3, incorporating Augustus and Genemark)[60], GeneID (2021
531 version)[61] and GlimmerHMM (2014 version)[62] were used for the *ab initio* gene prediction,
532 leveraging the transcriptome evidence as a guide.

533 4. Gene model integration and filtering

534 The EvidenceModeler (v2.1.0)[63,64] was used to integrate predictions from the above methods
535 into consensus gene structures. The following weights were assigned: GeneID (1),
536 GlimmerHMM (1), GeneMark (2), Augustus (4), GeMoMa (7), StringTie (5), and
537 PASA_transdecoder (10).

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

540 transposon_ORF_lib/transposon_db.pep) using TransposonPSI (v1.0.0), which employs PSI-
541 BLAST to identify TE-related sequences[65] with default parameters. Proteins with > 30%
542 similarity and > 60% coverage to known TE sequences were removed. In addition, gene
543 models with incomplete ORFs or ORFs ≤ 100 amino acids were excluded, if they met all the
544 three following conditions: lacking RNA-seq expression support, having < 50% coverage
545 overlap with GeMoMa models, and lacking functional annotation by InterProScan (v5.67-
546 99.0)[66].

547 5. Model optimization with extended RNA-seq data

548 The resulting gene models (v1.0) from above steps were further supplemented using
549 additional RNA-seq data from other 28 samples (**Table S6**). First, to avoid potential mapping
550 errors between tandem duplicates with high sequence similarities, RNA-seq reads were first
551 mapped to the v1.0 transcriptome, and unmapped reads were further mapped to the genome,
552 using TopHat2 (v2.1.1)[67] with the parameter '--max-multihits 1'. Next, individual bam files
553 were used to assemble transcripts (including alternative splicing isoforms) using StringTie.
554 The resulting GTF files from all samples were merged using StringTie to create a

555 comprehensive transcript annotation. TransDecoder was used to infer the ORFs, producing a
556 GFF file (v2.0) that included alternative splicing and five prime untranslated regions (5'UTR)
557 and 3'UTR information. Finally, the v2.0 annotations were compared with and used to
558 correct the v1.0 annotations, prioritizing v2.0 evidence in overlapping regions. This process
559 resulted in a final set of 93,708 protein-coding genes.

560 Protein-coding gene functional annotation

561 *P.* sa protein sequences were first searched against the eggNOG database (v5.0.2)[68] using
562 eggNOG-mapper (v2.1.12)[69], with parameters '-m diamond --seed_ortholog_evalue 1e-5 --
563 tax_scope Viridiplantae', for annotations including Gene Ontology (GO), Kyoto
564 Encyclopedia of Genes and Genomes (KEGG) Orthology (KO), and Clusters of Orthologous
565 Groups (COG). In order to obtain the functional description of related protein-coding genes
566 in model plant species, protein sequences of *P.* sa genome were searched against the UniProt
567 database (uniprot_sprot_plants.fasta that contains manually annotated protein sequences from
568 sequenced plant genomes, updated on March 27th, 2024, https://www.uniprot.org/) and
569 protein sequences from Arabidopsis (Araport11_pep_20240409, downloaded from The
570 Arabidopsis Information Resource) using BLASTP (v2.12.0+)[70] with parameters '-max_hsps
571 1 -max_target_seqs 1 -evalue 1e-5'. Furthermore, *P.* sa protein sequences were searched
572 against databases such as Pfam, SMART, CDD, Superfamily and others, using InterProScan
573 (v5.67-99.0) with parameters '-goterms -iprlookup'. Gene Functional Annotation for Plants
574 (GFAP, v3.1)[71] was used to annotate GO, KEGG and Pfam items by searching *P.* sa protein
575 sequences against a plant-specific database with the parameter '-awd psd -go -kegg -pfam',
576 and to obtain homology-based annotations by searching against the plant-specific NCBI non-
577 redundant (nr) database with the parameter '-awd nr'. Lastly, protein sequences were
578 searched against the Reference Protein Sequence Dataset (RPSD, release_2024_07_31) using
579 Ensemble Enzyme Prediction Pipeline (E2P2, v5)[72] with default parameters for the enzyme
580 function annotation. Transcription factors in *P.* sa genome were annotated in the Plant
581 Transcription Factor Database (PlantTFDB)[73].

