Sector: Transforming Bioenergy & Agriculture with Synthetic Biology

Use Case for Cells for Conversion: Synthetic Sexual Reproduction - Apomixis

End Product: Technologies that could fix genetic complexity in plants/crops to create hybrid vigor

Organism(s) if applicable: Plants/crops

Terrestrial plants do not have the ability to move so their ability to adapt to changing environments (temperature, water, and sunlight) requires precise genetic regulatory mechanisms that control their growth and development, nutrient composition, and yield. Can mixing of these regulatory mechanisms across plants species generate plants and crops that have the potential to generate optimal food, feed, and fiber that would support the agricultural needs of U.S. and the world in the future?

Sexual reproduction mixes genes among parents and provides genetic diversity in subsequent progeny, key for survival and fodder for evolution. Sexual reproduction is detrimental to the propagation of hybrid crops though, as mixing the genes leads to progeny that will be inferior to the hybrid parent. Khanday et al. (see supplemental reading) have developed an asexual-propagation trait that allows a hybrid plant to reproduce clonally, with the progeny also carrying the desirable genome-wide heterogeneity. They accomplish this by silencing genes encoding the BABYBOOM transcription factor (*BBM1-3*) and introducing a *BBM1* gene expressed solely in the egg cell. This leads to embryo development in the absence of fertilization and clonal progeny, although seed production still requires fertilization to generate the endosperm. This work, which builds on previous studies that have shown the capacity of BBM1 to induce somatic embryos from vegetative tissues, shows that seed propagation from hybrid rice varieties can occur without genetic segregation. Apomixis (asexual seed formation) is the process by which a plant bypasses meiosis and fertilization, resulting a plant that develops as a maternal clone. Many flowering plants have shown this trait; however no major seed crops have been shown capable of apomixis.

Desired outcome(s) that stretch current capabilities

- Develop genetic or molecular tools to fix genetic hybrids across outcrossing species
- Understand the mechanisms of apomixis
- Understanding of hybrid vigor in plants and relationship to yield
- Understanding of self-compatibility and self-incompatibility in plant sexual reproduction

Sector: Transforming Bioenergy & Agriculture with Synthetic Biology

Use Case for Cells for Conversion: Metabolic Engineering of C3 Plants to C4 plants

End Product: Ability to maximize plants ability to fix CO2 leading to C3 plants that can function in like C4 plants in photosynthetic efficiency

Organism(s) if applicable: Plants/crops

Terrestrial plants do not have the ability to move so their ability to adapt to changing environments (temperature, water, and sunlight) requires precise genetic regulatory mechanisms that control their growth and development, nutrient composition, and yield. Can mixing of these regulatory mechanisms across plants species generate plants and crops that have the potential to generate optimal food, feed, and fiber that would support the agricultural needs of U.S. and the world in the future.

Over 3 billion people depend on rice for survival across the globe. Due to predicted population increases and a general trend towards urbanization, land that currently provides enough rice to feed 27 people will need to support 43 by 2050. In this context, rice yields need to increase by 50% over the next 35 years. Given that traditional breeding programs have hit a yield barrier, the world (South Asia and sub-Saharan Africa in particular) is facing an unprecedented level of food shortages.

Rice is evolutionarily a C3 plant. In C3 plants, the bundle sheath cells do not contain chloroplasts. In C4 plants, the bundle sheath cells contain chloroplasts. In C3 plants, the carbon dioxide fixation takes place only at one place. In C4 plants, the carbon dioxide fixation takes places twice (one in mesophyll cells, second in bundle sheath cells). Because of this inherent difference, C4 plants are more efficient.at fixing CO2 than C3 plants. Therefore, the introduction of C4 traits into current C3 rice is predicted to increase photosynthetic efficiency by 50%, improve nitrogen use efficiency and double water use efficiency.

Desired outcome(s) that stretch current capabilities

- Develop genetic or molecular tools to transform C3 plants to C4 plants for enhanced CO2 assimilation
- Ability to enhanced grain and biomass yield in C3 crops (such as rice)

Sector: Transforming Bioenergy & Agriculture with Synthetic Biology

Use Case for Cells for Conversion: Metabolic Engineering of Lignin Synthesis and Composition

End Product: Ability to efficiently control the composition of lignin synthesis or regioselect lignin types

Organism(s) if applicable: bioenergy crops or other plants where lignin composition is important

Lignin provides structural support, a mechanical barrier against microbial infestation and facilitates movement of water inside plant systems. It is the second most abundant natural polymer in the terrestrial environments other than cellulose. Lignin is one of the most important secondary metabolite which is produced by the phenylalanine/tyrosine metabolic pathway in plant cells. It possesses unique routes for the production of bulk and specialty chemicals with aromatic/phenolic skeletons. Natural lignin composition in plants is widely diverse in many respects : (1) type and frequency of monomer (p-hydroxyphenyl(H)-, guaiacyl (G) - and syringyl(S) -propane) units, (2) type and frequency of interunit bonds, (3) shape, size, and (4) linkages between lignin and polysaccharides. This diversity is found also in different kinds of cells of different plant species.

The commercial applications of lignin are limited due to this diversity and it is often recognized for its negative impact on the biochemical conversion of lignocellulosic biomass to fuels and chemicals. Understanding of the structure of lignin monomers and their interactions among themselves, as well as with carbohydrate polymers in biomass, is vital for the development of innovative biomass deconstruction processes and thereby valorization of all biopolymers of lignocellulosic residues, including lignin.

Desired outcome(s) that stretch current capabilities

- Chemical or regulator mechanisms that can limit the diversity of specific types of lignin that are produced in plants
- Capabilities of post-harvest triggers that allow lignin populations to convert to a specific type that is conducive to sustainably conversion of biomass for biofuels and bio-products production (via engineered microbes or enzymes)

Clonal seeds in hybrid rice using CRISPR/Cas9

Chun Wang[1], Qing Liu[1], Yi Shen[2], Yufeng Hua[1], Junjie Wang[1], Jianrong Lin[1], Mingguo Wu[1], Tingting Sun[1], Zhukuan Cheng[2], Raphael Mercier[3], Kejian Wang[1]

[1]State Key Laboratory of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China.

[2]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China.

[3]Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, RD10, 78000 Versailles, France.

Correspondence should be addressed to K.W. (wangkejian@caas.cn)

Heterosis, the observation that first generation hybrids outcompete the parental lines, is widely used in increasing the productivity and yield of agricultural crops[1,2]. However, heterosis is lost in the following generations because of genetic segregation. In addition, the high cost of hybrid seed production hinders the application of heterosis in many crops. Clonal reproduction through seeds could be revolutionary for agriculture by allowing self-propagation of F_1 hybrids[3,4]. Here we show that heterozygosity of F_1 hybrid rice can be fixed and thus propagated without additional crossing. First, we showed that multiplex editing of three key meiotic genes[5,6] in hybrid rice leads to the production of clonal diploid gametes and tetraploid seeds. Next, editing of the *MATRILINEAL* (*MTL*) gene that involved in fertilization[7,8] results in the induction of haploid seeds in hybrid rice. By simultaneous editing of these four endogenous genes in hybrid rice using the CRISPR/Cas9 system, we obtained in one generation plants able to propagate clonally through seeds. This opens the possibility to fix heterozygosity of hybrid varieties in food crops.

Heterosis (also known as hybrid vigor) is a phenomenon whereby hybrid offspring of genetically diverse individuals display increased vigor relative to their homozygous parents. Heterosis has been widely applied in agriculture to dramatically improve the production and to broaden adaptability of crops[1,2]. However, the essential process of hybrid seed production increases the seed cost and even prohibits its application in many crops. It has been proposed to fix the heterosis of hybrid crop by introduction of apomixis[3]. Apomixis is an asexual reproductive strategy where the

offspring were generated through seeds, but without meiosis and fertilization. Although it has been described in many flowering plant taxa[9], apomixis has not been reported in major crops. Previously, it was revealed that combined mutations of three genes that affect key meiotic processes created a genotype called *MiMe* (*Mitosis instead of Meiosis*) in which meiosis is totally replaced by mitotic-like division, leading to the production of male and female clonal diploid gametes in Arabidopsis and rice[5,6]. However, the self-fertilization of *MiMe* resulted in doubling of ploidy at each generation. By crossing Arabidopsis *MiMe* with CenH3-mediated chromosome elimination line, clonal diploid offspring were obtained[4]. However, the system still relies on the crossing between different plants and the CENH3-mediated chromosome elimination appeared to be difficult to transfer to other species[10]. Therefore, further work is required to achieve the aim of heterosis fixation in self-fertilized hybrids.

Firstly, to test the feasibility of *MiMe* technology in hybrid rice varieties, we performed experiments on Chunyou84 (CY84), an elite inter-subspecific hybrid rice from a cross between the maternal Chunjiang 16A (16A), a *japonica* male sterile line, and the paternal C84, an *indica-japonica* intermediate type line (Extended Data Fig.1). To ensure rapid generation of *MiMe* in the hybrid CY84 background, we simultaneously edited the *REC8*, *PAIR1* and *OSD1* genes using our previously developed multiplex CRISPR/Cas9 system[11] (Fig. 1a). In the primary transformed plants, 7 of 32 plants were identified as frameshift triple mutants, and three of them were analyzed (Extended Data Fig.2). The triple mutant (*MiMe*) could not be distinguished from the wild-type CY84 based on its growth or morphology (Extended

Data Fig.3). To test whether the meiosis was turned into a mitotic-like division, we investigated the male meiotic chromosome behavior in both wild type and *MiMe*. In the wild-type CY84 (Extended Data Fig.4a-f), 12 bivalents were scattered at diakinesis and aligned along the equatorial plate at metaphase I. The 12 pairs of homologous chromosomes separated at anaphase I and produced tetrad spores after the second meiotic division. In *MiMe* (Extended Data Fig.4g-i), 24 univalents were found in diakinesis and aligned at metaphase I. In anaphase I, 24 pairs of chromatids segregated into two groups and produced dyads of spores, suggesting that the meiosis has been turned into a mitotic-like division. We next examined the ploidy of spores of *MiMe* by performing fluorescent in situ hybridization (FISH) analyses using a 5S rDNA-specific probe, which identifies chromosome 11 of rice. Only one signal was observed in CY84 spores (n=30), while two signals were constantly observed in *MiMe* spores (n=40, Fig. 1b), showing that diploid gametes were generated in *MiMe*. We also investigated the fertility of *MiMe* mutant and found that the panicle seed setting rate in *MiMe* was ~81.2% (n=4043), which is comparable to that of wild type (~79.1%, n=3876), (Fig.1c, Table 1), suggesting that simultaneously editing of these three genes do not obviously affect fertility in this hybrid variety. The ploidy of the progeny of *MiMe* plant was investigated by flow cytometry and all (n=123) were found to be tetraploid plants (Fig.1d, Table 1). Further, we found that these progenies (n=123) retained completely the heterozygosity of their parent CY84 for 10 tested Insertion-deletion (Indel) makers (Fig.1e). And these progenies of *MiMe* displayed reduced fertility, increased grain size and elongated awn length compared to wild type,

all of which being typical characteristics of tetraploid rice (Fig.1f). These results show that the *MiMe* phenotype can be rapidly introduced into hybrid rice varieties using CRISPR/Cas9 genome editing technique.

