Science Bulletin 62 (2017) 1304–1314

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Article

Single-cell transcriptome analysis reveals widespread monoallelic gene expression in individual rice mesophyll cells

Yingying Han [a,c,1,2], Xiao Chu [a,b,c,2], Haopeng Yu [a,c], Ying-Ke Ma [a,b,c,3], Xiu-Jie Wang [a,b,c], Wenfeng Qian [a,b,c,*], Yuling Jiao [a,c,*]

[a] State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[b] Key Laboratory of Genetic Network Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 10 August 2017
Received in revised form 13 September 2017
Accepted 14 September 2017
Available online 22 September 2017

Keywords:
Monoallelic gene expression
Single-cell RNA-seq
Rice mesophyll cells
Independent allelic expression

ABSTRACT

Monoallelic gene expression refers to the phenomenon that all transcripts of a gene in a cell are expressed from only one of the two alleles in a diploid organism. Although monoallelic gene expression has been occasionally reported with bulk transcriptome analysis in plants, how prevalent it is in individual plant cells remains unknown. Here, we developed a single-cell RNA-seq protocol in rice and investigated allelic expression patterns in mesophyll cells of *indica* (93-11) and *japonica* (Nipponbare) inbred lines, as well as their F1 reciprocal hybrids. We observed pervasive monoallelic gene expression in individual mesophyll cells, which could be largely explained by stochastic and independent transcription of two alleles. By contrast, two mechanisms that were proposed previously based on bulk transcriptome analyses, parent-of-origin effects and allelic repression, were not well supported by our data. Furthermore, monoallelically expressed genes exhibited a number of characteristics, such as lower expression levels, narrower H3K4me3/H3K9ac/H3K27me3 peaks, and larger expression divergences between 93-11 and Nipponbare. Taken together, the development of a single-cell RNA-seq protocol in this study offers us an excellent opportunity to investigate the origins and prevalence of monoallelic gene expression in plant cells.

1. Introduction

Although a gene has two copies of DNA in a diploid cell, sometimes all mRNA molecules present in a cell are expressed from only one of them, a phenomenon often referred to as monoallelic gene expression [1–3]. Monoallelic expression in plants has been investigated with allele-specific RT-PCR, microarray, or bulk RNA-seq [4–14], and has been suggested to increase the phenotypic diversity [4,8,15,16]. Based on the observations in these bulk transcriptome analyses, two possible molecular mechanisms, parent-of-origin effects (only maternal or paternal allele is expressed in a cell) and allelic repression (the expression of one allele can repress that of the other), have been proposed to cause monoallelic gene expression [1].

It is of importance to note that monoallelic gene expression in individual cells may not always be observed in bulk transcriptome analyses, especially when the allele expressed in individual cells is random and dynamic [2,3]. In individual plant cells, monoallelic gene expression can be a potential regulatory mechanism or a genetic constraint in development, but the prevalence of it remains largely unknown. Recent advancement of RNA detection techniques has enabled studies of single-cell transcriptome in animals [17–22]. However, such approach has not been applied to plant cells due to technical hurdles.

Furthermore, we do not yet understand how monoallelic gene expression is established in individual plant cells. In diploid organisms, the two copies of DNA may exhibit substantial difference in expression level due to the stochasticity in gene expression [16,23–26]. Alternatively, parent-of-origin effects and allelic repression may also play a role in plant cells [1]. It remains interesting and critical to scrutinize the relative importance of these molecular mechanisms in individual plant cells.

* Corresponding authors.
E-mail addresses: wfqian@genetics.ac.cn (W. Qian), yljiao@genetics.ac.cn (Y. Jiao).

[1] Present address: Department of Biological Chemistry, School of Medicine, University of California, Irvine, CA 92697, USA.

[2] These authors contributed equally to this work.

[3] Present address: Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.

https://doi.org/10.1016/j.scib.2017.09.011
2095-9273/

Here, we developed a single-cell RNA-seq protocol in rice and applied this approach in mesophyll cells to address these questions. Mesophyll cell is a representative type of leaf cells responsible for photosynthesis, without further differentiation or endoreduplication in rice [27]. Furthermore, it is morphologically distinguishable from other types of leaf cells, such as bundle sheath cells. Whereas many dicots have distinct palisade mesophyll and spongy mesophyll, rice mesophyll cells are homogenous [28]. In addition, high quality reference genomes are available for representative varieties from both cultivated *indica* (93-11) and *japonica* (Nipponbare, NPB hereafter) subspecies [29,30]. More importantly, whereas their reciprocal hybrids (93-11 × NPB and NPB × 93-11) are viable, the evolutionarily divergence between 93-11 and NPB is large enough (0.45% for one-to-one orthologs at the DNA level) to distinguish the expression of two alleles in the hybrids. All make rice mesophyll cell an excellent system to study the origins and prevalence of monoallelic gene expression in individual plant cells.

2. Materials and methods

2.1. Plant materials, growth conditions, and tissue collection

Rice cultivar Nipponbare (NPB, *Oryza sativa* ssp *japonica*) and 93-11 (*O. sativa* ssp *indica*) and their F1 hybrids (93-11 × NPB and NPB × 93-11) were used. Plants were grown on 1/2 Murashige Skoog (MS) medium under long-day conditions (16 h light and 8 h dark at 26 °C).

2.2. Isolation of single rice cells

Rice protoplasts of mesophyll cells were isolated as previously described [31]. Briefly, the middle part of the second leaf from a 5 days-after-germination (DAG) rice seedling [32] was cut into 0.5 mm strips using a scalpel. The leaf strips were submerged into enzyme solution (0.6 mol/L Mannitol, 1 mmol/L $CaCl_2$, 0.1% BSA, 10 mmol/L MES, pH 5.7, 1.5% Cellulase RS (Yakult), 0.75% Macerozyme R-10 (Yakult), 5 mmol/L β-mercaptoethanol) for protoplasting. After a 3 h incubation, mesophyll cells were released by gently pipetting using a mouth pipette. Single mesophyll cells were identified under a dissection microscope based on morphology. We selected single cells of similar size. Because the ploidy of a cell often correlates with its cell size [33], by doing this, we could exclude the potential and occasional events of endoreduplication in individual mesophyll cells. After washing three times with PBS-BSA, cells were transferred within a volume of <0.1 μL of PBS-BSA to PCR tubes containing the lysis buffer composed of 1× PCR buffer II (Thermo Fisher), 1.35 mmol/L $MgCl_2$, 0.45% NP-40 (Roche), 4.5 mmol/L DTT, 0.045 mmol/L dNTP, 0.18 U/μL SUPERase-In (Thermo Fisher), 0.36 U/μL RNase Inhibitor (Thermo Fisher), 12.5 nmol/L UP1 Primer (Table S1 online).

2.3. RNA-seq library preparation and sequencing

Single cells and pool-and-split cDNA libraries were constructed as published previously [34]. In brief, individual cells were seeded into lysate buffer by mouth pipette, and reverse transcription reacted directly on the whole cell lysate. We then applied exonuclease I (New England Biolabs) to remove free primers. Next, a poly(A) tail was added to the 3′ end of the first-strand cDNAs using terminal deoxynucleotidyl transferase (Thermo Fisher). Single-cell cDNAs were then amplified by 33 cycles of PCR. The resulting 100–200 ng of amplified cDNAs were used to construct a sequencing library. Final cDNA libraries (200–2,000 ng depending on the amount of input material) were checked for the expression of *OsRBC* (*Os12g17600*, primers see Table S1) using qPCR. After passing quality control, libraries were fragmented with a Covaris S2 system. After fragmentation, we constructed cDNA libraries using the NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's instructions. Samples were sequenced using Illumina HiSeq 2000 to obtain >20 million single-end 100-bp reads per sample.