582 Homoeologous gene identification

583 The Synteny and Rearrangement Identifier (SyRI, v1.7.0)[74] and MCScanX v1.0.0[75] were used
584 to conduct synteny analysis among different HChrs with default parameter settings, based on
585 genomic and protein sequences, respectively. Later on, the homoeologous genes were
586 identified with in-house scripts (see **Code availability**) using the outputs of SyRI

587 preferentially and MCScanX supplementarily, to avoid improper assignment of multiple
588 MCScanX short syntenic blocks between HChrs.

589 Transcriptome analysis

590 RNA-seq reads were mapped to the transcriptome and genome using Tophat2 (v2.1.1)[67], with
591 parameter '--max-multihits 1', using indexes from both the genome and the final
592 transcriptome. Read counts for protein-coding genes were obtained using htseq-count
593 (v2.0.5)[76] and TPM was calculated using the "calculateTPM" function in R package scater
594 (v1.0.4)[77]. The TPM values were normalized across samples using quantile normalization,
595 followed by averaging across the three biological replicates to obtain a mean TPM value for
596 each tissue.

597 To categorize genes according to their expression levels, we defined four thresholds based on
598 TPM values: 1) TPM = 0 indicating no expression; 2) 0 < TPM ≤ 2 indicating low expression;
599 3) 2 < TPM ≤ 10 indicating moderate expression; 4) TPM > 10 indicating high expression.
600 The threshold of TPM = 10 distinguishing high expression was determined by modeling the
601 observed gene expression data as a mixed distribution. This model comprises a normal
602 distribution, representing true biological expression, and an exponentially decaying
603 distribution, representing technical noise or background. These two components were fitted
604 using the nls function from the R "stats" package[78]. The value of TPM = 10 corresponds
605 approximately to the 50th percentile of the fitted normal distribution, suggesting that roughly
606 half of the genuine gene expression occurs below this level.

607 Weighted gene co-expression network analysis (WGCNA)

608 Co-expression analysis for HG type 1 genes that were highly expressed (TPM > 10) in ≥ 1
609 tissue and had no non-self blast hist in *P.* sa genome was performed using WGCNA
610 (v1.72.1)[79], under a signed network model, with a minimum module size of 10.

611 Gene tree building

612 To build a robust phylogeny tree for *MYB2*, *MYB11* and *MYB12* genes, coding sequences
613 (CDS) were obtained from multiple orchid species—*P. aphrodite* (*P. ap*), *P. equestris* (*P. eq*),
614 *Apostasia shenzhenica* (*A. sh*), *Cremastra appendiculata* (*C. ap*), *Cymbidium sinense* (*C. si*),
615 *Dendrobium chrysotoxum* (*D. ch*), *Gastrodia elata* (*G. el*), *Platanthera guangdongensis* (*P.*
616 *gu*), *Vanilla planifolia* (*V. pl*)—via the Orchid Multiomics Database
617 (http://orchidcomics.com/, access on June 11th, 2025)[80]. For each query gene, the top 20

618 BLAST hits per species (top 80 hits for *P.* sa) were extracted, translated into amino acid
619 sequences, and aligned using MEGA v7.0[81]. The resulting amino acid alignment was
620 manually curated to retain only reliably alignable regions, which were then converted back to
621 a corresponding CDS alignment. The best-fit nucleotide substitution model (GTR+I+G4) was
622 selected using ModelTest-NG (v0.1.7)[82]. Finally, a maximum likelihood phylogenetic tree
623 was constructed with RAxML-NG (v1.2.2)[83], employing 1000 bootstrap replicates.