MiMe clonal gametes participate in normal self-fertilization, giving rise to progeny with doubled ploidy. This ploidy doubling must be prevented to achieve apoximis. Recently, it was reported that the *MATRILINEAL* (*MTL*) gene, a sperm-specific phospholipase, triggers haploid induction in maize[7,8]. To test whether the homologous gene could be manipulated to induce haploid in self-fertilized hybrid rice, we edited the *MTL* gene in CY84 (Fig. 2a). 11 of 32 transformed plants were identified as frameshift mutants, and three of them were analyzed (Extended Data Fig.5). The *mtl* mutants showed normal vegetative growth (Extended Data Fig.3), but the seed-setting rates significantly reduced to ~11.5% (n=5180, Fig. 2b, Table 1). 12 Indel markers (1 per chromosome) that were polymorphic between the two parents were used to determine the genotype of the progenies of *mtl* plants (Extended Data Table1). In the wild-type CY84 progeny, no plants homozygous at all markers were found (n=220, Table 1). In contrast, 11 plants among 248 *mtl* progenies appeared to be homozygous for all markers (Fig. 2c, Table 1). Flow cytometry results showed that 9 of these plants were indeed haploid, while 2 were diploid, presumably resulting from spontaneous doubling of haploid embryos (Fig. 2d, Table 1). To further classify the genotype of those identified plants, the whole genomes of 2 haploids, 2 doubled haploids of *mtl* progenies, and 2 offspring plants of wild-type CY84 were resequenced with a depth of 30-fold. A total of 78,909 single nucleotide polymorphisms (SNPs)

that differed between two parents were screened out for detailed genotype analysis. Whole genome sequencing revealed that the haploids and doubled haploids were homozygous at all loci (Fig. 2e), and recombinant compared to the parental genome, suggesting that they are respectively derived from a single gamete. The haploid plants showed reduced plant height, decreased glume size and loss of fertility, while the doubled haploid plant displayed normal vegetative and reproductive growth (Fig. 2f). The results demonstrated that haploid plants can be generated by self-fertilization of hybrid varieties.

Since turning meiosis into mitosis and paternal genome elimination is possible in self-fertilized hybrid rice, we next test the possibility of inducing heterozygosity fixation without additional crossing in hybrid rice by simultaneously editing four genes, namely *OSD1*, *PAIR1*, *REC8* and *MTL* in CY84 (Fig. 3a-b). Among 22 transgenic plants, three were identified by DNA sequencing as *osd1*/*pair1*/*rec8*/*mtl* quadruple mutants (namely *Fix*, *Fixation of hybrids*) and used for further analysis (Extended Data Fig.6). The *Fix* mutants grew normally during the vegetative stage (Fig. 3c). During reproductive stage, the male meiotic chromosome behavior was investigated and found to be indistinguishable from that of *MiMe* (Extended Data Fig.4j-l). The panicle seed setting percentage was found to be ~4.5% (n=5850) (Table 1, Fig. 3c), which is slightly lower than that of the *mtl* mutant. In the progeny seedlings, the ploidy was investigated using flow cytometry. Among 145 progeny of *Fix* mutants, 136 were identified as tetraploid and 9 as diploid (Fig. 3d, Table 1). To investigate whether the heterozygosity was fixed in these diploid offspring, the

124 genomes of 2 diploid and 2 tetraploid offspring plants of *Fix* were resequenced with
125 an average of 30× coverage. Bioinformatic analysis revealed that all the 78,909
126 SNPs were heterozygous in both these diploid and tetraploid progeny plants, and were
127 thus genetically identical to the hybrid rice CY84 (Fig. 3e). Finally, we investigated
128 the phenotype of the potential clonal plants of *Fix*. All these 9 diploid plants displayed
129 normal glume size and awn length, and showed a reduced seed setting (~10%,
130 n=2726), which were similar to their parent *Fix* plants (Fig. 3f). Taken together, the
131 diploid progeny of *Fix* plant displayed the same ploidy, the same heterozygous
132 genotype, and the similar phenotype with the parent *Fix* plants, implying that *Fix* is
133 able to produce clonal seeds and fix the heterozygosity of hybrid rice.

134 Our findings revealed that hybrids can be self-propagated through seeds by
135 targeted editing of four endogenous genes in rice hybrid varieties. Simultaneous
136 editing of *REC8*, *PAIR1* and *OSD1* genes does not have obvious adverse effects on the
137 growth and reproduction of the hybrid. On contrast, the *MTL* gene used to induce
138 paternal genome elimination has impacts on rice fertility and is not fully penetrant;
139 further work is thus required to allow this technology to reach the rice fields. However,
140 the findings in this study revealed a strategy to fix heterozygosity in rice. Considering
141 the establishment of multiplex genome editing technology in many other crops along
142 with the conservation of these four genes, the strategy might extend heterosis
143 application in agriculture.
144

Methods

Plasmid construction. The plasmids expressing the CRISPR/Cas9 system were constructed *via* the isocaudamer ligation method, as previously described[11]. The modified single guide RNAs (sgRNAs) scaffold and *ACTIN1* promoter-driven Cas9 were used to increase the mutation rate in this study[12]. Briefly, the double-stranded overhangs of target oligoes (listed in Extended Data Table1) were ligated into the SK-sgRNA vectors digested with *Aar*I. Then the sgRNAs of *OSD1* (digested with *Kpn*I and *Sal*I), *PAIR1* (digested with *Xho*I and *Bgl*II) and *REC8* (digested with *Bam*HI and *Nhe*I) were assembled into one pC1300-ACT:Cas9 binary vector (digested with *Kpn*I and *Xba*I) using T4 ligase to obtain the vector pC1300-ACT:Cas9-sgRNA^{OSD1}-sgRNA^{PAIR1} -sgRNA^{REC8} for generation of *MiMe*. The sgRNA of *MTL* (digested with *Kpn*I and *Nhe*I) was assembled into pC1300-ACT:Cas9 binary vector (digested with *Kpn*I and *Xba*I) to obtain the vector pC1300-ACT:Cas9-sgRNA^{MTL} for generation of *mtl*. The sgRNA of *MTL* (digested with *Kpn*I and *Nhe*I) was assembled into pC1300-ACT:Cas9-sgRNA^{OSD1}-sgRNA^{PAIR1}-sgRNA^{REC8} vector (digested with *Kpn*I and *Xba*I) to obtain the vector pC1300-ACT:Cas9-sgRNA^{OSD1}-sgRNA^{PAIR1}-sgRNA^{REC8}-sgRNA^{MTL} for generation of *Fix*.

Rice transformation and growth conditions. The hybrid rice Chunyou 84 (CY84) was used as the host variety in this study. The generation of transgenic rice, by *Agrobacterium*-mediated transformation with the strain EHA105, was performed by the Biogle company (Hangzhou, China).

The T_0 generation of transgenic plants were grown in the transgenic paddy fields of the China National Rice Research Institute in Hangzhou, China (at N 30.32° , E 120.12°) in the summer of 2017. The T_1 plants were grown in greenhouse in the winter of 2017.

Detection of genome modifications. Genomic DNA was extracted from approximately 100 mg of rice leaf tissue *via* the CTAB method. PCR was conducted with KOD FX DNA Polymerase (Toyobo, Osaka, Japan) to amplify the genomic regions surrounding the target sites. The primers are listed in Extended Data Table1. The fragments were sequenced by the Sanger method and decoded by the degenerate sequence decoding method[13].

Cytological analyses. Young panicles of meiosis stage were harvested and fixed in Carnoy's solution (ethanol:glacial acetic, 3:1). Microsporocytes undergoing meiosis were squashed in an acetocarmine solution. Slides were frozen in liquid nitrogen and the coverslips were removed with a blade quickly. Chromosomes were counterstained with 4',6-diamidinophenylindole (DAPI) in an antifade solution (Vector Laboratories, Burlingame, CA). Microscopy was conducted using an Olympus BX61 fluorescence microscope with a microCCD camera.

Fluorescence *in situ* hybridizaiton (FISH) analysis was conducted as described previously[14]. The plasmid pTa794 was used as FISH probe to quantify the 5S rDNA.

Genotyping with Indel Markers. Insertion-deletion (Indel) markers to distinguish genotypes of heterozygote and homozygote were designed based on the whole-genome sequences of C84 and 16A. The primers are listed in Extended Data

Table1. The genotyping was performed by normal PCR program using 2× Taq Master Mix (Novoprotein Scientific, China), and the PCR products were detected using 5% agarose gels.

Flow cytometry determination of DNA content in leaf cell nuclei. The ploidy of leaf cell was determined by estimating nuclear DNA content using flow cytometry. All procedures were done at 4 ºC or on ice. Approximately ~ 2 cm^2 of leaf tissue was chopped using a new razor blade for 2 to 3 minutes in 1 ml LB01 Buffer (15 mM Tris, 2 mM Na_2EDTA, 0.5 mM spermine tetrahydrochloride, 80 mM KCl, 20 mM NaCl, 0.1% Triton X-100, 15 mM β-mercaptoethanol, pH 7.5, filter through a 0.22 μm filter). The homogenate was filtered through the 40-μm nylon filter followed by centrifugation (1200× rpm, 5 min) to collect the nuclei. The supernatant was discarded and the pellet was resuspended with 450 μL of fresh LB01 Buffer, then 25 μl of 1 mg/ml propidium iodide (PI, Sigma P4170) and 25 μl of 1 mg/ml DNase-free RNase A (Sigma V900498) were added to stain the DNA. The stained samples were incubated on ice in darkness for 10 minutes prior to analysis. The samples were analyzed using BD Accuri C6 flow cytometer, with the laser illumination at 552 nm and 610/20 nm filter. The gating strategy was provided in Supplementary Information. Samples with the same result of CY84 were deemed as diploids, which the first peak of relative fluorescence at ~100 (x10,000). And the samples with the first peak of relative fluorescence at ~50 (x10,000) were deemed as haploids, while samples with the first peak of relative fluorescence at ~200 (x10,000) were deemed as tetraploids.