2.4. RNA-FISH

Rice protoplasts were isolated, washed with PBS buffer (137 mmol/L NaCl, 2.7 mmol/L KCl, 10 mmol/L Na_2HPO_4, 1.8 mmol/L KH_2PO_4, pH 7.4) for three times, and transferred onto poly-lysine treated glass slides by a mouth pipette. Cells were fixed with PBS solution containing 3.6% PFA and 10% acetic acid for 30 min on ice. Cells were then permeabilized with 0.5% Triton X-100 in PBS for 5–10 min, washed with 1× PBS for 2 min for three times, and rinsed at 4 °C in 70% ethanol for 5 min for two times, dehydrated after an 80%, 90%, and 100% ethanol gradient, and air dried. FISH was then performed with digoxin-labeled probes (see Table S1 for sequences) at a concentration of 4 μg/mL. Cells were incubated with probes at 37 °C overnight. We then used 2× SSC (60 mmol/L $Na_3C_6H_5O_7$, 200 mmol/L NaCl, pH 7.0) with 50% formamide to wash cells for 5 min at 42 °C for three times, washed cells with 2× SSC for 5 min at 65 °C for three times, washed with 2× SSC for an additional 10 min at room temperature, and finally washed with 1× PBST for 5 min for four times. We then put slides with cells into PBST (containing 5% BSA) to block for at least 1 h, and added sheep anti-DIG (1:1,000, Roche), and incubated for 2 h. We next used PBST to wash cells for 5 min for two times, incubated with Alexa Fluor 488-conjugated AffiniPure donkey anti-sheep IgG (H+L) (Jackson ImmunoResearch) for 1 h in the dark, and washed with PBST for 5 min for two times before a 2 min stain with propidium iodide (PI, 25 μg/mL). We finally washed cells with PBST for 5 min for two times, and mounted slides with Prolong Gold antifade (Thermo Fisher). Confocal microscopy images were taken with a Nikon A1 confocal microscope.

2.5. Read alignment and gene expression quantification

We first removed primer and adaptor contaminations, low quality reads (Q20 < 70), and reads corresponding to rRNA or tRNA annotated in Rfam database release 12.0 [34]. Then, we mapped the remaining reads to the *O. sativa* spp. *japonica* (cv. Nipponbare) version 7.1 (MSU v7.1) reference genome using STAR2 version 2.4.1d [35], allowing for up to 3 mismatch and 20 alignment hits. The expression counts of gene locus were calculated from uniquely aligned reads by HTseq-count version 0.6.0 [36], and then normalized to RPKM (Reads Per Kilobase per Million mapped reads) by edgeR version 3.10.3 [37] with the Trimmed Mean of M-values (TMM).

2.6. SNP identification

Reads from two parental lines (93-11 and NPB) were used to identify SNPs. Specifically, the standard RNA-seq variant analysis workflow imbedded in GATK version 3.3 [38] was performed on 93-11 and NPB respectively. Five criteria were used to filter unreliable SNPs: (i) we replaced candidate SNPs in the reference genome, performed alignment again, and only kept the SNPs with the recalculated quality scores greater than 20; (ii) the SNP was supported by 90% of the SNP-covering reads in a parental line; (iii) the quality by depth (QD) of the SNP should be greater than 2.0 [39]; (iv) the strand bias score (Fisher strand, FS) should be less than 30; (v) if at least 3 SNPs exist within a 35-base window in the genome, all these SNPs were discarded. After that, we removed the SNPs identified in both parental lines.

2.7. Classification of biallelic, monoallelic and silenced genes in a cell

Uniquely aligned SNP-covering reads were classified into 93-11 origin or NPB origin in each cell. We counted the number of reads from both alleles and calculated the average read count per SNP for each gene. Genes with the average read count >10 were considered to be expressed which could further be classified as monoallelic and biallelic ones. Following a previous study in animals [25], two criteria were used in the identification of monoallelic expression genes: (i) at least 98% of the reads were expressed from one allele and (ii) false discovery rate (FDR, or *Q* value) <0.001. *P* values were estimated with *G*-test, against the null hypothesis that a sequencing read is equally likely from both alleles. Note that only one sequencing read was counted if multiple ones exhibiting an identical starting site and the same allelic origin, to correct for the potentially biased amplification during library preparation [40]. *Q* values were estimated with an *R* package "qvalue", which corrected for multiple comparisons.

2.8. Estimation of the breadths of histone modification peaks

H3K4me3, H3K9ac and H3K27me3 peak locations in 93-11, NPB, 93-11 × NPB and NPB × 93-11 were retrieved from a previous study [7]. For each gene, the total peak breadths were calculated within the range between the upstream 1000 base pairs to the transcriptional start site and the transcriptional termination site of the gene.

2.9. Estimation of the expression divergence between NPB and 93-11

For each gene, the average expression levels (in the unit of RPKM) were calculated in 93-11 and NPB. The expression divergence was defined as the log-transformed ratio between them.

$$\text{Expression divergence} = \left|\log_2\left(\frac{\frac{1}{8}\sum_{i=1}^{8}(expr_{93\text{-}11,i})}{\frac{1}{8}\sum_{i=1}^{8}(expr_{NPB,i})}\right)\right|,$$

where $expr_{93\text{-}11,i}$ ($expr_{NPB,i}$) represents the expression level of a gene in the *i*th 93-11 (NPB) cell.

2.10. Evaluation of the independence of allelic expression

We used two parameters, $k_{93\text{-}11}$ and k_{NPB}, to describe the probabilities an allele is expressed in single cells, which was estimated from the proportion of cells that express this allele.

$$k_{NPB} = \frac{N_{NPB\,monoallelic\,cells} + N_{biallelic\,cells}}{8}$$

$$k_{93\text{-}11} = \frac{N_{93\text{-}11\,monoallelic\,cells} + N_{biallelic\,cells}}{8}$$

If the expression of two alleles is independent, the proportion of cells exhibiting monoallelic gene expression (mono%) can be predicted by $k_{93\text{-}11} \times (1 - k_{\text{NPB}}) + k_{\text{NPB}} \times (1 - k_{93\text{-}11})$. Note that our formula is different from a previous study on this topic [24], in which the authors assumed the transcription efficacies of the two alleles were equal.

2.11. GO and KEGG enrichment analysis

The gene ontology annotation of *Oryza sativa* was downloaded from Ensembl Plants (http://plants.ensembl.org). GO and KEGG pathway analyses were performed with *R* packages "GOstats" and "KEGGREST", respectively.

2.12. Accession numbers

Raw reads of this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP072415.

3. Results

3.1. Single-cell RNA-seq in rice mesophyll cells

We isolated individual mesophyll cells from 5-days-after-germination seedling leaves of 93-11, NPB, and their reciprocal hybrids (93-11 × NPB and NPB × 93-11, Fig. 1a-b). Different from animal cells, plant cells are surrounded by cell walls that are shared by neighboring cells. We therefore used enzyme digestion to remove cell walls and isolated individual protoplasts. With modification of a single-cell RNA-seq protocol in mammalian cells [34], we extracted mRNA from individual mesophyll protoplasts and acquired their transcriptomes using high-throughput sequencing (Fig. 1b). In total, we gauged single-cell transcriptomes for 32 cells, with at least 13 million clean sequencing reads for each cell (Table S2 online). We performed principal component analysis (PCA) based on the expression levels of 39,045 non-TE-related genes and confirmed that gene expression patterns largely recapitulated genetic background (Fig. 1c).

A closer inspection, however, revealed non-ignorable expression variation among single cells with identical genotype (Fig. 1c). Intriguingly, some genes exhibited substantial expression variation (e.g., *Os06g35700* and *Os02g50690*, Fig. 2a), whereas the expression of other genes were more homogenous (e.g., *Os06g05880*, Fig. 2a). Presumably, the variance in expression level among cells could result from either experimental errors in single-cell RNA-seq [41] or the genuine stochasticity of gene expression [42]. We performed two additional analyses to distinguish these two possibilities. First, we mixed cell lysate from six 93-11 mesophyll cells and split it evenly into six aliquots. Three of them were individually subject to library preparation and high-throughput sequencing (pool-and-split experiment, Fig. S1a). Therefore, the variation among these pool-and-split transcriptomes is attributable to technical error generated during the preparation of sequencing libraries. We observed that the cell-to-cell variability in expression level was much smaller in the pool-and-split experiments albeit the average expression levels were similar (Fig. S1b-d). Second, we performed RNA-fluorescence *in situ* hybridization (RNA-FISH, Table S1), a method that does not rely on mRNA amplification as in single-cell RNA-seq, and thus, is more accurate in quantifying mRNA levels in individual cells [21] (Fig. 2b). We found that *Os06g05880* indeed exhibited smaller expression variation than *Os06g35700* and *Os02g50690*. Therefore, both analyses suggest that the cell-to-cell expression variability observed in our single-cell RNA-seq experiment is unlikely a result of technical errors (Fig. 2b-c).