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

628 Genetic variant calling

629 Total DNA was extracted for flower stalks of 46 H80 (*P.* '1071-13' × *P.* 'CH1804-01')
630 individuals. Sequencing services of partial H80 individuals were provided by BGI Genomics
631 (Shenzhen, China), and the others were provided by Novogene Co., Ltd (Beijing, China). All
632 whole-genome resequencing in this study was performed on the DNBSEQ-T7 platform using
633 a 150-bp paired-end strategy, generating ~60 Gb of data per sample. Raw reads were
634 processed using fastp (v0.24.1)[52] with parameters '-g -q 5 -u 50 -n 15 -l 150 --
635 overlap_diff_limit 1 --overlap_diff_percent_limit 10'. Clean reads were aligned to the *P.* sa
636 genome using BWA-MEM algorithm in Burrow-Wheeler Aligner (BWA, 0.7.18-r1243-
637 dirty)[84] with default settings, followed by sorting with SAMtools (v1.21)[85]. Duplicate reads
638 were marked using the MarkDuplicates module in the Genome Analysis Toolkit (GATK,
639 v4.6.1.0)[86].

640 Population-level SNP and indel calling was conducted using GATK. First, variants were
641 called per sample with HaplotypeCaller using '-ERC GVCF -ploidy 4'. Individual gVCF files
642 were merged with CombineGVCFs and jointly genotyped with GenotypeGVCFs. SNPs and
643 indels were separated using SelectVariants and filtered with VariantFiltration using the
644 following criteria: for SNPs, 'QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum
645 < -12.5 || ReadPosRankSum < -8.0'; for indels, 'QD < 2.0 || FS > 200.0 || MQ < 40.0 || SOR >
646 10.0 || ReadPosRankSum < -20.0'. Multiallelic sites were split into biallelic records using
647 BCFtools norm (v1.21)[85] with '--multiallelics -both'.

648 SVs were called and genotyped using smoove (v0.2.8)[87] with default parameters. CNVs were
649 detected with CNVpytor (v1.3.2)[88] using a bin size of 100. Finally, all variants (SNPs, indels,
650 SVs and CNVs) were annotated using ANNOVAR (v2020-06-07)[89].

651 Genome-wide association study (GWAS) analysis

652 GWAS was performed using the R package GWASpoly (v2.13)[90] based on the Q + K mixed
653 model. First, the VCF files were transformed into dosage files using VCF2dosage function
654 with parameters '-ploidy 4 -min.DP 8 -max.missing 0.5 -min.minor 3'. To control for
655 population structure, the covariance matrix for polygenic effect was calculated using the
656 set.K function with the parameter 'LOCO=TRUE', where a different covariance matrix was
657 calculated for each chromosome according to markers from all other chromosomes[91]. During
658 this step, missing data were imputed using the k-Nearest Neighbors approach implemented
659 by the KNNImputer function from the sklearn.impute module[92]. The imputation was
660 performed five times with different k values (2, 5, 10, 15, and 20), and the average of the five
661 imputed values was used. The set.params function was used to set the threshold of minor
662 allele frequency with the parameter 'geno.freq = 1 - 5/N, MAF=0.05'. The GWASpoly
663 function was then used to assess the significance of each marker associated with target trait
664 variation under the additive model and two simplex dominance models (1-dom-ref and 1-
665 dom-alt) hypothesis. In the additive model, the marker effect scales linearly with the alternate
666 allele dosage. For the simplex dominance models, a single copy of the dominant allele
667 (reference or alternate allele) is sufficient for the complete dominance phenotype, thereby all
668 heterozygotes possessing this allele exhibit equivalent marker effects. Two significance
669 thresholds were adopted: p-value<0.05/variant counts and p-value<0.01/variant counts,
670 according to the Bonferroni adjustment approach. Lastly, the CMplot (v4.5.1)[93] package was
671 used to generate the Manhattan plots. For better visualization, $-\log_{10}(p)$ values greater than 30
672 were truncated to 30 in the Manhattan plots.