Whole genome re-sequencing and genotype calling. The 150-bp paired-end reads

were generated by Illumina Hiseq2500, covering approximately an average depth of 30× for each sample. The short-read sequence data have been deposited in the NCBI Sequence Read Archive (SRP149641, SRP149677). The raw paired-end reads were first filtered into clean data using NGSQCtookit v2.3.3[15]. The cutoff value for PHRED quality score was set to 30. Clean reads of each accession were aligned against the rice reference genome (IRGSP 1.0) using the software SOAPaligner (soap version 2.21) [16] with the parameters of '-m 200, -x 1000, -l 35, -s 42, -v 5' and '–p 8'. To get high-quality SNPs, reads that could be mapped to different genomic positions were excluded by SOAPsnp[17]. Uniquely mapped single-end and paired-end results were used in the SNP calling. Genotype calling was carried out in the whole genome region using these SNPs which are heterozygous in the parent. The window size (the number of n consecutive SNPs in a window) was 0.1 K. And the recombination map was constructed for each chromosome.

Data availability. Whole genome sequencing data are deposited in the NCBI Sequence Read Archive (SRP149641, SRP149677). Patent applications have been filed relating to work in this manuscript.

1 Schnable, P. S. & Springer, N. M. Progress toward understanding heterosis in crop plants. *Annu Rev Plant Biol* **64**, 71-88 (2013).

2 Birchler, J. A., Auger, D. L. & Riddle, N. C. In search of the molecular basis of heterosis. *Plant Cell* **15**, 2236-2239 (2003).

3 Spillane, C., Curtis, M. D. & Grossniklaus, U. Apomixis technology development-virgin births in farmers' fields? *Nat Biotechnol* **22**, 687-691 (2004).

4 Marimuthu, M. P. A. *et al.* Synthetic clonal reproduction through seeds. *Science* **331**, 876

(2011).

5 d'Erfurth, I. *et al.* Turning meiosis into mitosis. *PLoS Biology* **7**, e1000124 (2009).

6 Mieulet, D. *et al.* Turning rice meiosis into mitosis. *Cell Research* **26**, 1242-1254 (2016).

7 Kelliher, T. *et al.* MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. *Nature* **542**, 105-109 (2017).

8 Li, X. *et al.* Single nucleus sequencing reveals spermatid chromosome fragmentation as a possible cause of maize haploid induction. *Nat Commun* **8**, 991 (2017).

9 Sailer, C., Schmid, B. & Grossniklaus, U. Apomixis Allows the Transgenerational Fixation of Phenotypes in Hybrid Plants. *Curr Biol* **26**, 331-337 (2016).

10 Karimi-Ashtiyani, R. *et al.* Point mutation impairs centromeric CENH3 loading and induces haploid plants. *Proc Natl Acad Sci U S A* **112**, 11211-11216 (2015).

11 Wang, C., Shen, L., Fu, Y., Yan, C. & Wang, K. A Simple CRISPR/Cas9 System for Multiplex Genome Editing in Rice. *Journal of genetics and genomics* **42**, 703-706 (2015).

12 Hu, X., Meng, X., Liu, Q., Li, J. & Wang, K. Increasing the efficiency of CRISPR-Cas9-VQR precise genome editing in rice. *Plant Biotechnology Journal* **16**, 292-297 (2018).

13 Ma, X., Chen, L., Zhu, Q., Chen, Y. & Liu, Y. G. Rapid Decoding of Sequence-Specific Nuclease-Induced Heterozygous and Biallelic Mutations by Direct Sequencing of PCR Products. *Mol Plant* **8**, 1285-1287 (2015).

14 Zhang, W. *et al.* The transcribed 165-bp CentO satellite is the major functional centromeric element in the wild rice species Oryza punctata. *Plant Physiol* **139**, 306-315 (2005).

15 Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7**, e30619 (2012).

16 Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966-1967 (2009).

17 Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res* **19**, 1124-1132 (2009).

265 **Acknowledgements** This research was supported by the Agricultural Science and
266 Technology Innovation Program of Chinese Academy of Agricultural Sciences, and
267 the National Natural Science Foundation of China (No. 31401363).

268 **Author Contributions** C.W., and K.W. conceived and designed the study. C.W., Y.S.,
269 and Z.C. performed the lab experiments. Q.L., and T.S. conducted the computational
270 analyses. Y.H., and J.W. carried out the field experiments. J.L., and M.W. provided the
271 rice varieties and helped with the field management. C.W., R.M., and K.W. wrote the
272 manuscript.
273

274 **Table1 Ploidy analysis of the progeny of CY84, *MiMe*, *mtl* and *Fix* lines**

275

	Line	Seed setting percentage	Progeny tested	Haploid+DH (%)	Diploid (%)	Tetraploid (%)
CY84	#1	77.2% (1151/1490)	65	0	65	0
	#2	81.3% (951/1170)	73	0	73	0
	#3	79.1% (962/1216)	82	0	82	0
MiMe	#1	81.9% (1178/1439)	35	0	0	35 (100%)
	#2	79.2% (877/1108)	43	0	0	43 (100%)
	#3	82.1% (1228/1496)	45	0	0	45 (100%)
mtl	#1	9.1% (101/1103)	77	6+0 (7.8%)	71	0
	#2	13.6% (217/1601)	90	2+1 (3.3%)	87	0
	#3	11.3% (280/2476)	81	1+1 (2.5%)	79	0
Fix	#1	3.7% (63/1725)	39	0	2 (5.1%)	37
	#2	5.2% (124/2373)	64	0	3 (4.7%)	61
	#3	4.3% (76/1752)	42	0	4 (9.5%)	38

276

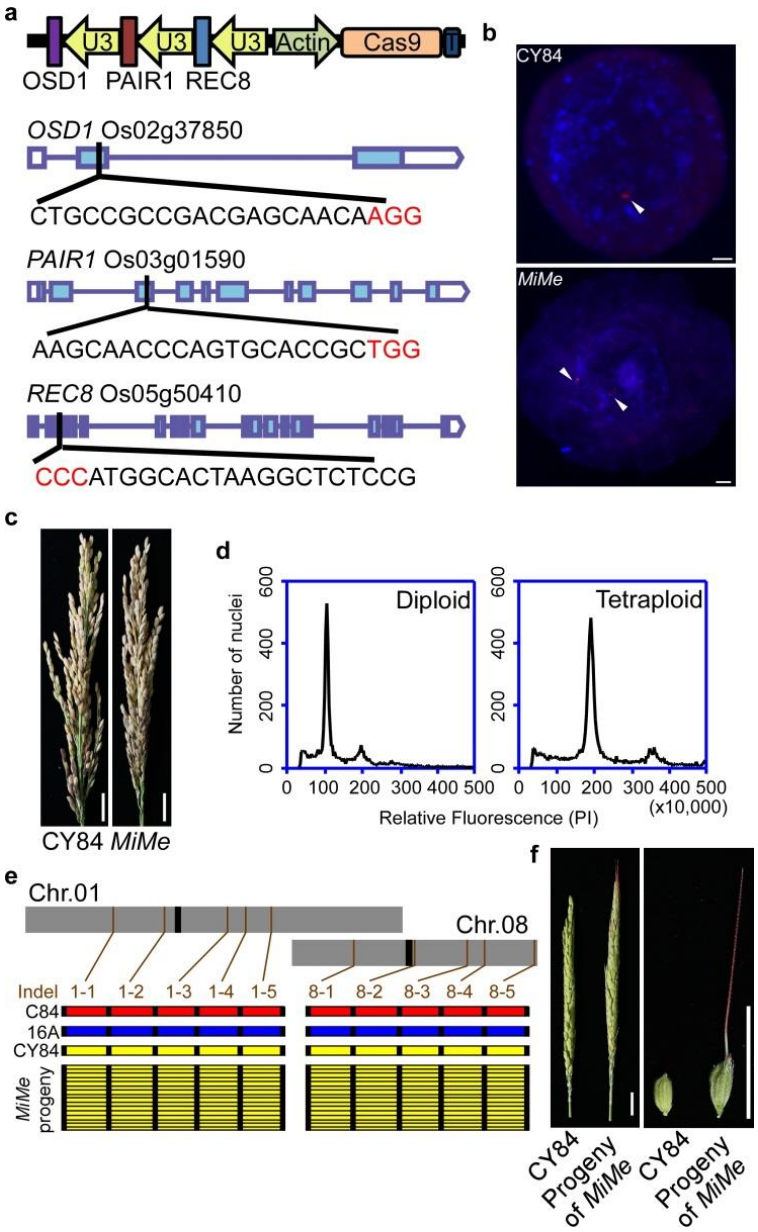
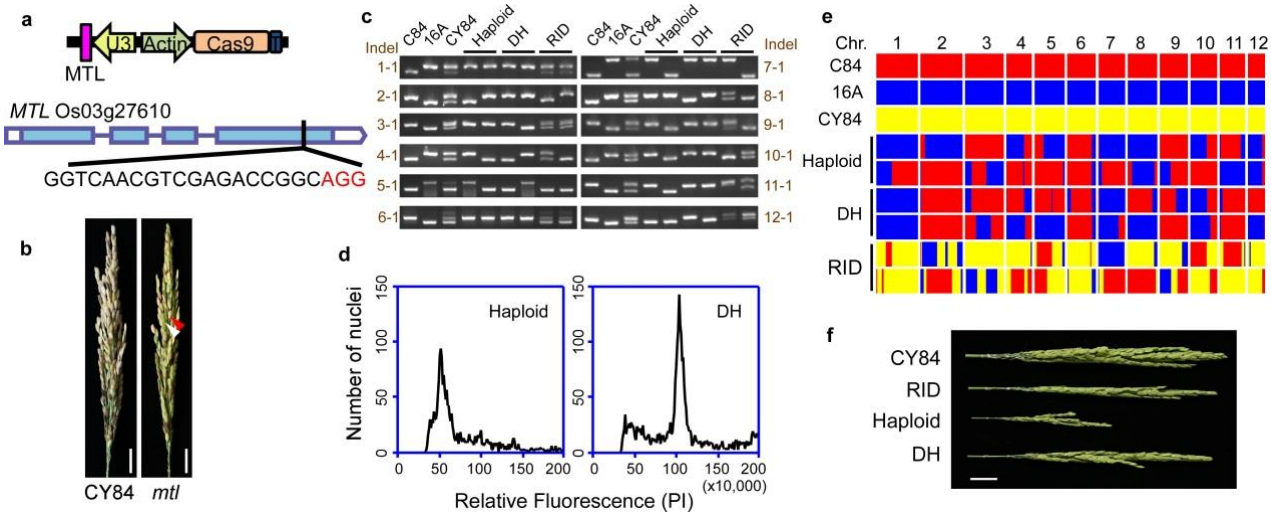