3.2. Pervasive monoallelic gene expression in rice mesophyll cells

Transcriptomes of individual mesophyll cells isolated from hybrids offer us a unique opportunity to investigate the prevalence of monoallelic gene expression because the allelic gene expression can be distinguished (Fig. 3a). We identified 13,606 single-nucleotide polymorphisms (SNPs) on 5161 genes between 93-11 and NPB, among which 2046 genes contain one SNP and 3115 genes contain at least two SNPs (Fig. S2). For each of the 16 hybrid

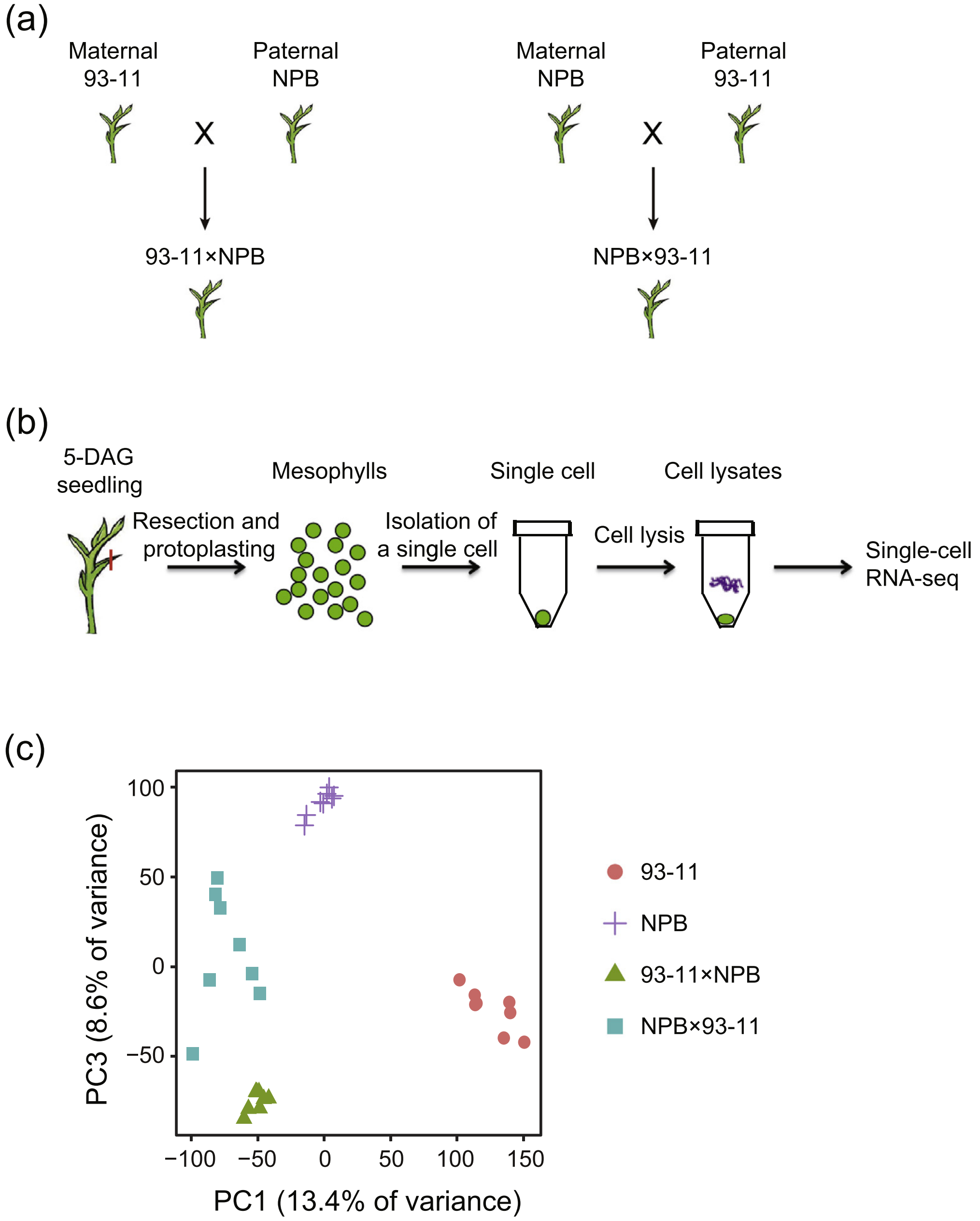


Fig. 1. (Color online) Single-cell transcriptome analysis in rice mesophyll cells. (a) Reciprocal cross of two inbred rice lines (93-11 and NPB). (b) Schematic illustration for acquisition of single-cell RNA-seq profiles. The center section of the second leaf in a 5-DAG rice seedling was isolated and was subject to single-cell RNA-seq experiments. (c) Principal component analysis of 32 single-cell transcriptomes shows expression difference among genotypes and single cells.

cells, we calculated the average number of SNP-covering reads per SNP for each gene (n) and defined a gene as expressed in a cell when n is equal to or larger than 10 (equivalent to ~15% most highly expressed genes in a cell). Following a previous study [25], we used two additional criteria to define monoallelic gene expression in a cell: (1) The fraction of sequencing reads in support of one allele is greater than 98% of all SNP-covering reads on this gene. (2) We used G-test to estimate P values, with the null hypothesis that the gene is equally expressed from both alleles. Sequencing reads mapped to the same start and stop positions could be resulted from PCR amplification during the preparation of high-throughput sequencing libraries. To make our test more rigorous, these reads were counted only once (defined as non-redundant reads). We further calculated false discovery rate (FDR, or Q value) to correct for multiple comparisons and required a Q value smaller than 0.001 to call a gene monoallelically expressed in a cell. Using these criteria, all expressed genes in a cell were classified into three categories: biallelic, 93-11 monoallelic, and NPB monoallelic (Fig. 3).

In cell #1 of the 93-11 × NPB hybrid, 563 (49%), 325 (29%), and 257 (22%) genes were classified as biallelic, 93-11 monoallelic, and NPB monoallelic, respectively (Fig. 3b-c). For example, in total 523 non-redundant reads covering one of the three SNPs in *Os06g40490*, among which 522 were from the NPB allele ($Q < 10^{-100}$, Fig. 3b). We therefore classified this gene as NPB monoallelic. By contrast, in a 93-11 monoallelically expressed gene *Os02g52314*, 416 and 0 non-redundant reads were in support of the expression of the 93-11 allele and the NPB allele, respectively ($Q < 10^{-100}$, Fig. 3b). In a biallelically expressed gene *Os03g18810*, 286 and 316 reads were in support of the 93-11 and NPB alleles, respectively ($Q = 0.66$, Fig. 3b). On average, the expression of 32% ± 3%, 36% ± 1%, 32% ± 2% genes were biallelic, 93-11 allelic,