673 Virus-induced gene silencing

674 Seedlings of *P.* 'Jinbianlinglong' (*P.* ji) were cultivated in a growth room under a 14-h
675 light/10-h dark photoperiod, with a light intensity of 220 $\mu mol \cdot m^{-2} \cdot s^{-1}$ and a temperature of
676 25–27 °C. Partial sequences of *PjPDS* and *PjMYBx1* were introduced into the TRV2-based
677 pYL156 vector, as illustrated in **Fig. S21**. The resulting TRV2 plasmids were transformed
678 into *Agrobacterium tumefaciens* strain GV3101. Bacterial cultures containing either TRV2-
679 *PDS* or TRV2-*PDS-MYBx1*, and cultures containing the TRV1 plasmids were incubated

separately. When the OD_{600} reached approximately 2.0, the cells were harvested and resuspended in an induction buffer (10 mM morpholineethanesulfonic acid, 10 mM $MgCl_2$, and 200 uM acetosyringone) to a final OD_{600} of 2.0. Each TRV2 suspension was mixed with an equal volume of the TRV1 suspension, and Silwet L-77 was added to a final concentration of 0.005%. The mixture was incubated statically at room temperature for 3 h.

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

To confirm the silencing of target genes in plants showing phenotypic changes, total RNA was extracted from the corresponding floral tissues using the FastPure Universal Plant Total RNA Isolation Kit (Vazyme, Nanjing, China). First-strand cDNA was synthesized using MightyScript Plus First Strand cDNA Synthesis Master Mix (gDNA digester) (Sangon Biotech, Shanghai, China). Quantitative reverse-transcription PCR (qRT-PCR) was performed using SupRealQ Ultra Hunter SYBR qPCR Master Mix (U+) (Vazyme). The $2^{-\Delta\Delta Ct}$ method was employed to calculate the relative expression of target and related genes, with *PeActin* serving as the internal control. Primer sequences for VIGS and qRT-PCR were 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 genes; LZ, FM and PW conducted the genome assembly quality assessment and syntenic analysis; PW and FW conducted the transcriptome analysis; PW, XZ and FW conducted the gene tree building; XZ conducted the genetic variant calling and GWAS analyses; ZW, LLin (Letian Lin) and ZL conducted the VIGS experiments; LLin performed the qRT-PCR; PW conducted all the other analyses; PW, RJ, XZ, LZ, LW, ZW and LLin wrote the draft with contributions from all the other authors. All authors read and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Supplementary figure legends

Fig. S1 Phenotypes of H80 individuals. Each panel shows an individual plant, with its classifications for the P/A of stripe, background pink color and spot/patch, and lip morphology provided on the top-left, top-right, bottom-left and bottom-right corners per

740 panel, respectively. Red arrowheads indicate stripes, while blue ones indicate venation-
741 associated shadow.

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

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

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

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

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

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

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

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

781 **Fig. S10 Gene tree for MIKC-type MADS-box genes in orchids.** NCBI accession ID (such
782 as JL017169) was listed between gene name (such as *CeMADS32*) and the reference (such as
783 2021 Ai).

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

790 **Fig. S12 IGV images showing reads of seven H80 individuals with or without stripes**
791 **that were mapped onto the *PsMYB12* gene-containing regions.** Red arrowheads indicate
792 stripes, while blue ones indicate venation-associated shadow in individuals which had no
793 intact reads mapped onto *PsMYB12* genes.

794 **Fig. S13 Evolution of color patterning-related *MYB* genes.** (**A**) Gene tree of *MYB* genes
795 from orchids with genome assemblies. (**B**) Evolutionary history of stripes on perianths in the
796 genus *Phalaenopsis*. The phylogeny of *Phalaenopsis* species were modified from our
797 previous review paper [28]. The red star indicates the origination of stripes in the most recent
798 common ancestor of *P. equestris* and *P. lindenii*. Three flower images show the venation-
799 associated distribution of spots in other *Phalaenopsis* species. (**C**) Integrative Genomics
800 Viewer images showing reads of *Vanda* 'Pakchong Blue' that were mapped onto genic
801 regions of two *PsMYB12* genes as examples.