Figure 1 | Turning meiosis into mitosis in hybrid rice variety Chunyou84 (CY84). **a**, Schematic diagram of the structure of CRISPR/Cas9 vector targeting *OSD1*, *PAIR1* and *REC8*. **b**, The chromosomes of CY84 and *MiMe* were probed by digoxige-nin-16-dUTP-labled 5S rDNA (red signal, indicated with white arrow) in spores, showing one signal in wild-type CY84 and two signals in *MiMe*. Scale bars, 5 μm. **c**, Panicles of wild-type CY84 and *MiMe*. The fertility of *MiMe* is as high as that of wild-type CY84. Scale bars, 2 cm. **d**, Ploidy analysis of CY84 (left) and the

progeny of *MiMe* (right) by flow cytometry, which is found to be diploid and tetraploid, respectively (Table 1). **e**, Genotype analysis of the paternal C84, maternal Chunjiang 16A (16A), hybrid variety Chunyou84 (CY84) and the progeny siblings of *MiMe*. 10 Indel markers distributed on chromosomes 1 and 8 were used to identify the genotype of the offspring of *MiMe*. Positions of markers (brown) and centromeres (black) are indicated along the chromosomes. For each marker, plants carrying the C84 allele are in red, plants carrying the 16A allele are in blue, while plants with both C84 and 16A alleles appear in yellow. Each row represents one plant, and each column indicates a locus. **f**, Panicles and grain shape of CY84 and the progeny of *MiMe*. The progeny of *MiMe* displayed reduced fertility, increased glume size and elongated awn length. Scale bars, 2 cm.



297

298 **Figure 2|Generation of haploid inducer line by editing the *MTL* gene in hybrid**
299 **rice variety CY84. a**, Schematic diagram of the structure of CRISPR/Cas9 vector
300 targeting *MTL*. **b**, Panicles of the WT and *mtl* in CY84 background. The fertility was
301 decreased in *mtl*, white arrow indicates aborted seed, and red arrow shows fertile seed.
302 Scale bars, 2 cm. **c**, Determination of the genotype of haploids, doubled haploids (DH)
303 and recombinant inbred diploids (RID) using 12 Indel markers (1 per chromosome).
304 Plants homozygous at all markers in the progeny siblings of *mtl* were identified as
305 haploid or DH. **d**, Ploidy analysis of the haploid and DH by flow cytometry (Table 1).
306 **e**, Whole genome sequencing of the haploid, DH and RID plants. 12 blocks represent
307 12 chromosomes. The SNPs of C84 allele are in red, the SNPs of 16A allele are in
308 blue, and co-existence of both alleles are in yellow. **f**, Panicles of wild-type CY84 and
309 *mtl* progeny, including RID, haploid and DH plants. Scale bars, 2 cm.

310

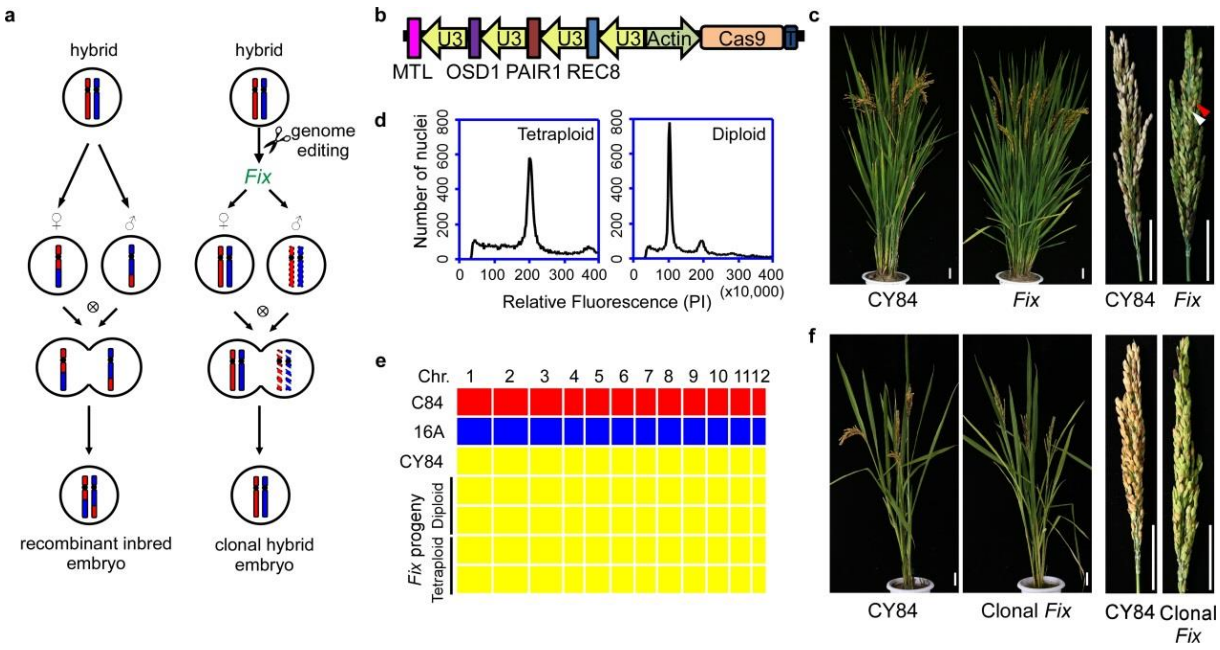


Figure 3|Fixation of rice heterozygosity by multiplex gene editing in hybrid rice variety CY84. a, The model of fixation of heterozygosity of hybrid. In normal sexual reproduction (left), recombinant inbred embryos are generated by fusion of recombined haploid gametes. The clonal reproduction strategy (right) is based on two components: meiosis is turned into mitosis to produce clonal diploid gametes (*MiMe*), and the genome of male gamete is eliminated by knocking out the *MTL* gene. The progeny of self-fertilized *Fix* is genetically identical to the hybrid parent. **b**, Schematic diagram of the structure of CRISPR/Cas9 vector simultaneously targeting *OSD1*, *PAIR1*, *REC8* and *MTL*. **c**, Comparison of the morphology and panicles of CY84 and *Fix* (*osd1 pair1 rec8 mtl*). The fertility was decreased in *Fix.* An aborted seed is indicated with white arrow, and a normally developed seed is indicated with red arrow. Scale bars, 5 cm. **d**, Ploidy analysis of the progeny of *Fix* by flow cytometry, including tetraploid (left) and diploid (right), respectively. **e**, Whole genome sequencing of the diploid and tetraploid progenies of *Fix*. The SNPs of C84

allele are in red, the SNPs of 16A allele are in blue, and co-existence of both alleles are in yellow. 12 blocks represent 12 chromosomes. The diploid and tetraploid progenies of *Fix* are heterozygous, identical to CY84. **f**, Comparison the morphology and panicles of wild-type CY84 and the diploid progeny of *Fix*. Both plants were grown in the glasshouse. The clonal *Fix* displayed normal growth except the reduced fertility, which is similar to that of parent *Fix* plant. Scale bars, 5 cm.

LETTER

https://doi.org/10.1038/s41586-018-0785-8

A male–expressed rice embryogenic trigger redirected for asexual propagation through seeds

Imtiyaz Khanday[1,2], Debra Skinner[1], Bing Yang[3], Raphael Mercier[4] & Venkatesan Sundaresan[1,2,5]*

The molecular pathways that trigger the initiation of embryogenesis after fertilization in flowering plants, and prevent its occurrence without fertilization, are not well understood[1]. Here we show in rice (*Oryza sativa*) that BABY BOOM1 (BBM1), a member of the AP2 family[2] of transcription factors that is expressed in sperm cells, has a key role in this process. Ectopic expression of *BBM1* in the egg cell is sufficient for parthenogenesis, which indicates that a single wild-type gene can bypass the fertilization checkpoint in the female gamete. Zygotic expression of *BBM1* is initially specific to the male allele but is subsequently biparental, and this is consistent with its observed auto-activation. Triple knockout of the genes *BBM1*, *BBM2* and *BBM3* causes embryo arrest and abortion, which are fully rescued by male-transmitted *BBM1*. These findings suggest that the requirement for fertilization in embryogenesis is mediated by male-genome transmission of pluripotency factors. When genome editing to substitute mitosis for meiosis (*MiMe*)[3,4] is combined with the expression of *BBM1* in the egg cell, clonal progeny can be obtained that retain genome-wide parental heterozygosity. The synthetic asexual-propagation trait is heritable through multiple generations of clones. Hybrid crops provide increased yields that cannot be maintained by their progeny owing to genetic segregation. This work establishes the feasibility of asexual reproduction in crops, and could enable the maintenance of hybrids clonally through seed propagation[5,6].

Understanding the molecular pathway that underlies the initiation of embryogenesis by a fertilized egg cell is a major unresolved problem in plant development[1]. In animals, the initiation of embryogenesis depends upon defined maternal factors[7]. In plants, two contrasting models have been proposed: one suggests that the two parental genomes contribute equally[8], whereas the other considers that the maternal genome has the primary role in early embryogenesis[9,10]. The identity and parental origin of the specific factors in plants that trigger zygotic development are as yet undetermined. We have previously used rice to elucidate transcriptome dynamics during the zygotic transition[11] and found that *BABY BOOM* (*BBM*)-like transcription factors of the APETALA 2/ETHYLENE RESPONSE FACTOR (AP2/ERF) superfamily[12] are expressed in zygotes after fertilization, which suggests a potential role in the initiation of embryogenesis (Extended Data Table 1a). *BBM* genes from *Arabidopsis thaliana* and *Brassica napus* can ectopically induce somatic embryos[13]; however, a role for these genes in the initiation of zygotic embryos has not been established[2]. We first determined that ectopic expression of *BBM1*—a *BBM*-like gene expressed in rice zygotes—also resulted in somatic embryos, both by examining their morphology and by using embryo marker genes (Extended Data Fig. 1a–d). Because *BBM1* expression increases with the age of the zygote[11] (Extended Data Table 1a), we investigated whether its expression is autoregulated, by inducing a constitutive BBM1–glucocorticoid receptor (GR) fusion in somatic tissues using dexamethasone (DEX) (Extended Data Fig. 1e). Quantitative PCR after reverse transcription (RT–qPCR), using allele-specific primers, showed that the expression of endogenous *BBM1*—but not the *BBM1-GR* fusion transgene—was highly induced after 24 h of DEX treatment (Extended Data Fig. 1f–h). This expression was maintained in the presence of the protein-biosynthesis inhibitor cycloheximide (CYC), indicating that BBM1 auto-activation is likely to be direct (Extended Data Fig. 1h). Auto-activation might be a conserved feature of *BBM* genes, because *B. napus* BABY BOOM can activate the expression of *Arabidopsis BBM*[14].