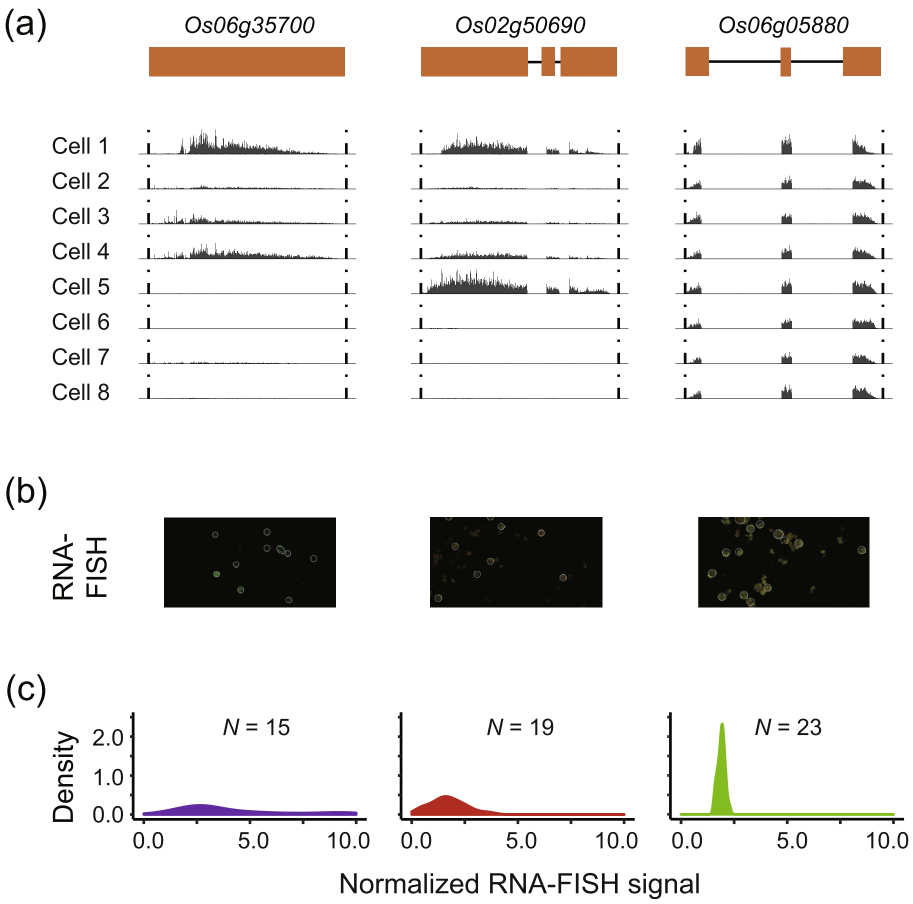


Fig. 2. (Color online) RNA-FISH analysis validated the gene expression variability among single cells. (a) The distribution of the 5′ ends of sequencing reads from single-cell RNA-seq analysis in each 93-11 cell. Exon (orange blocks)-intron (black lines) structures are shown. Dashed lines mark the boundaries of transcripts. (b) Overlaid images of RNA-FISH experiments. False-color overlays of PI straining (red) and RNA-FISH (green) fluorescence micrographs from 93-11 mesophyll cells. (c) The distribution of normalized RNA-FISH signals (the ratio between the intensity of Alexa Fluor 488 and that of PI) among cells. Example cells are circled in (b) and the number of cells (*N*) used to estimate this distribution is shown.

and NPB monoallelic (Fig. 3d, mean ± standard error of the mean) among the 16 hybrid cells, respectively. The average proportion of monoallelic expressed genes was higher in NPB × 93-11 cells than that in 93-11 × NPB cells (60% and 77%, respectively, $P = 0.002$, *t*-test). It remains unclear whether this difference is due to genetic reasons (such as the difference in the mitochondrial genome) or technical reasons (such as the lower efficiency of cell lysis in NPB × 93-11 single-cell samples). Nevertheless, monoallelic gene expression was consistently observed in both reciprocal hybrids.

3.3. Monoallelically expressed genes exhibit lower expression levels, narrower H3K4me3/H3K9ac/H3K27me3 peaks, and larger expression divergences between 93-11 and NPB

We next investigated characteristics associated with allelic expression patterns. To this end, we compared genes that show biallelic expression in at least two of the 16 hybrid cells (B genes) and genes that are not biallelically expressed in any of these 16 cells (M genes, examples are shown in Fig. 4a). M genes exhibited significant lower expression levels in all four genetic backgrounds (Fig. 4b), suggesting that stochasticity in gene expression, which is more prominent among more lowly expressed genes, may contribute to monoallelic gene expression in individual cells (discussed in detail in the next section). Furthermore, M genes exhibited narrower H3K4me3 peaks in rice mesophyll cells (Fig. 4c, an example is shown in Fig. 4a), echoing a recent report that H3K4me3 breadth elevates expression consistency among mammalian cells [43]. In addition, M genes also exhibited narrower peaks of another activating histone marker H3K9ac (Fig. 4a and d). Intriguingly, the peak of H3K27me3, a suppressive histone modification that was reported to be associated with gene exhibiting lower expression divergence between homologous genes in plants [44], was also narrower in M genes (Fig. 4a and e). Concordantly, M genes exhibited larger expression divergence between one-to-one orthologous genes in 93-11 and NPB (Fig. 4a and f, see also Section 4).

3.4. Independent allelic expression partly explains the widespread monoallelic gene expression in rice mesophyll cells

Presumably, monoallelic expression may result from mutual allelic repression and/or from the independent stochastic expression of two alleles [2,3]. To investigate these two mechanisms, we defined $k_{93\text{-}11}$ and k_{NPB} as the expression probability of the 93-11 and NPB alleles (i.e., the fraction of cells with an allele expressed) in a hybrid and tested whether the proportion of cells exhibiting monoallelic gene expression (mono%) can be predicted under the assumption of independent allelic expression. The predicted mono% is equal to $k_{93\text{-}11} \times (1\text{-}k_{\text{NPB}}) + k_{\text{NPB}} \times (1\text{-}k_{93\text{-}11})$, which is the sum of the proportion of 93-11 monoallelic cells ($\text{mono}_{93\text{-}11}\%$, before "+") and that of NPB monoallelic cells ($\text{mono}_{\text{NPB}}\%$, after "+"). For example, in the 93-11 × NPB hybrid, the 93-11 allele of gene *Os02g53040* was expressed in 4 out of 8 cells ($k_{93\text{-}11} = 4/8$, Fig. 5a) and the NPB allele was expressed in 2 cells ($k_{\text{NPB}} = 2/8$, Fig. 5a). Therefore, $4/8 \times (1 - 2/8) + 2/8 \times (1 - 4/8) = 50\%$ cells are