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

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

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

820 **Fig. S17 Manhattan plot for GWAS of background color in H80 using the 1-dom-alt**
821 **marker-effect model.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs that were
822 called by referring to HChr set b were used, respectively. Red and orange dots denote
823 variants significantly associated with the P/A variation of pink color; blue dots represent
824 variants located within or adjacent (±2 kb) to *PsMYB2b*.

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

830 **Fig. S19 Manhattan plot for GWAS of patch P/A in H80 using HChr set b as the**
831 **reference.** (**A-D**) GWAS results when SNPs, indels, SVs and CNVs were used, respectively,
832 with the additive marker-effect model. Red and orange dots denote variants significantly

833 associated with the P/A variation of patch; blue dots represent variants located within or
834 adjacent (±2 kb) to *PsMYB11b*.

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

840 **Fig. S21 *PjPDS* and *PjMYBx1* sequences and the VIGS constructions.** (**A**) CDS and
841 3'UTR sequence alignment for four *PjPDS* HGs. (**B**) CDS sequence alignment for four
842 *PjMYBx1* HGs. (**C**) Diagrams showing the structure of TRV2 plasmid PYL156, and the
843 insertion locations of *PjPDS* and *PjMYBx1* sequences. *PjPDS_1* and *PjPDS_2* are sequences
844 from two *PjPDS* HGs marked with green shadow.

845 # Supplementary tables

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

848 **Table S2** Statistics of repetitive sequences in *P.* sa genome

849 **Table S3** Statistics of *P.* sa genes and the functional annotations

850 **Table S4** Statistics of noncoding RNAs in *P.* sa genome

851 **Table S5** Structural variation summary

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

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

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

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

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

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

859 # Reference

860 1. Hsiao, Y. *et al.* A novel homodimeric geranyl diphosphate synthase from the orchid
861 *Phalaenopsis bellina* lacking a DD(X)$_{2–4}$D motif. *Plant J* **55**, 719–733 (2008).
862 2. Huang, H. *et al. Terpene synthase-b* and *terpene synthase-e/f* genes produce
863 monoterpenes for *Phalaenopsis bellina* floral scent. *Front Plant Sci* **12**, 700958 (2021).
864 3. Tsai, W.-C., Kuoh, C.-S., Chuang, M.-H., Chen, W.-H. & Chen, H.-H. Four *DEF*-like
865 MADS box genes displayed distinct floral morphogenetic roles in *Phalaenopsis* orchid.
866 *Plant Cell Physiol* **45**, 831–844 (2004).
867 4. Acri-Nunes-Miranda, R. & Mondragón-Palomino, M. Expression of paralogous *SEP*-,
868 *FUL*-, *AG*- and *STK*-like MADS-box genes in wild-type and peloric *Phalaenopsis*
869 flowers. *Front Plant Sci* **5**, 76 (2014).
870 5. Pan, Z. *et al.* Flower development of *Phalaenopsis* orchid involves functionally divergent
871 *SEPALLATA*-like genes. *New Phytol* **202**, 1024–1042 (2014).
872 6. Xu, Q. *et al.* PeNAC67-PeKAN2-PeSCL23 and B-class MADS-box transcription factors
873 synergistically regulate the specialization process from petal to lip in *Phalaenopsis*
874 *equestris*. *Mol Hortic* **4**, 15 (2024).
875 7. Hsu, C.-C., Chen, Y.-Y., Tsai, W.-C., Chen, W.-H. & Chen, H.-H. Three R2R3-MYB
876 transcription factors regulate distinct floral pigmentation patterning in *Phalaenopsis* spp.
877 *Plant Physiol* **168**, 175–191 (2015).
878 8. Fu, Z. *et al.* An R3-MYB gene of *Phalaenopsis*, *MYBx1*, represses anthocyanin
879 accumulation. *Plant Growth Regul* **88**, 129–138 (2019).
880 9. Hsu, C.-C., Su, C.-J., Jeng, M.-F., Chen, W.-H. & Chen, H.-H. A *HORT1* retrotransposon
881 insertion in the *PeMYB11* promoter causes harlequin/black flowers in *Phalaenopsis*
882 orchids. *Plant Physiol* **180**, 1535–1548 (2019).
883 10. Lu, T.-W., Chen, W.-H., Chen, P.-Y., Shu, Y.-C. & Chen, H.-H. Correction: Perturbation