Our previous study of hybrid zygote transcriptomes[11] indicated that, although most zygotic transcripts were from the female genome, a few de novo transcription factors—including *BBM1*—had male-derived transcripts. We used RT–PCR amplification across single nucleotide polymorphisms (SNPs) in *BBM1* to confirm that, at 2.5 h after pollination (HAP) (corresponding to karyogamy), only the male *BBM1* allele is expressed in reciprocal crosses of *indica* and *japonica* cultivars[11] (Extended Data Fig. 2a). These results were confirmed in isogenic zygotes in the *japonica* Kitaake cultivar. We reciprocally crossed wild-type plants to transgenic plants that carried a translational fusion of the *BBM1* genomic locus to GFP (*BBM1-GFP*) (Extended Data Fig. 2b). Zygotes at 2.5 HAP displayed GFP expression only if the *BBM1-GFP* transgene was transmitted from the male parent (Fig. 1a). Consistent with this observation, in *BBM1-GFP* selfed progeny, GFP was detected in only about half of the zygotes, instead of the three-quarters ratio that would be expected if there is no parent-of-origin bias (Fig. 1a). Subsequently, GFP expression can be detected from the female allele in 6.5 HAP zygotes, corresponding to mid-to-late G2 phase (Extended Data Fig. 2c, d). Because BBM1 is capable of auto-activation of its own promoter (Extended Data Fig. 1h), the late expression of *BBM1* from the female allele might result from earlier expression of *BBM1* from the male allele. Other redundantly acting *BBM* genes might also contribute to this delayed activation (see below). *BBM1* expression continues through the later stages of embryo development (Extended Data Fig. 2e). In gametes, *BBM1* RNA can be detected by RT–PCR in sperm cells but not in egg cells (Extended Data Fig. 2f), which is consistent with RNA sequencing data[15] (Extended Data Table 1a). Furthermore, the BBM1–GFP fusion protein was expressed in sperm cells, which suggests that both transcription and translation of *BBM1* can occur in male gametes before fertilization (Extended Data Fig. 2g).

The expression of *BBM1* specifically from the male genome after fertilization, together with its capability to induce somatic embryogenesis, suggested that BBM1 could be a trigger of embryo development in the zygote (Extended Data Fig. 3a). In naturally apomictic (asexually reproducing) *Pennisetum squamulatum*, an apospory-specific locus contains multiple copies of a *BABY BOOM*-like gene that is expressed in egg cells before fertilization and induces parthenogenesis[16,17]. However, it is not known whether the BBM protein from the apomict has evolved novel capability in functional domains and interactions with other factors[16,17], or whether parthenogenesis might simply be a consequence of the expression pattern. To test whether wild-type rice BBM1 could initiate embryo development without fertilization, we ectopically expressed *BBM1* under an *Arabidopsis* egg-cell-specific promoter (*pDD45*)[18] that has previously been shown to confer egg-cell expression in rice[19] (Extended Data Fig. 3b, c). In emasculated flowers,

[1]Department of Plant Biology, University of California, Davis, CA, USA. [2]Innovative Genomics Institute, Berkeley, CA, USA. [3]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA. [4]Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, Versailles, France. [5]Department of Plant Sciences, University of California, Davis, CA, USA. *e-mail: sundar@ucdavis.edu

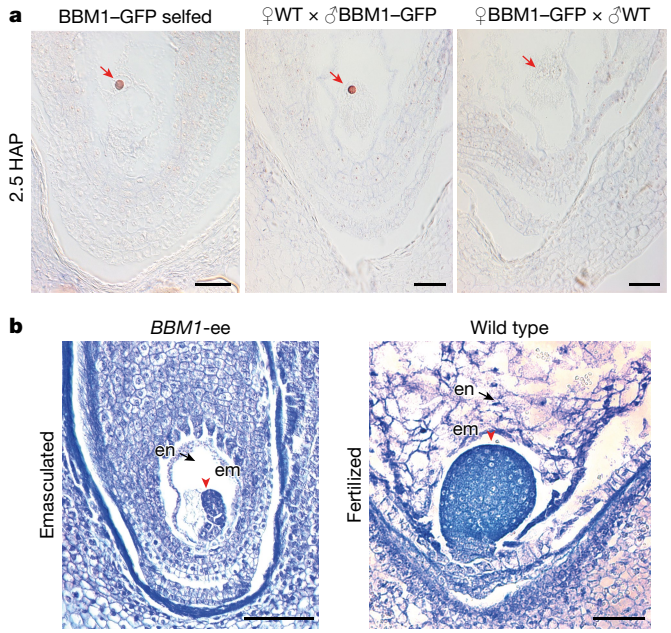


Fig. 1 | Paternal expression of *BBM1* in zygotes. a, Paternal allele-specific expression of *BBM1* in isogenic zygotes at 2.5 HAP. Expression of BBM1 fused to a GFP reporter was detected by antibody staining. GFP expression is observed only when BBM1–GFP is transmitted by the male parent ($n = 20$ for each panel, χ^2 test $P = 0.039$). Left, $n = 11/20$; middle, $n = 9/20$; right, $n = 0/20$. Red arrows point to zygote nuclei. WT, wild type. Scale bars, 25 μm. **b**, Development of parthenogenetic embryos (red arrowhead) by egg-cell-specific expression of *BBM1* in carpels of an emasculated *BBM1*-ee plant at nine days after emasculation ($n = 12/98$). In the absence of fertilization, endosperm development is not observed (black arrow). In fertilized control wild-type (4 days after pollination (DAP)) carpels, the development of both embryo (em; red arrowhead) and endosperm (en; black arrow) is observed ($n = 30$). Scale bars, 100 μm.

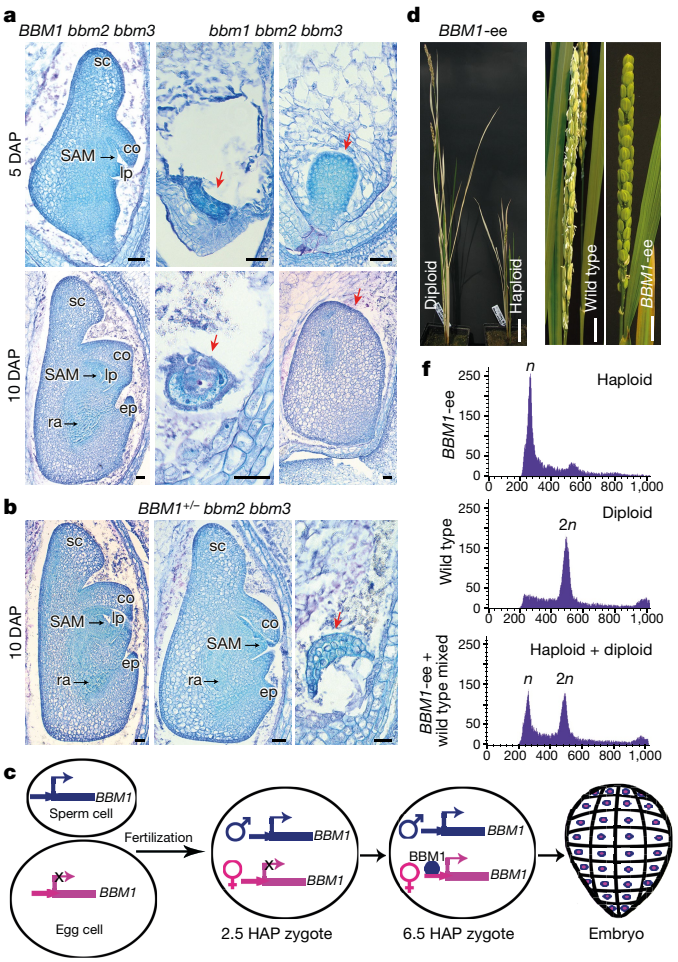


Fig. 2 | Phenotypes of *bbm1 bbm2 bbm3* mutant embryos and haploid induction. a, Embryos at 5 DAP (top) and 10 DAP (bottom). Embryos develop normally with wild-type *BBM1* ($n = 50$; left) but show an early arrest ($n = 24/82$; middle) or undergo a number of divisions without organ formation ($n = 58/82$; right) in *bbm1 bbm2 bbm3* triple homozygous mutant embryos. **b**, 10 DAP embryos that are heterozygous for *BBM1* but homozygous mutants for *bbm2* and *bbm3*. They show normal development ($n = 38/53$, left), are delayed ($n = 8/53$; middle), or show early arrest ($n = 4/53$; right). Scale bars, 100 μm. co, coleoptile; ep, epiblast; lp, leaf primordia; ra, radicle; SAM, shoot apical meristem; sc, scutellum. **c**, Schematic model of *BBM1* function in rice embryogenesis. **d**–**f**, Characterization of *BBM1*-ee induced haploids. **d**, Difference in height between parthenogenetic haploid and sexual diploid siblings ($n = 555$). Scale bar, 5 cm. **e**, A *BBM1*-ee parthenogenetic haploid panicle showing no anthesis (right) compared to an anthesis stage control wild-type panicle (left) ($n = 113$). **f**, Flow-cytometric DNA histograms for ploidy determination. Parthenogenetic haploid showing a 1*n* peak ($n = 19$, top), wild-type diploid with a 2*n* peak (middle) and a mixed sample of *BBM1*-ee and wild type showing 1*n* and 2*n* peaks (bottom).

we observed embryonic structures without endosperm development (Fig. 1b) in around 12% ($n = 98$) of ovules of *pDD45::BBM1* transformants (hereafter referred to as *BBM1*-ee, to denote *BBM1*-egg-cell expressed); these structures were absent in wild-type ovules ($n = 109$). Thus, the expression of a single wild-type transcription factor, BBM1, can overcome the requirement of fertilization for embryo initiation by an egg cell. The observation that a wild-type gene from a sexually reproducing plant is sufficient to induce parthenogenesis when misexpressed suggests that asexual reproduction could potentially evolve from the altered expression of existing genes within the sexual pathway.