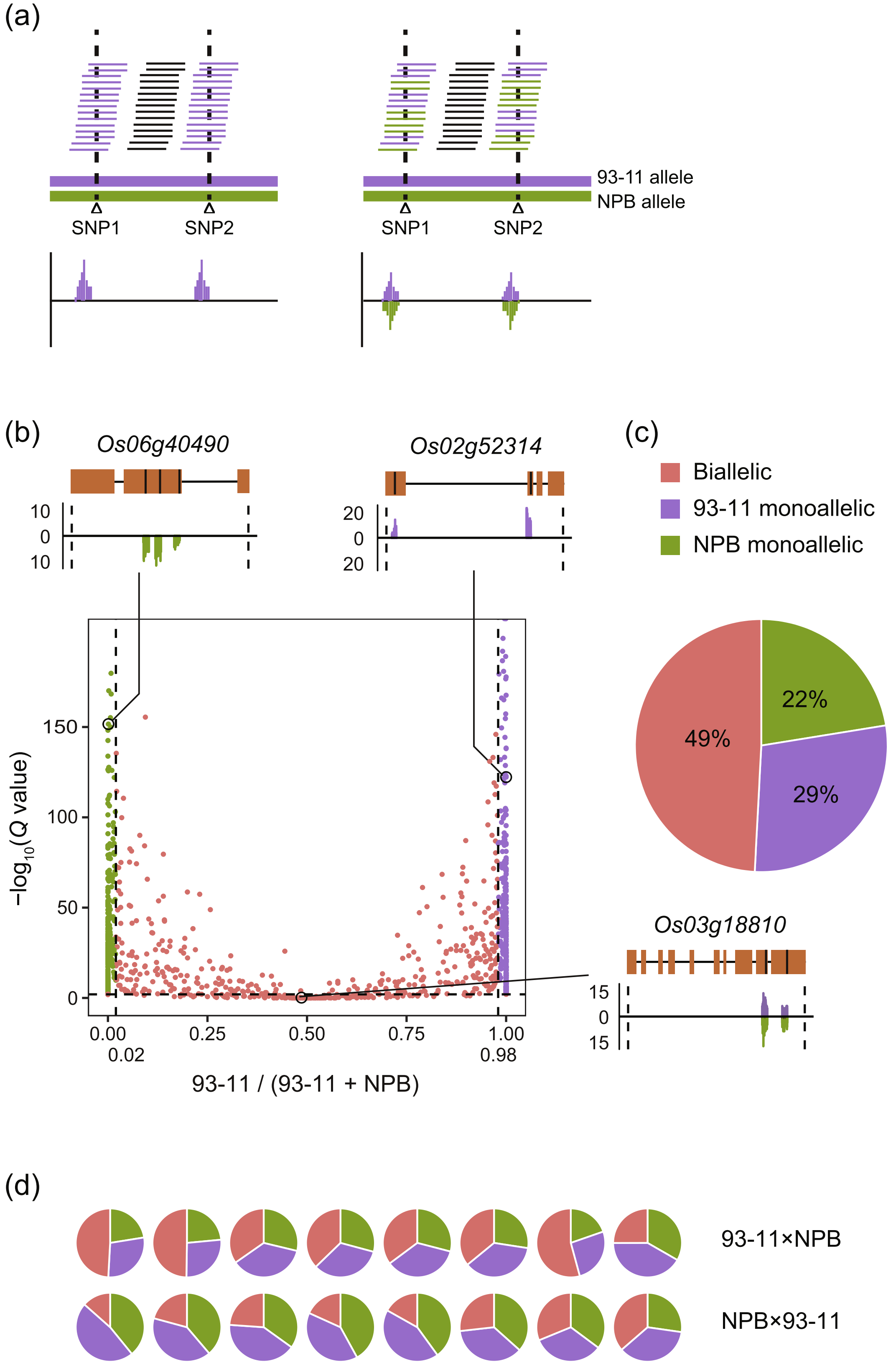


Fig. 3. Allelic expression patterns in single cells. (a) Schematic illustration for the quantification of SNP-covering reads in hybrid cells. Dashed lines mark the positions of SNPs on a genome region in the hybrid. Reads do not cover a SNP are in black. Reads from 93-11 and NPB alleles are shown in purple and green, respectively. Distributions of the 5′ ends of SNP-covering reads are shown. (b) Volcano plot shows the allelic expression patterns in one 93-11 × NPB cell. Expressed genes were classified as monoallelic when two criteria were met: (1) at least 98% SNP-covering reads were expressed from one allele (beyond the vertical dashed lines), and (2) Q value <0.001 (above the horizontal dashed line). Three genes with 93-11 monoallelic (*Os02g52314*), NPB monoallelic (*Os06g40490*) and biallelic (*Os03g18810*) expression are shown as examples. Black vertical lines in gene models stand for the positions of SNPs. (c) A pie chart shows the proportions of 93-11 monoallelic, NPB monoallelic and biallelic genes among expressed genes in the same cell of (b). (d) The proportions of 93-11 monoallelic, NPB monoallelic and biallelic genes in 16 hybrid cells.

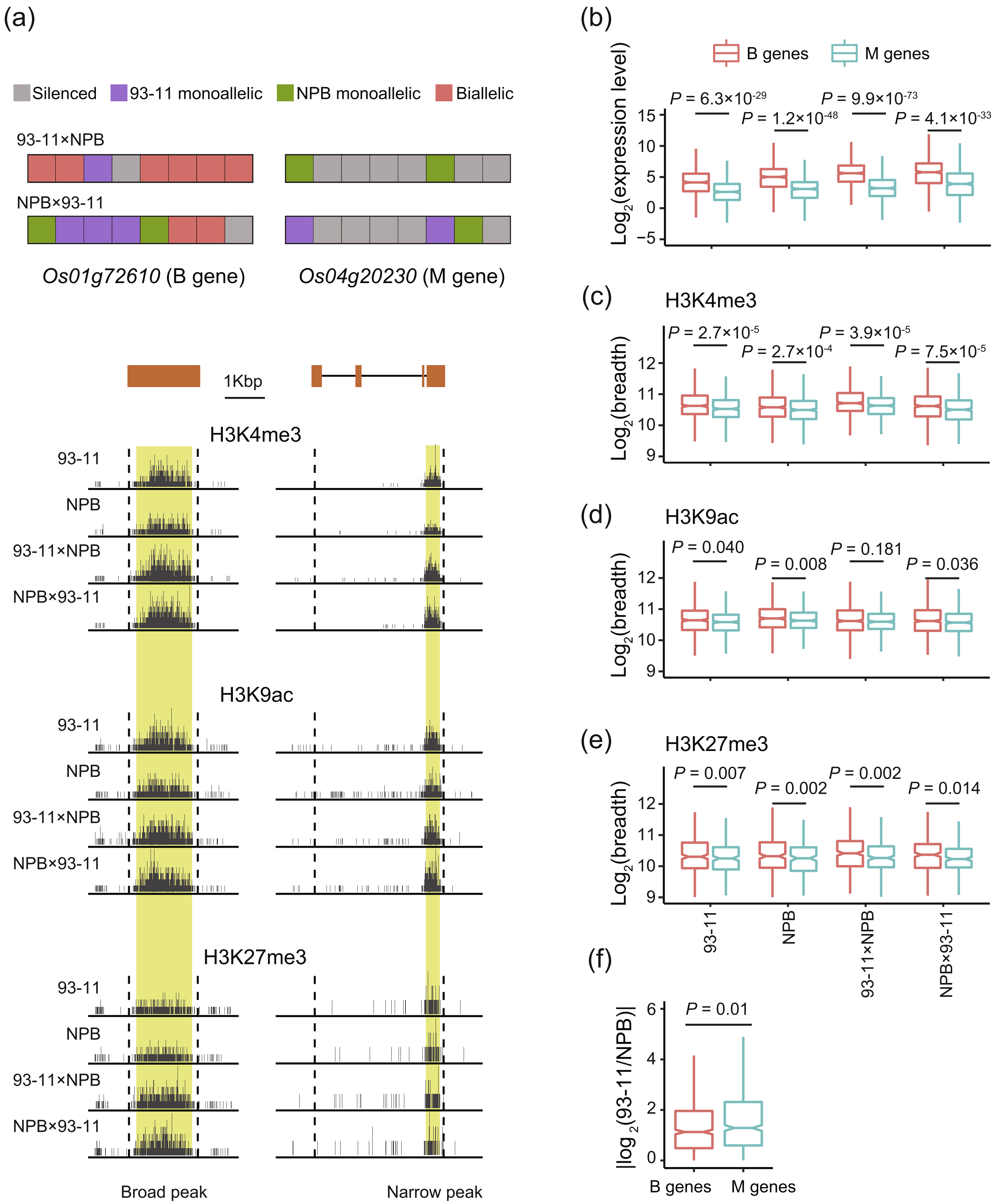


Fig. 4. Properties of monoallelically expressed genes. (a) Top panel: allelic expression patterns of a representative B gene (*Os01g72610*) and a representative M gene (*Os04g20230*). Bottom panel: the distributions of three histone modifications (H3K4me3, H3K9ac, and H3K27me3) in 93-11, NPB, 93-11 × NPB, and NPB × 93-11, respectively. Yellow blocks highlight the histone modification peaks. (b) M genes exhibit significantly lower expression levels than B genes in 93-11, NPB, 93-11 × NPB, and NPB × 93-11, respectively. The average expression level of a gene was calculated from 8 isogenic cells and *P* values were estimated from the Mann-Whitney *U* test. (c–e) M genes exhibit narrower peaks of H3K4me3 (c), H3K9ac (d), and H3K27me3 (e) than B genes in 93-11, NPB, 93-11 × NPB, and NPB × 93-11, respectively. *P* values were estimated from the Mann-Whitney *U* test. (f) M genes exhibit larger expression divergence between 93-11 and NPB. *P* value was estimated from the Mann-Whitney *U* test.

predicted to exhibit monoallelic gene expression of this gene. Indeed, the 93-11 allele of this gene was monoallelically expressed in 3 cells and the NPB allele was monoallelically expressed in one cell. In total, the observed mono% is equal to (4/8=) 50% (Fig. 5a-b). Two additional examples (*Os06g47340* and *Os10g37710*) were shown in Fig. 5a-b.