884 of periodic spot-generation balance leads to diversified pigmentation patterning of
885 harlequin *Phalaenopsis* orchids: in silico prediction. *BMC Plant Biol* **24**, 782 (2024).
886 11. Chen, W., Hsu, W., Hsu, H. & Yang, C. A tetraspanin gene regulating auxin response
887 and affecting orchid perianth size and various plant developmental processes. *Plant*
888 *Direct* **3**, e00157 (2019).
889 12. Lu, H.-C. *et al. R2R3-MYB* genes coordinate conical cell development and cuticular wax
890 biosynthesis in *Phalaenopsis aphrodite*. *Plant Physiol* **188**, 318–331 (2022).
891 13. Chuang, Y.-C. *et al.* A dual repeat cis-element determines expression of *GERANYL*
892 *DIPHOSPHATE SYNTHASE* for monoterpene production in *Phalaenopsis* orchids. *Front*
893 *Plant Sci* **9**, 765 (2018).
894 14. Cai, J. *et al.* The genome sequence of the orchid *Phalaenopsis equestris*. *Nat Genet* **47**,
895 65–72 (2015).
896 15. Chao, Y. *et al.* Chromosome-level assembly, genetic and physical mapping of
897 *Phalaenopsis aphrodite* genome provides new insights into species adaptation and
898 resources for orchid breeding. *Plant Biotechnol J* **16**, 2027–2041 (2018).
899 16. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome
900 assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat*
901 *Commun* **14**, 6556 (2023).
902 17. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality,
903 completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245
904 (2020).
905 18. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and
906 annotation completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
907 19. Bashton, M. & Chothia, C. The generation of new protein functions by the combination
908 of domains. *Structure* **15**, 85–99 (2007).

909 20. Vishwanath, S., de Brevern, A. G. & Srinivasan, N. Same but not alike: Structure,
910 flexibility and energetics of domains in multi-domain proteins are influenced by the
911 presence of other domains. *PLoS Comput Biol* **14**, e1006008 (2018).
912 21. Liu, Y. *et al. Arabidopsis FHY3* and *FAR1* regulate the balance between growth and
913 defense responses under shade conditions. *Plant Cell* **31**, 2089–2106 (2019).
914 22. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of *de novo* genes in
915 *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
916 23. Dion-Côté, A.-M. Spermatogenesis: A hotspot for new genes. *eLife* **8**, e50136 (2019).
917 24. Yang, H. *et al.* Global modulation of gene expression and transcriptome size in aneuploid
918 combinations of maize. *Proc Natl Acad Sci U S A* **122**, e2426749122 (2025).
919 25. Pavelka, N. *et al.* Aneuploidy confers quantitative proteome changes and phenotypic
920 variation in budding yeast. *Nature* **468**, 321–325 (2010).
921 26. Selmecki, A., Forche, A. & Berman, J. Aneuploidy and isochromosome formation in
922 drug-resistant *Candida albicans*. *Science* **313**, 367–370 (2006).
923 27. Hsu, H.-F. *et al.* Multifunctional evolution of B and *AGL6* MADS box genes in orchids.
924 *Nat Commun* **12**, 902 (2021).
925 28. Wang, F. *et al.* Molecular mechanisms underlying floral trait formation in *Phalaenopsis*
926 orchids. *Hortic Res* uhaf340 (2025) doi:10.1093/hr/uhaf340.
927 29. Chen, W.-H., Jiang, Z.-Y., Hsu, H.-F. & Yang, C.-H. Silencing of *FOREVER YOUNG*
928 *FLOWER* -like genes from *Phalaenopsis* orchids promotes flower senescence and
929 abscission. *Plant Cell Physiol* **62**, 111–124 (2021).
930 30. Wang, R., Mao, C. & Ming, F. PeMYB4L interacts with PeMYC4 to regulate
931 anthocyanin biosynthesis in *Phalaenopsis* orchid. *Plant Sci* **324**, 111423 (2022).
932 31. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of
933 occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