Loss-of-function mutants of *BBM*-like genes in *Arabidopsis* and related plants have no embryonic phenotypes; consequently, their functions in early embryogenesis are as yet undefined[2]. Of the multiple *BBM*-like genes in rice, at least three—*BBM1*, *BBM2* and *BBM3* (Os11g19060, Os02g40070 and Os01g67410, respectively)—are consistently expressed in early zygotes (Extended Data Table 1a). We used the CRISPR–Cas9 system to generate *bbm1 bbm3* and *bbm2 bbm3* double mutants (Extended Data Fig. 4a, b), both of which were fully fertile. Crossing the double mutants and selfing (Extended Data Fig. 4c; see Methods) yielded no *bbm1 bbm2 bbm3* triple homozygous plants ($n = 52$). However, *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* plants were recovered and selfed (Extended Data Fig. 4d). Analysis of the progeny showed that approximately 36% failed to germinate (Extended Data Table 1b). Genotyping of the germinated seedlings suggested that the viability of the *bbm1 bbm2 bbm3* triple-mutant seeds was severely affected (2 out of 191 viable compared with the expected 48 out of 191; Extended Data Table 1b). *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* seedlings were also under-represented, which suggests that the viability of this genotype is also compromised (Extended Data Table 1b). A subset of the non-germinating seeds could be genotyped using their endosperm, and were found to be either homozygous or heterozygous for *bbm1* but not homozygous for *BBM1* (Extended Data Fig. 4e). The two *bbm1 bbm2 bbm3* triple homozygotes showed normal growth with no obvious vegetative or floral defects and produced normal seed sets, indicating that the *BBM1–BBM3* genes are not required for post-embryonic development. However, their progeny seeds failed to germinate (Extended Data Fig. 4f), confirming the requirement of *BBM1–BBM3* genes for seed viability.

To test whether the parent of origin affects seed viability, we performed reciprocal crosses of *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* to *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* plants. When the mutant *bbm1* allele was provided by the male parent, approximately 31% of the *bbm1/BBM1* progeny seeds failed to germinate (Extended Data Table 1c), whereas all progeny germinated when the *bbm1* allele was inherited from the female parent (Extended Data Table 1d). Thus,

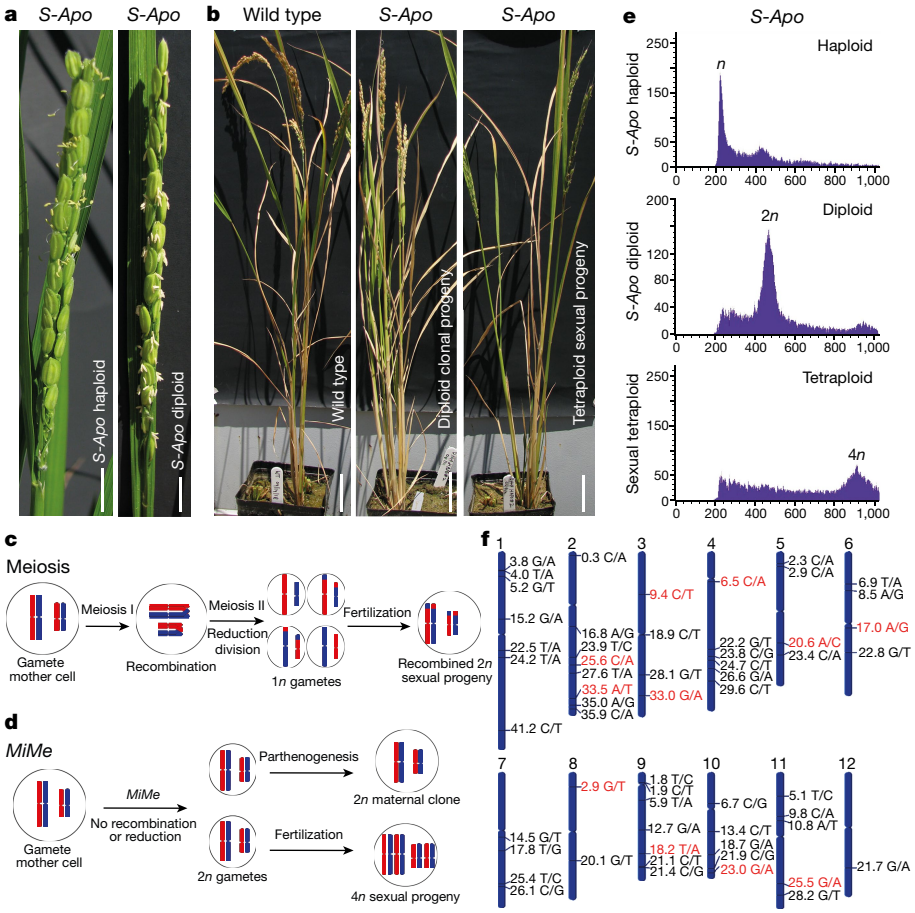


Fig. 3 | Characterization of asexually derived (apomictic) haploids and diploids. a, An *S-Apo* haploid (left; $n = 45$) and *S-Apo* diploid (right; $n = 57$) panicle undergoing anthesis. Scale bars, 1 cm. **b**, Comparison of wild-type (left), *S-Apo* diploid (middle; $n = 57/381$) and sexual tetraploid (right; $n = 324/381$) progeny plants. Scale bars, 5 cm. **c**, **d**, Schematics showing the difference between natural meiosis and *MiMe*. Whereas meiosis and fertilization produce recombined haploid gametes and diploid progeny, *MiMe* leads to the formation of diploid gametes that are clones of the mother plant. Parthenogenesis of a diploid egg cell produces clonal progeny and fertilization of diploid gametes leads to 4*n* sexual progeny. **e**, Flow-cytometric DNA histograms for ploidy determination of *S-Apo* plants. An *S-Apo* haploid (1*n*, top, $n = 30$), an *S-Apo* diploid progeny of a diploid *S-Apo* parent showing a 2*n* peak (middle; $n = 26$) and a sexual tetraploid progeny of a diploid *S-Apo* parent shows a 4*n* peak (bottom; $n = 90$). The *x* axis is the measure of relative fluorescence and the *y* axis shows the number of nuclei. **f**, Chromosomal view showing 57 heterozygous SNPs (position in Mb) identified in the T_0 *S-Apo* mother plant of line 1. The SNPs labelled in red are those additionally confirmed by PCR.

seed viability depends upon a functional *BBM1* allele from the male parent, consistent with male-specific expression of *BBM1* in zygotes. Next we investigated the embryo phenotypes of *bbm2 bbm3* progeny seeds segregating for the *bbm1* mutation. The *bbm1 bbm2 bbm3* embryos were either arrested early or underwent growth by cell division without any corresponding developmental patterning (Fig. 2a). By contrast, embryos that were heterozygous (*BBM1/bbm1 bbm2/bbm2 bbm3/bbm3*) showed a range of phenotypes—from normal to delayed development (Fig. 2b)—as well as the early arrest or unstructured growth phenotypes observed in the triple mutant (Fig. 2b, Extended Data Fig. 4g). This range of phenotypes might occur by partial rescue from late expression of the female *BBM1* allele. Additionally, *BBM4* (Os04g42570)—a fourth *BBM*-like gene that also shows detectable expression in male gametes (Extended Data Table 1a)—might provide sufficient residual function for partial rescue. The recovery of around 0.7% of the *bbm1 bbm2 bbm3* triple homozygous plants is consistent with the hypothesis of residual *BBM* function being provided by *BBM4* (Extended Data Table 1b).

Together, these data suggest that male-genome-derived expression of *BBM1*—acting redundantly with other *BBM* genes—triggers the embryonic program in the fertilized egg cell. Subsequent activation of expression of the female *BBM1* allele by the male BBM1 results in biallelic expression, with both parental alleles eventually contributing to embryo patterning and organ morphogenesis (Fig. 2c). *BBM*-like genes have been shown to promote regeneration from tissue culture, suggesting that they act as pluripotency factors[20]. Our study supports a model in which the requirement of fertilization to initiate embryogenesis in rice arises from the dependency of the zygote on the male gamete for the expression of pluripotency factors after fertilization. This is in contrast to embryogenesis in vertebrate animals, in which pluripotency factors are maternally provided[7]. As demonstrated below, the requirement for fertilization can therefore be bypassed by driving the expression of one such factor from the female gamete.

Haploid plants are efficient tools for the acceleration of plant breeding, because homozygous isogenic lines can be produced in one generation after chromosome doubling[21]. The expression of *BBM1* in the egg cell initiated parthenogenesis in emasculated flowers (Fig. 1b), but the seeds aborted in the absence of endosperm (Extended Data Fig. 3d). Self-pollinated T_1 progeny from *BBM1*-ee transgenic plants were analysed to determine whether endosperm development by fertilization could produce viable seeds containing parthenogenetically derived haploid embryos. We identified haploids by their small size compared with their diploid siblings, as well as by their sterile flowers owing to defective meiosis[22] (Fig. 2d, e, Extended Data Fig. 5a–d). The ploidy of haploid T_1 plants was confirmed by flow cytometry (Fig. 2f). The haploid induction frequency was 5–10% (T_1 plants) and reached around 29% in homozygous T_2 line 8C—this frequency was maintained through multiple generations (Extended Data Table 2a). Thus, misexpression of

the wild-type *BBM1* gene in the egg cell is sufficient for the production of haploid plants.

Crop yields can be improved markedly by the use of F_1 hybrid plants that exhibit enhanced vigour ('hybrid vigour'). If meiosis and fertilization are bypassed, hybrids could be propagated through seeds without segregation. Asexual propagation through seeds—known as apomixes—is known to occur naturally in more than 400 species, although not in the major crop plants[23,24]. The development of a method to introduce apomixis into crop plants has been described as 'the holy grail of agriculture'[5] as it can enable fixation of hybrid vigour and stabilization of superior heterozygous genotypes in breeding programs[6,25]. A genetic approach called *MiMe*, which eliminates recombination and substitutes mitosis for meiosis (Fig. 3c, d), has been reported in *Arabidopsis*[3] and rice[4]. In *MiMe*, a triple knockout of the meiotic genes *REC8*, *PAIR1* and *OSD1* produces unrecombined diploid male and female gametes. We tested the possibility that *BBM1*-ee-induced parthenogenesis in rice combined with *MiMe* could result in asexual propagation through seeds (Extended Data Fig. 5f). The three rice *MiMe* genes[4] were subject to genome editing by CRISPR–Cas9 in haploid and diploid plants carrying the *BBM1*-ee transgene (Extended Data Fig. 6a). Unlike *BBM1-ee* haploids, the *MiMe* + *BBM1*-ee haploids were fertile (Extended Data Fig. 6c, d) with normal anther development (Fig. 3a), suggesting that meiosis was successfully replaced by mitosis. Self-pollination of *MiMe* plants invariably results in doubling of the chromosome number[22], so the progeny of haploid *MiMe* plants should be diploid (double haploid). However, we obtained haploid progeny from two *MiMe* + *BBM1*-ee (hereafter denoted *S-Apo*, for *Synthetic-Apomictic*) haploid mother plants at frequencies of 26% and 15%, due to parthenogenesis (Fig. 3e, top, Extended Data Table 2b). These haploid induction frequencies were maintained for the next two generations (Extended Data Table 2b). These results show that haploid *S-Apo* plants can be propagated asexually through seeds. Additionally, the sexual T_1 double-haploid ($2n$) progeny from the haploid *S-Apo* plants yielded both diploid and tetraploid plants in the T_2, T_3 and T_4 generations; the former class is expected from the successful asexual propagation of double haploids (Extended Data Table 2b).