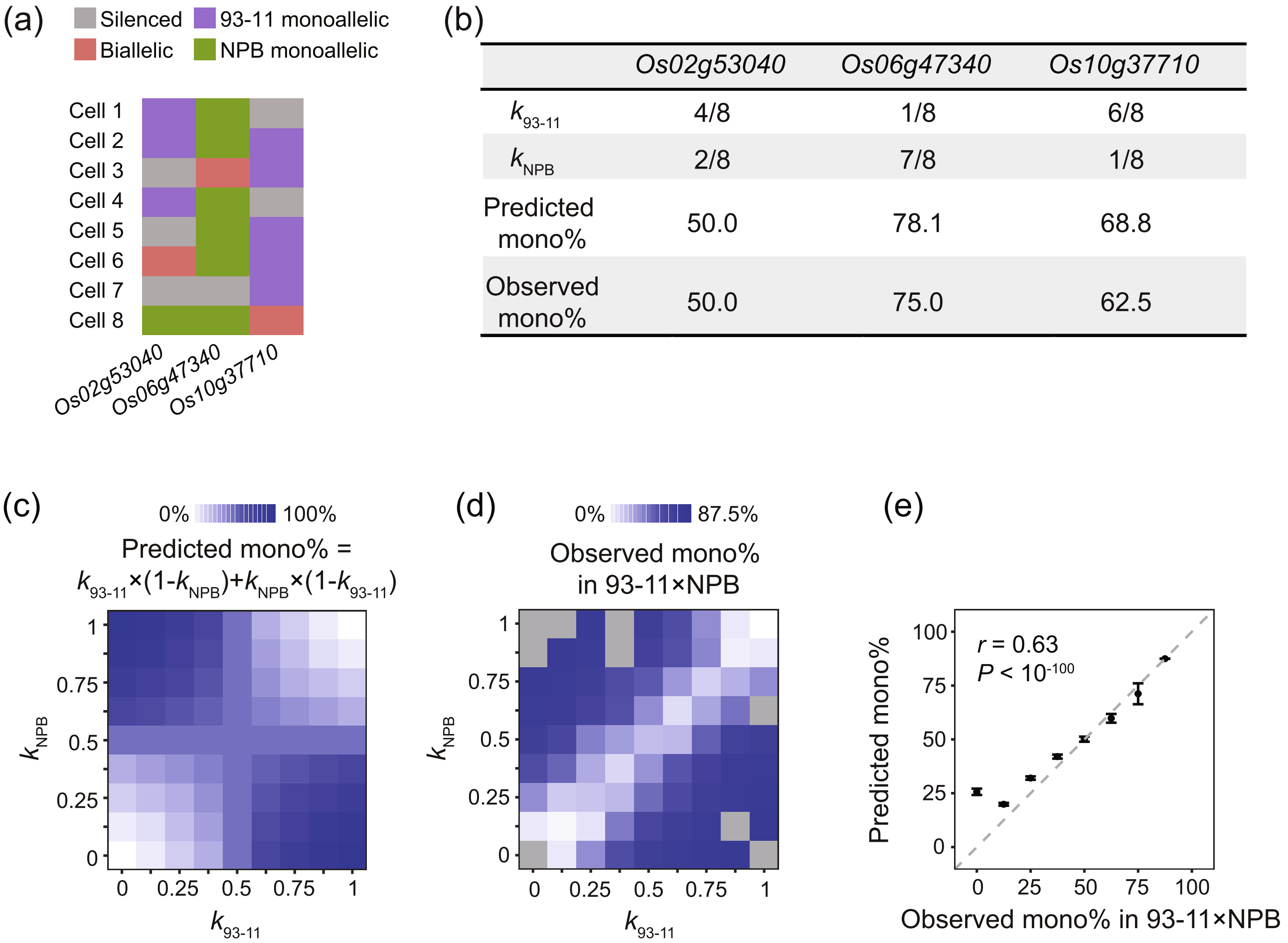
We predicted mono% with a variety of $k_{93\text{-}11}$ and k_{NPB} values under the assumption of independent allelic expression (Fig. 5c). To test the accuracy of this prediction, we calculated observed mono% among eight 93-11 × NPB cells for each gene. We further grouped these genes based on their $k_{93\text{-}11}$ and k_{NPB} values and calculated the average observed mono% among genes within each group. Intriguingly, the observed mono% (Fig. 5d) largely recapitulated the prediction ($r = 0.63$, $P < 10^{-100}$, $N = 2925$, Pearson's correlation, Fig. 5d-e). A similar pattern was also observed in the NPB × 93-11 hybrid ($r = 0.58$, $P < 10^{-100}$, $N = 2251$, Pearson's correlation, Fig. S3a-b), suggesting that independent allelic expression largely explains the widespread monoallelic gene expression in individual rice mesophyll cells. Note that the conclusion remained valid when we dropped cells with the lowest unique-aligned ratio in both reciprocal hybrids (Fig. S3c-d).

3.5. Lack of evidence in support of mutual allelic repression

To further understand whether mutual allelic repression plays an important role in monoallelic gene expression for some genes, we performed permutation among the hybrid cells. For example, in the 93-11 × NPB hybrid, the 93-11 allele of *Os02g53040* is expressed in cells #1, #2, #4 and #6 and the NPB allele is expressed in cells #6 and #8 (Fig. 5a). In a permutation, we randomly assigned 4 cells expressing the 93-11 allele and 2 cells expressing the NPB allele, and then estimated mono% for this gene. We performed the permutation for 1000 times, calculated one-tailed *P* values, and estimated *Q* values to correct for multiple comparisons. We would observe smaller mono% in permutations if mutual allelic repression is one of the mechanisms causing monoallelic gene expression in individual cells. With the cut-off of $Q < 0.05$, we did not identify any gene in support of mutual allelic repression.

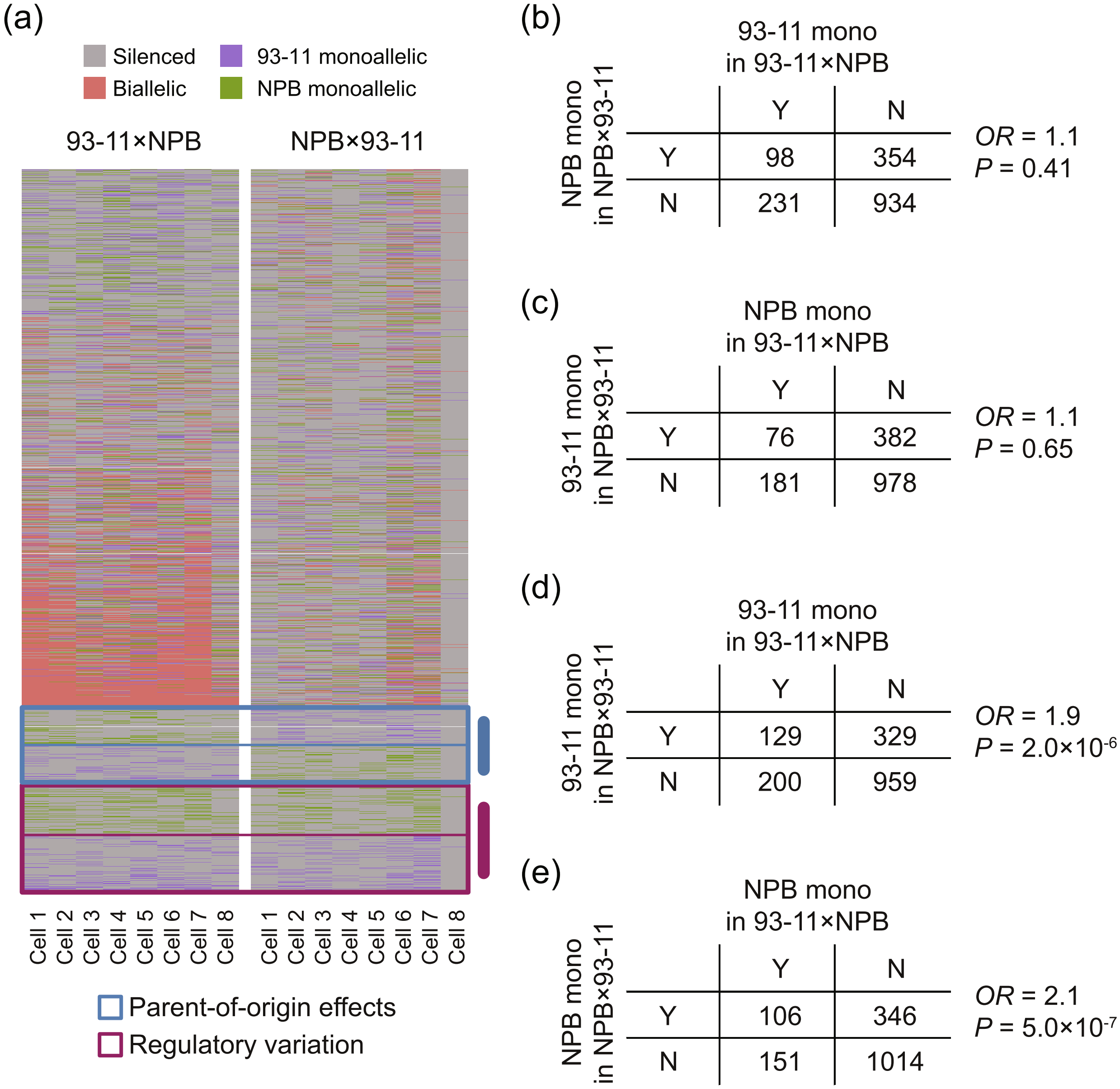
3.6. Lack of evidence in support of parent-of-origin effects