934 32. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and
935 Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432
936 (2020).
937 33. Liu, B. *et al.* Estimation of genomic characteristics by analyzing k-mer frequency in de
938 novo genome projects. Preprint at https://doi.org/10.48550/ARXIV.1308.2012 (2013).
939 34. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo
940 assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
941 35. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration
942 of BWA-MEM for multicore systems. in *2019 IEEE International Parallel and*
943 *Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, Rio de Janeiro, Brazil,
944 2019). doi:10.1109/IPDPS.2019.00041.
945 36. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using
946 Hi-C data without reference genomes. *Nat Plants* **10**, 1184–1200 (2024).
947 37. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with
948 unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
949 38. Jia, K. *et al.* SubPhase: a robust allopolyploid subgenome phasing method based on
950 subgenome-specific *k*-mers. *New Phytol* **235**, 801–809 (2022).
951 39. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
952 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
953 40. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software
954 for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
955 41. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling
956 rapid identification of long terminal repeat retrotransposons. *Mobile DNA* **10**, 48 (2019).
957 42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR
958 Assembly Index (LAI). *Nucleic Acids Res* **46**, e126 (2018).

959 43. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for
960 identification of long terminal repeat retrotransposons. *Plant Physiol* **176**, 1410–1422
961 (2018).
962 44. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and
963 centromeric repeat identification. *Hortic Res* **10**, uhad127 (2023).
964 45. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection
965 and functional classification of transfer RNA genes. *Mol Biol Evol* **49**, 9077–9096 (2021).
966 46. Nawrocki, E. P. Annotating functional RNAs in genomes using Infernal. *Methods Mol*
967 *Biol* **1097**, 163–197 (2014).
968 47. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a
969 streamlined, comprehensive pipeline. *Genome Biol* **20**, 275 (2019).
970 48. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo
971 classification of transposons with convolutional neural network. *Bioinformatics* **36**,
972 4269–4275 (2020).
973 49. Zhang, W. *et al.* Genome sequence of *Apostasia ramifera* provides insights into the
974 adaptive evolution in orchids. *BMC Genomics* **22**, 536 (2021).
975 50. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
976 *Res* **27**, 573–580 (1999).
977 51. Smit, A. & Hubley, R. RepeatModeler Open-1.0. 2008-2015
978 <http://www.repeatmasker.org>.
979 52. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and
980 deduplication using fastp. *iMeta* **2**, e107 (2023).
981 53. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction
982 utilizing intron position conservation and RNA-seq data. *Methods Mol Biol* **1962**, 161–
983 177 (2019).

984 54. Zhang, G.-Q. *et al.* The *Apostasia* genome and the evolution of orchids. *Nature* **549**, 379–
985 383 (2017).
986 55. Zhang, G.-Q. *et al.* The *Dendrobium catenatum* Lindl. genome sequence provides
987 insights into polysaccharide synthase, floral development and adaptive evolution. *Sci Rep*
988 **6**, 19029 (2016).
989 56. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
990 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–
991 915 (2019).
992 57. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
993 RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).
994 58. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a
995 genome from RNA-Seq data. *Nat Biotechnol* **29**, 644–652 (2011).
996 59. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript
997 alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
998 60. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and
999 protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res* **34**,
1000 769–777 (2024).
1001 61. Blanco, E., Parra, G. & Guigó, R. Using geneid to identify genes. *Curr Protoc*
1002 *Bioinformatics* **18**, (2007).
1003 62. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open
1004 source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
1005 63. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
1006 EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7
1007 (2008).
1008 64. Haas, B. J., Zeng, Q., Pearson, M. D., Cuomo, C. A. & Wortman, J. R. Approaches to