For the clonal propagation of diploid *S-Apo* plants, we obtained two fertile transformants with the requisite six null mutations in three *MiMe* genes (Extended Data Fig. 7a, b). Diploid *MiMe* rice plants have been previously shown—despite reduced seed sets—to produce exclusively tetraploid progeny by sexual reproduction and no diploids[4] (Extended Data Fig. 6c). However, we obtained diploids at frequencies of 11% and 29% (Extended Data Table 2b) from the progeny of two diploid *S-Apo* (that is, *MiMe* + *BBM1*-ee) T_0 transformants (Fig. 3b–e, Extended Data Fig. 6e). The rest of the progeny were tetraploid (Fig. 3e). The progeny of a control *MiMe* diploid plant were all determined to be tetraploid (Extended Data Fig. 6b, c). Because T_1 diploid progeny of T_0 diploid *S-Apo* parents are predicted to arise from the parthenogenesis of unreduced female gametes, they should be clonal with the parent and should not exhibit genetic segregation. The T_1 diploids were propagated, and two more generations (T_2 and T_3) of diploid clones were identified by flow cytometry screening.

To demonstrate clonal propagation, we performed whole-genome sequencing on a diploid T_0 *S-Apo* mother plant (line 1), two diploid T_1 progeny, two T_2 diploid progeny of diploid T_1 plants and a control untransformed wild-type plant. Analysis for sequence variants identified 57 heterozygous SNPs in unique sequences distributed over the genome in the T_0 mother plant (Fig. 3f, Supplementary Table 1) that are non-variant in the wild-type plant (see Methods). These 57 SNPs were determined to be heterozygous in all four T_1 and T_2 diploid progeny sequenced. The probability of any single progeny retaining heterozygosity by random segregation for just a subset of 22 unlinked SNPs on different chromosome arms is $P = 2.4 \times 10^{-7}$. The maintenance of heterozygosity at all 57 loci for two generations confirms that the diploid progeny are clonally generated by asexual reproduction. The T_0 *S-Apo* mother (line 1) is additionally biallelic for mutations in the *PAIR1* and *REC8* genes, as were all T_1, T_2 and two T_3 diploid progeny tested (Extended Data Fig. 7a). For SNP validation, 11 randomly selected SNPs were amplified by PCR followed by Sanger sequencing[26] and found to be conserved in the T_0 mother plant and all the T_1, T_2 and T_3 progeny tested (Extended Data Fig. 8). The second diploid *S-Apo* transformant (line 5) is biallelic for all three *MiMe* genes (Extended Data Fig. 7b) and also heterozygous for one of the 11 SNPs confirmed by PCR for line 1. Five T_1 diploid progeny carried an identical set of alleles to the T_0 mother (Extended Data Fig. 7b). The probability that all five progeny would inherit heterozygosity at these four loci by random segregation is $P = 1.8 \times 10^{-5}$. These findings from an independently generated apomictic parent provide further support for successful clonal propagation.

This study demonstrates that asexual propagation without genetic segregation can be engineered in a sexually reproducing plant, and illustrates the feasibility of clonal propagation of hybrids through seeds in rice. Seed formation in this system still requires fertilization to make endosperm (Extended Data Fig. 5f). This endosperm is expected to be hexaploid owing to fertilization of a tetraploid central cell by a diploid sperm cell, whereas the parthenogenetic embryo is diploid, giving a 3:1 ploidy ratio. This deviation from the normal 3:2 ploidy ratio of endosperm to embryo does not appear to be consequential for viability or seed size (Extended Data Fig. 6f, g). Additionally, the clonally propagated seeds preserve the 2:1 maternal-to-paternal genome ratio in endosperm that is required for seed viability[27,28]. To engineer a completely asexual system involving autonomous endosperm formation may not be straightforward in a sexually reproducing crop, and nor is it essential, as many natural apomicts also form seeds with fertilized endosperm[23]. The efficiency of clonal propagation in our system is in part limited by the frequency of parthenogenesis, which could potentially be improved in the future, for example with different promoters. An important factor to consider for future rice-breeding strategies is that genome-wide heterozygosity may be less critical for yield than the incorporation of specific alleles that exhibit full or partial dominance[29,30]. Nevertheless, hybrids can provide a rapid route to higher yields from favourable gene combinations, and have been extensively exploited in maize. Because homologous *BBM*-like and *MiMe* genes are found in other cereal crops, including maize[2,20], the methods described here for asexual propagation through synthetic apomixis should be generally extendible to most cereal crops.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0785-8.

Received: 8 June 2018; Accepted: 29 October 2018;
Published online 12 December 2018.

1. Palovaara, J., de Zeeuw, T. & Weijers, D. Tissue and organ initiation in the plant embryo: a first time for everything. *Annu. Rev. Cell Dev. Biol.* **32**, 47–75 (2016).
2. Horstman, A., Willemsen, V., Boutilier, K. & Heidstra, R. AINTEGUMENTA-LIKE proteins: hubs in a plethora of networks. *Trends Plant Sci.* **19**, 146–157 (2014).
3. d'Erfurth, I. et al. Turning meiosis into mitosis. *PLoS Biol.* **7**, e1000124 (2009).
4. Mieulet, D. et al. Turning rice meiosis into mitosis. *Cell Res.* **26**, 1242–1254 (2016).
5. Sailer, C., Schmid, B. & Grossniklaus, U. Apomixis allows the transgenerational fixation of phenotypes in hybrid plants. *Curr. Biol.* **26**, 331–337 (2016).
6. Vielle Calzada, J.-P., Crane, C. F. & Stelly, D. M. Apomixis—the asexual revolution. *Science* **274**, 1322–1323 (1996).
7. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014).
8. Nodine, M. D. & Bartel, D. P. Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. *Nature* **482**, 94–97 (2012).
9. Autran, D. et al. Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. *Cell* **145**, 707–719 (2011).
10. Del Toro-De León, G., García-Aguilar, M. & Gillmor, C. S. Non-equivalent contributions of maternal and paternal genomes to early plant embryogenesis. *Nature* **514**, 624–627 (2014).
11. Anderson, S. N. et al. The zygotic transition is initiated in unicellular plant zygotes with asymmetric activation of parental genomes. *Dev. Cell* **43**, 349–358. e4, (2017).
12. Kim, S., Soltis, P. S., Wall, K. & Soltis, D. E. Phylogeny and domain evolution in the *APETALA2*-like gene family. *Mol. Biol. Evol.* **23**, 107–120 (2006).

13. Boutilier, K. et al. Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. *Plant Cell* **14**, 1737–1749 (2002).
14. Passarinho, P. et al. BABY BOOM target genes provide diverse entry points into cell proliferation and cell growth pathways. *Plant Mol. Biol.* **68**, 225–237 (2008).
15. Anderson, S. N. et al. Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J.* **76**, 729–741 (2013).
16. Conner, J. A., Mookkan, M., Huo, H., Chae, K. & Ozias-Akins, P. A parthenogenesis gene of apomict origin elicits embryo formation from unfertilized eggs in a sexual plant. *Proc. Natl Acad. Sci. USA* **112**, 11205–11210 (2015).
17. Conner, J. A., Podio, M. & Ozias-Akins, P. Haploid embryo production in rice and maize induced by *PsASGR-BBML* transgenes. *Plant Reprod.* **30**, 41–52 (2017).
18. Steffen, J. G., Kang, I. H., Macfarlane, J. & Drews, G. N. Identification of genes expressed in the *Arabidopsis* female gametophyte. *Plant J.* **51**, 281–292 (2007).
19. Ohnishi, Y., Hoshino, R. & Okamoto, T. Dynamics of male and female chromatin during karyogamy in rice zygotes. *Plant Physiol.* **165**, 1533–1543 (2014).
20. Lowe, K. et al. Morphogenic regulators *Baby boom* and *Wuschel* improve monocot transformation. *Plant Cell* **28**, 1998–2015 (2016).
21. Murovec, J. & Bohanec, B. in *Plant Breeding* (ed. Abdurakhmonov, I. Y.) Ch. 5 (IntechOpen, London, 2012).
22. Cifuentes, M., Rivard, M., Pereira, L., Chelysheva, L. & Mercier, R. Haploid meiosis in *Arabidopsis*: double-strand breaks are formed and repaired but without synapsis and crossovers. *PLoS ONE* **8**, e72431 (2013).
23. Hand, M. L. & Koltunow, A. M. The genetic control of apomixis: asexual seed formation. *Genetics* **197**, 441–450 (2014).
24. Ozias-Akins, P. & van Dijk, P. J. Mendelian genetics of apomixis in plants. *Annu. Rev. Genet.* **41**, 509–537 (2007).
25. Khush, G. S. *Apomixis: exploiting hybrid vigor in rice*. (International Rice Research Institute, 1994).
26. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
27. Lafon-Placette, C. & Köhler, C. Endosperm-based postzygotic hybridization barriers: developmental mechanisms and evolutionary drivers. *Mol. Ecol.* **25**, 2620–2629 (2016).
28. Sekine, D. et al. Dissection of two major components of the post-zygotic hybridization barrier in rice endosperm. *Plant J.* **76**, 792–799 (2013).
29. Hua, J. et al. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl Acad. Sci. USA* **100**, 2574–2579 (2003).
30. Huang, X. et al. Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629–633 (2016).