We next examined whether parent-of-origin effects contribute to the monoallelic gene expression observed in rice mesophyll cells. By comparing the expression pattern among the sixteen 93-11 × NPB and NPB × 93-11 cells, we identified 98 (6.1%) genes that only maternal allele was expressed (Fig. 6a, in cyan). A closer inspection, however, revealed that this number was not significantly larger than the random expectation. Specifically, we observed 452 and 329 (28% and 20%) genes, of which only maternal allele was expressed in the eight NPB × 93-11 cells and the eight 93-11 × NPB cells, respectively. Therefore, we expected that the expression patterns of (28% × 20% =) 5.7% genes were consistent



	Os02g53040	*Os06g47340*	*Os10g37710*
$k_{93\text{-}11}$	4/8	1/8	6/8
k_{NPB}	2/8	7/8	1/8
Predicted mono%	50.0	78.1	68.8
Observed mono%	50.0	75.0	62.5

Fig. 5. Independent and stochastic allelic expression predicts the proportion of cells exhibiting monoallelic expression. (a) Allelic expression patterns of 3 genes across eight 93-11 × NPB cells. (b) Estimation of the expression parameters. See METHODS for details. (c–d) Predicted (c) and observed (d) proportions of cells exhibiting monoallelic expression of a gene. Genes were binned by $k_{93\text{-}11}$ and k_{NPB}, and the average proportion among all genes in a bin is shown. Grey blocks indicate data not available. (e) The proportion of cells exhibiting monoallelic expression could be successfully predicted from $k_{93\text{-}11}$ and k_{NPB}, under the assumption of independent allelic expression. Pearson's correlation coefficient *r* and *P* value were calculated from the raw data ($N = 2925$). Error bars represent standard errors.



(b) 93-11 mono in 93-11×NPB; NPB mono in NPB×93-11

	Y	N
Y	98	354
N	231	934

OR = 1.1
P = 0.41

(c) NPB mono in 93-11×NPB; 93-11 mono in NPB×93-11

	Y	N
Y	76	382
N	181	978

OR = 1.1
P = 0.65

(d) 93-11 mono in 93-11×NPB; 93-11 mono in NPB×93-11

	Y	N
Y	129	329
N	200	959

OR = 1.9
$P = 2.0\times10^{-6}$

(e) NPB mono in 93-11×NPB; NPB mono in NPB×93-11

	Y	N
Y	106	346
N	151	1014

OR = 2.1
$P = 5.0\times10^{-7}$

Fig. 6. Parent-of-origin effects and regulatory variation. (a) Overview of the allelic expression patterns among 16 hybrid cells. Each row represents a gene and each column represents a cell. The genes exhibiting parent-of-origin effects and regulatory variation are shown within cyan and purple boxes, respectively. (b–e) Statistical analyses on the numbers of genes exhibiting parent-of-origin effects (b–c) and regulatory variation (d–e). A gene was classified into one of the 4 classes in the 2 × 2 contingency tables and the number of genes in each class was labeled. "93-11 mono" ("NPB mono") represents genes that the NPB (93-11) allele is not expressed in any of the eight cells. Odds ratios (*OR*) and *P* values were given by Fisher's exact test.

with parent-of-origin effects, which was not significantly different from 6.1% (Fig. 6b, odds ratio = 1.1, *P* = 0.41, Fisher's exact test). Similarly, we did not observe significantly more genes, of which only paternal allele was expressed (Fig. 6c, odds ratio = 1.1, *P* = 0.65, Fisher's exact test). These observations suggest that parent-of-origin effects are unlikely a major cause leading to the widespread monoallelic gene expression in individual rice mesophyll cells, echoing some previous observations in bulk-based analyses [4–8,10].

4. Discussion

We studied monoallelic gene expression in cells from reciprocal hybrids, where allelic expression can be detected. However, in addition to the mechanisms mentioned above, which are applied to both the inbred lines and their hybrids, the phenomenon of monoallelic gene expression in the hybrid cells may be partly caused by the regulatory variation between 93-11 and NPB as well. This mechanism predicts that the same allele (93-11 or NPB) is monoallelically expressed in both reciprocal hybrids (Fig. 6a, in purple). Indeed, we identified 129 such alleles in 93-11 and 106 such alleles in NPB. These numbers were significantly larger than the random expectation (Fig. 6d-e, odds ratio = 1.9 and 2.0, the difference between observed and expected gene numbers = 36 and 34, $P = 2\times10^{-6}$ and 5×10^{-7}, respectively, Fisher's exact test). Therefore, regulatory variation partly contributes to the monoallelic gene expression observed in hybrid mesophyll cells. One of the possibilities is that one allele contains a stronger promoter, consistent with the phenomenon that M genes exhibited larger expression divergence between 93-11 and NPB (Fig. 4f). We expect this regulatory variation being smaller in inbred lines, and therefore predict to observe less monoallelically expressed genes in

93-11 or NPB, if we were able to detect monoallelic gene expression experimentally in these lines. Nevertheless, the number of genes that can be explained by regulatory variation is small (36 + 34 = 70), and therefore, our major conclusion that monoallelic gene expression is widespread likely persists in inbred lines.

Different from animal cells, plant cells are surrounded by cell walls, making protoplasting a necessary step for single-cell isolation. In addition, mature plant cells, such as mesophyll cells, have a large central vacuole that maintains the cell's turgor. Therefore, cell wall removal during protoplasting makes protoplasts vulnerable to damage due to the high turgor pressure, which is at the range of ∼1 MPa, i.e. ∼10 times of the atmospheric pressure [45]. To overcome these limitations, we used a manual cell isolation approach to ensure the consistency of the mesophyll cell identity/size and the quality of the next generation sequencing. However, this approach limited the number of single cells obtained in this study. Nevertheless, we observed a consistent pattern in all 16 cells from reciprocal hybrids that monoallelic gene expression was pervasive (Fig. 3d), suggesting that 16 single-cell "replicates" is statistically powerful to gauge a robust conclusion that monoallelic gene expression is prevalent. We further examined whether the molecular properties of B and M genes observed in Fig. 4 were sensitive to the number of single cells used in our analysis. To this end, we (i) randomly dropped one or two cells from these 16 cells (Fig. S4a-b), (ii) dropped cells with the lowest unique-aligned ratio in reciprocal hybrids (Fig. S4c), to mimic a decrease of sample size. Our analysis showed that the molecular properties identified in all hybrid cells persisted when down-sampling was performed (Fig. S4). In addition, although mesophyll cells from a mature leaf are fully differentiated and do not likely vary in cell cycle, morphologically indistinguishable and unknown cell differentiation may exist among mesophyll cells. Such cell differentiation, however, is unlikely to alter our conclusion, because the pervasive monoallelic expression was observed in all hybrid cells. More importantly, different from many other plants in which polyploid cells were reported, mesophyll cells in rice do not exhibit endoreduplication [27]. Furthermore, the proportions of monoallelically expressed genes did not decrease with the increase of the cut-off used in defining gene expression (Fig. S5), and therefore, the observed monoallelic gene expression was unlikely caused by technical noise generated during single-cell RNA-seq experiments [46]. Last, in the transcriptomes of individual hybrid cells, only SNP-covering reads are informative for identifying allelic expression. Because they are only a small fraction of sequencing reads, one may wonder if they can accurately reflect gene expression level. Nevertheless, the number of SNP-covering reads and the number of all sequencing reads were highly correlated (Fig. S6), suggesting that SNP-covering reads can faithful reflect mRNA abundance.

We further performed in silico pooling of the transcriptomes of single cells with identical genotype (NPB, 93-11, 93-11 × NPB, or NPB × 93-11) and confirmed that the transcriptomes pooled *in silico* were highly correlated with bulk transcriptomes (ρ = 0.65, 0.65, 0.63 and 0.56, respectively, Spearman's correlation) quantified in a previous study [7]. This suggests that the single-cell RNA-seq data generated in this study largely reflect the actual transcriptome of plant cells in vivo. More importantly, single-cell RNA-seq analysis offers a new approach to identify the fingerprints of expression in individual cells that could never be observed in bulk RNA-seq analysis. In fact, although previous studies with bulk transcriptome analyses reported that the expression of a fraction of genes was biased (Table S3), monoallelically expressed genes were rarely identified. For example, only 3%–4% genes were classified as monoallelic expressed in a previous rice study [8]. By contrast, we identified on average 68% gene that were monoallelically expressed in individual mesophyll cells (Fig. 3). Concordantly, the prevalence of monoallelic gene expression was dramatically reduced when we performed *in silico* pool-and-split (Fig. S7), highlighting the power of single-cell RNA-seq in detecting the transcriptomic behaviors at the single-cell resolution.