1009 fungal genome annotation. *Mycology* **2**, 118–141 (2011).
1010 65. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
1011 database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).
1012 66. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
1013 *Bioinformatics* **30**, 1236–1240 (2014).
1014 67. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with
1015 RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
1016 68. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
1017 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
1018 *Res* **47**, D309–D314 (2019).
1019 69. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J.
1020 eggNOG-mapper v2: functional annotation, orthology assignments, and domain
1021 prediction at the metagenomic scale. *Mol Biol Evol* **38**, 5825–5829 (2021).
1022 70. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1023 (2009).
1024 71. Xu, D. *et al.* GFAP: ultrafast and accurate gene functional annotation software for plants.
1025 *Plant Physiol* **193**, 1745–1748 (2023).
1026 72. Schläpfer, P. *et al.* Genome-wide prediction of metabolic enzymes, pathways, and gene
1027 clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
1028 73. Jin, J. *et al.* PlantTFDB 4.0: toward a central hub for transcription factors and regulatory
1029 interactions in plants. *Nucleic Acids Res* **45**, D1040–D1045 (2017).
1030 74. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic
1031 rearrangements and local sequence differences from whole-genome assemblies. *Genome*
1032 *Biol* **20**, 277 (2019).
1033 75. Wang, Y. *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene

1034 synteny and collinearity. *Nucleic Acids Res* **40**, e49–e49 (2012).
1035 76. Putri, G. H., Anders, S., Pyl, P. T., Pimanda, J. E. & Zanini, F. Analysing high-
1036 throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics* **38**, 2943–2945
1037 (2022).
1038 77. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing,
1039 quality control, normalization and visualization of single-cell RNA-seq data in R.
1040 *Bioinformatics* **33**, 1179–1186 (2017).
1041 78. R Core Team. R: a language and environment for statistical computing. R foundation for
1042 statistical computing, Vienna. (2023).
1043 79. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
1044 analysis. *BMC Bioinformatics* **9**, 559 (2008).
1045 80. Wei, Y. *et al.* OrchidMD: an integrated and user-interactive orchid multi-omics database
1046 for mining genes and biological research. *Plant Biotechnol J* pbi.70445 (2025)
1047 doi:10.1111/pbi.70445.
1048 81. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis
1049 version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870–1874 (2016).
1050 82. Darriba, D. *et al.* ModelTest-NG: a new and scalable tool for the selection of DNA and
1051 protein evolutionary models. *Mol Biol Evol* **37**, 291–294 (2020).
1052 83. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast,
1053 scalable and user-friendly tool for maximum likelihood phylogenetic inference.
1054 *Bioinformatics* **35**, 4453–4455 (2019).
1055 84. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1056 Preprint at https://doi.org/10.48550/ARXIV.1303.3997 (2013).
1057 85. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008
1058 (2021).

1059 86. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of
1060 samples. Preprint at https://doi.org/10.1101/201178 (2017).
1061 87. Pedersen, B. S. smoove: structural-variant calling and genotyping with existing tools.
1062 (2020).
1063 88. Suvakov, M., Panda, A., Diesh, C., Holmes, I. & Abyzov, A. CNVpytor: A tool for copy
1064 number variation detection and analysis from read depth and allele imbalance in whole-
1065 genome sequencing. *GigaScience* **10**, giab074 (2021).
1066 89. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic
1067 variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
1068 90. Rosyara, U. R., Jong, W. S. D., Douches, D. S. & Endelman, J. B. Software for genome-
1069 wide association studies in autopolyploids and its application to potato. *Plant Genome* **9**,
1070 plantgenome2015.8.73 (2016).
1071 91. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and
1072 pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100–106
1073 (2014).
1074 92. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**,
1075 2825–2830 (2011).
1076 93. Yin, L. CMplot: Circle Manhattan Plot. (2024).