The widespread monoallelic expression may lead to heterogeneity among cells if protein sequences encoded by the two alleles in a diploid cell are not identical. This situation is especially astonishing when protein-protein interaction partners are under consideration. If two proteins (A and B) interact with each other and both are monoallelically expressed, there are 4 possible statuses in a cell, $A^{93\text{-}11}B^{93\text{-}11}$, $A^{NPB}B^{93\text{-}11}$, $A^{93\text{-}11}B^{NPB}$, and $A^{NPB}B^{NPB}$. Indeed, the products of two genes (*Os09g19700* and *Os01g22900*) can physically interact with each other [47] and all 4 possible statuses were observed in the single-cell transcriptomes (Fig. S8). When we consider a protein complex with N ($N > 2$) subunits, the number of possible statuses is even larger (2^N). Because a number of monoallelic genes are related to cytokinesis, cell differentiation, G-protein coupled receptor signaling pathway, trichome morphogenesis, circadian rhythm, starch and sucrose metabolism, cellulose metabolic process, and biosynthetic process, among many processes (Table S4), the widespread monoallelic expression may have profound morphological and physiological outcomes. For example, variability in the control of cell division and cell size underlies leaf and sepal epidermal patterning [48]. Stochastic expression of a key regulator, *ATML1*, patterns a field of identical epidermal cells into giant and small cells [49]. Stochastic expression of additional genes are expected to generate further hidden variations among a uniform cell population, which may regulate developmental patterning and physiological processes.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

YJ and WQ designed the work. YH carried out the experiments. XC, HY, Y-KM and XW performed the analytical and computational analysis. YJ, WQ and XC wrote the paper with inputs from all authors.

Acknowledgments

We thank Fuchou Tang for advice on single-cell RNA-seq, Hang He and Xing-Wang Deng for seeds, and Shaohuan Wu for reading the draft manuscript. This work was supported by the National Basic Research Program of China (2014CB943500), the National Program for Support of Top-Notch Young Professionals, the Recruitment Program of Global Youth Experts, and by the State Key Laboratory of Plant Genomics (SKLPG2011B0103 and SKLPG2016C0219-01).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.scib.2017.09.011.

References

[1] Tarutani Y, Takayama S. Monoallelic gene expression and its mechanisms. Curr Opin Plant Biol 2011;14:608–13.
[2] Reinius B, Sandberg R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. Nat Rev Genet 2015;16:653–64.
[3] Eckersley-Maslin MA, Spector DL. Random monoallelic expression: regulating gene expression one allele at a time. Trends Genet 2014;30:237–44.
[4] Guo M, Rupe MA, Zinselmeier C, et al. Allelic variation of gene expression in maize hybrids. Plant Cell 2004;16:1707–16.

[5] von Korff M, Radovic S, Choumane W, et al. Asymmetric allele-specific expression in relation to developmental variation and drought stress in barley hybrids. Plant J 2009;59:14–26.

[6] Zhang X, Borevitz JO. Global analysis of allele-specific expression in *Arabidopsis thaliana*. Genetics 2009;182:943–54.

[7] He G, Zhu X, Elling AA, et al. Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell 2010;22:17–33.

[8] Song G, Guo Z, Liu Z, et al. Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F1 hybrids. BMC Plant Biol 2013;13:221.

[9] Springer NM, Stupar RM. Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize. Plant Cell 2007;19:2391–402.

[10] Nodine MD, Bartel DP. Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. Nature 2012;482:94–7.

[11] Zhang M, Zhao H, Xie S, et al. Extensive, clustered parental imprinting of protein-coding and noncoding RNAs in developing maize endosperm. Proc Natl Acad Sci U S A 2011;108:20042–7.

[12] Wolff P, Weinhofer I, Seguin J, et al. High-resolution analysis of parent-of-origin allelic expression in the *Arabidopsis* endosperm. PLoS Genet 2011;7:e1002126.

[13] Waters AJ, Makarevitch I, Eichten SR, et al. Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. Plant Cell 2011;23:4221–33.

[14] Luo M, Taylor JM, Spriggs A, et al. A genome-wide survey of imprinted genes in rice seeds reveals imprinting primarily occurs in the endosperm. PLoS Genet 2011;7:e1002125.

[15] Knight JC. Allele-specific gene expression uncovered. Trends Genet 2004;20:113–6.

[16] Gendrel AV, Attia M, Chen CJ, et al. Developmental dynamics and disease potential of random monoallelic gene expression. Dev Cell 2014;28:366–80.

[17] Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods 2011;8:S6–11.

[18] Hashimshony T, Wagner F, Sher N, et al. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep 2012;2:666–73.

[19] Goetz JJ, Trimarchi JM. Transcriptome sequencing of single cells with Smart-Seq. Nat Biotechnol 2012;30:763–5.

[20] Islam S, Kjallquist U, Moliner A, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 2011;21:1160–7.

[21] Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 2013;498:236–40.

[22] Yan L, Yang M, Guo H, et al. Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 2013;20:1131–9.

[23] Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. Science 2004;304:1811–4.

[24] Deng Q, Ramskold D, Reinius B, et al. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 2014;343:193–6.

[25] Reinius B, Mold JE, Ramskold D, et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. Nat Genet 2016;48:1430–5.

[26] Eckersley-Maslin MA, Thybert D, Bergmann JH, et al. Random monoallelic gene expression increases upon embryonic stem cell differentiation. Dev Cell 2014;28:351–65.

[27] Endo M, Nakayama S, Umeda-Hara C, et al. CDKB2 is involved in mitosis and DNA damage response in rice. Plant J 2012;69:967–77.

[28] Sage TL, Sage RF. The functional anatomy of rice leaves: implications for refixation of photorespiratory CO_2 and efforts to engineer C4 photosynthesis into rice. Plant Cell Physiol 2009;50:756–72.

[29] International Rice Genome Sequencing P. The map-based sequence of the rice genome. Nature 2005;436:793–800.

[30] Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 2002;296:79–92.

[31] Shan Q, Wang Y, Li J, et al. Genome editing in rice and wheat using the CRISPR/Cas system. Nat Protoc 2014;9:2395–410.

[32] Kardailsky I, Shukla VK, Ahn JH, et al. Activation tagging of the floral inducer FT. Science 1999;286:1962–5.

[33] Melaragno JE, Mehrotra B, Coleman AW. Relationship between endopolyploidy and cell size in epidermal tissue of *Arabidopsis*. Plant Cell 1993;5:1661–8.

[34] Tang F, Barbacioru C, Nordman E, et al. RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc 2010;5:516–35.

[35] Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

[36] Anders S, Pyl PT, Huber W. HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics 2015;31:166–9.

[37] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–40.

[38] Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;11:11.10:11.10.1–11.10.33.

[39] Achim K, Pettit JB, Saraiva LR, et al. High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nat Biotechnol 2015;33:503–9.

[40] Marx V. How to deduplicate PCR. Nat Methods 2017;14:473–6.

[41] Brennecke P, Anders S, Kim JK, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 2013;10:1093–5.

[42] Marinov GK, Williams BA, McCue K, et al. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res 2014;24:496–510.

[43] Benayoun BA, Pollina EA, Ucar D, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell 2014;158:673–88.

[44] Berke L, Sanchez-Perez GF, Snel B. Contribution of the epigenetic mark H3K27me3 to functional divergence after whole genome duplication in *Arabidopsis*. Genome Biol 2012;13:R94.

[45] Serpe MD, Matthews MA. Growth, pressure, and wall stress in epidermal cells of *Begonia argenteo-guttata* L. leaves during development. Int J Plant Sci 1994;155:291–301.

[46] Kim JK, Kolodziejczyk AA, Ilicic T, et al. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat Commun 2015;6:8687.

[47] Ding X, Richter T, Chen M, et al. A rice kinase-protein interaction map. Plant Physiol 2009;149:1478–92.

[48] Hong L, Dumond M, Tsugawa S, et al. Variable cell growth yields reproducible organ development through spatiotemporal averaging. Dev Cell 2016;38:15–32.

[49] Meyer HM, Teles J, Formosa-Jordan P, et al. Fluctuations of the transcription factor ATML1 generate the pattern of giant cells in the *Arabidopsis* sepal. eLife 2017;6:e19131.