Acknowledgements We thank U. Vijayraghavan for providing pUN and pUGN vectors; S. Kappu, Z. Liechty and C. Santos-Medellín for advice and help with flow cytometry and sequence analysis; B. Van Bockern for rice transformations; and B. Nguyen and A. Yalda for technical assistance, including genotyping and transplantation. This research was supported by research grants from the National Science Foundation (NSF) (IOS-1547760) and the Innovative Genomics Institute to V.S., NSF grant IOS-1810468 to B.Y., National Institutes of Health grant 1S10OD010786-01to the University of California-Davis Genome Center, and by the United States Department of Agriculture Agricultural Experiment Station (project number CA-D-XXX-6973-H). R.M. acknowledges support from the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS) to the Institut Jean-Pierre Bourgin.

Reviewer information *Nature* thanks T. Dresselhaus and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.S. and I.K. designed the study. I.K. performed experiments and analysed data. D.S. performed analysis of the genome sequences. B.Y. provided pENTR-sgRNA and pUbi-Cas9 vectors for genome editing. V.S. and I. K. wrote the manuscript with input from R.M.

Competing interests The University of California-Davis has filed a patent application on haploid production (PCT/US2017/063249) and a provisional patent application on synthetic apomixis (US62/678,169) arising from this work. INRA has filed a patent application on the use of the MiMe system (EP2208790). The authors declare no other competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0785-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0785-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.S.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plant materials and growth conditions. Rice cultivar Kitaake (*O. sativa* L. subsp. *japonica*) was used for transformations for raising transgenic lines and as a wild-type control. Wild-type, mutant and transgenic seeds were germinated on half-strength Murashige and Skoog's (MS) medium[31] containing 1% sucrose and 0.3% phytagel in a growth chamber for 12 days, under a 16 h light:8 h dark cycle at 28 °C and 80% relative humidity. Seedlings were then transferred to a greenhouse and grown under natural light conditions in Davis, California.

Chemical treatments. Two-week-old wild-type and *BBM1-GR* seedlings were treated with 0.1% ethanol as mock, 10 μM DEX (Sigma-Aldrich), or 10 μM CYC (Sigma-Aldrich) alone or in combination with 10 μM DEX in liquid half-strength MS[31] salts. Seedlings that were of a similar size and had the same number of leaves were selected for the treatments. Individual biological replicates were constructed using similar leaf samples collected from four different plants, collected for RNA isolation after 24 h. CYC treatments were started 30 min before the DEX treatment in the samples that were treated with both reagents.

Plasmid constructs. Full-length coding sequence (CDS) of *BBM1* was amplified from cDNAs made from rice calli using two sets of primers (KitB1F1 5′-CGGATCCATGGCCTCCATCACC-3′, KitB1R1 5′-CCTTCGACCCCA TCCCAT-3′ and KitB1F2 5′-GGATGGGATGGGGTCGAAG-3′, KitB1R2 3′-GGTACCAGACTGAGAACAGAGGC-3′). The two fragments were fused together by an overlap PCR. The overexpression construct (*BBM1*-ox) was created by cloning *BBM1* coding sequence in pUN vector[32] (Extended Data Fig. 1a). To create the BBM1–GR plasmid (Extended Data Fig. 1e), *BBM1* coding sequence without the stop codon was cloned in pUGN vector[32] for translational fusion with rat glucocorticoid receptor[33]. The whole *BBM1* locus, approximately 3kb upstream sequences and the transcribed region until the stop codon were PCR-amplified in two fragments from genomic DNA using two primer pairs: pB1F1 5′-CTCGAGGTCAACACCAACGCCATC-3′, pB1R1 5′- GAAGTCCTCCAGCTTCGGCGC-3′ and pB1F2 5′-TTGATTGTGTTGATG TGCAGAGTGGGG-3′, pB1R2 5′-CTCGAGCGGTGTCGGCAAAACC-3′. The two fragments were joined at a unique restriction enzyme site, NotI, present downstream of the start codon in the sequence. The whole locus was moved to a pCAMBIA1300 vector already containing *Arabidopsis* histone H2B, eGFP and nopaline synthase gene terminator (Extended Data Fig. 2b). The construct for egg-cell-specific expression of BBM1 was made by cloning *BBM1* downstream to *Arabidopsis* DD45 promoter[18] and upstream of the nopaline synthase terminator (Extended Data Fig. 3b) in pCAMBIA1300.

For genome editing of *BBM1*, *BBM2* and *BBM3* genes, single-guide RNA (sgRNA) sequences 5′-GGAGGACTTCCTCGGCATGC-3′, 5′-GTATGCAATATACTCCTGCC -3′ and 5′-GACGGCGGGAGCTGATCCTG -3′, respectively, were designed by using the web tool https://www.genome.arizona.edu/crispr/ as described[34]. The sgRNAs were cloned in pENTR-sgRNA entry vector. The binary vectors for plant transformations (pCRISPR *BBM1* + *BBM3*, pCRISPR *BBM2* + *BBM3* and pCRISPR *BBM1* + *BBM2* + *BBM3*) were constructed by Gateway LR clonase (Life Technologies) recombination with pUbi–Cas9 destination vector as described[35]. Three candidate genes (*OSD1*, Os02g37850; *PAIR1*, Os03g01590 and *REC8*, Os05g50410) for creating *MiMe* mutations in rice were selected as previously described[4] and sgRNAs sequences 5′-GCGCTCGCCGACCCCTCGGG-3′, 5′-GGTGAG GAGGTTGTCGTCGA-3′ and 5′-GTGTGGCGATCGTGTACGAG-3′, respectively, for CRISPR–Cas9-based knockout were designed as described[34]. Vector pCAMBIA2300 *MiMe* CRISPR–Cas9 (Extended Data Fig. 6a) for plant transformations was constructed as described[35], except the resistance marker in the destination vector pUbi–Cas9 was changed to kanamycin (*Neomycin Phosphotransferase II*). pCAMBIA2300 *MiMe* CRISPR–Cas9 was transformed in embryogenic calli derived from *pDD45::BBM1#8c* haploid inducer lines (Extended Data Fig. 3b). Rice transformations were carried out as previously described[36] at the University of California-Davis plant transformation facility. T_0 plants were grown in a greenhouse and screened for *MiMe* mutations. T_1 plants obtained from seeds were subjected to ploidy determination and genotyping for *MiMe* mutations.

Generating *bbm1 bbm2 bbm3* mutants. Rice embryogenic calli were transformed with pCRISPR *BBM1* + *BBM3*, or pCRISPR *BBM2* + *BBM3*. The transformants that carried the *bbm1 bbm3* and *bbm2 bbm3* double mutations generated by genome editing (Extended Data Fig. 4a, b) did not show any phenotypic abnormalities and were fertile. The two double mutants were crossed and selfed; however, no *bbm1 bbm2 bbm3* triple-homozygous plants were recovered in the F_2 generation (Extended Data Fig. 4c). However, plants heterozygous for *BBM1 (bbm1/BBM1)* but homozygous mutant for both *bbm2* and *bbm3* could be recovered, and their progeny were analysed in detail (Extended Data Fig. 4d).

Genotyping. Genotyping of *BBM1*, *BBM2* and *BBM3* mutants was carried out by PCR-amplifying DNA at the mutation site with primers *BBM1* SeqF 5′-TTGATTGTGTTGATGTGC-3′ *BBM1* SeqR 5′-GAGAGACGACCTACTTG GTGAC-3′; *BBM2* SeqF 5′-TAGCTAGCTTGTTAATAGATCATAG-3′, *BBM2* SeqR 5′-TCATATCTCAGTGTGATAGTCTG-3′; and *BBM3* SeqF 5′-ATGCTGCTGCTCCGAGAAG-3′, *BBM3* SeqR 5′-GCTTAGTGCTCCAAACCTCTC-3′. Sanger sequencing[26] of the three PCR amplicons of 464 bp, 262 bp and 547 bp, respectively, for the three genes was carried out at the University of California-Davis DNA-sequencing facility. Because a 1-bp deletion mutation in *BBM1* disrupted an SphI restriction-enzyme site (Extended Data Fig. 4d), all further genotyping of *BBM1* for mutational analysis was performed with restriction digestion of the PCR amplicon with SphI (Extended Data Fig. 4e). For genotyping developing seeds of 5 DAP onwards, endosperm was used for genotyping and embryos were collected for mutant phenotype analysis. DNA fragments at the mutation sites of three *MiMe* genes were PCR-amplified with primers *OSD1* F 5′-TTACTTGGAAGAGGCAGGAGCC -3′, *OSD1* R 5′-ACCTTGACGACTGACGTGATGTC-3′; *PAIR1* F 5′-GTGG TGTGGTGTGTTCAGGAG-3′, *PAIR1* R 5′-TGGAATCCCCAA TCAGTAAGGCAC-3′; and *REC8* F 5′-GCACTAAGGCTCTCCGGAATTCTC-3′, *REC8* R 5′-AATGGATCAAGGAGGAGGCACC-3′. PCR amplicons of 364 bp, 344 bp and 326 bp—for *OSD1*, *PAIR1* and *REC8*, respectively—were subjected to Sanger sequencing[26] for mutation analysis.

Emasculation, crosses and pollinations. Flowers from *BBM1*-ee T_0 transgenic rice lines were emasculated around the anthesis stage, bagged and allowed to grow for another nine days after emasculation. Carpels were collected and fixed for analysis in formaldehyde (10%)–acetic acid (5%)–ethanol (50%). A translational fusion consisting of the *BBM1* genomic locus to GFP (BBM1–GFP; Extended Data Fig. 2b) was introduced into the inbred *japonica* (Kitaake) cultivar by transformation. Plants hemizygous for the BBM1–GFP transgene were then reciprocally crossed to wild-type plants. Flowers from wild-type or BBM1–GFP transgenic plants were hand-pollinated around the anthesis stage and carpels were collected 2.5 and 6.5 HAP.

For phenotypic analysis of mutant embryos, self-pollinated flowers from mutant plants were scored for anthesis, and collected 5 or 10 DAP. For crosses of *bbm1 bbm3* and *bbm2 bbm3* plants, only T_2 progeny plants in which the CRISPR–Cas9 transgene had already segregated out were used as parents. For all crosses of *bbm1 bbm3* with *bbm2 bbm3* plants, and for the reciprocal crosses between *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* and *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* plants, panicles used as females were emasculated and bagged with pollen donor panicles. The bags were gently finger-tapped (twice a day) for the next two days. Male panicles were removed, and female panicles were left bagged to make seeds. F_1 seeds were collected four weeks after pollination.

Immunohistochemistry and toluidine blue staining. Owing to the difficulty of imaging GFP fluorescence in early rice zygotes through the carpel tissue, we used antibodies against GFP to detect zygote expression in sectioned rice carpels. Collected carpels were fixed in formaldehyde (10%)–acetic acid (5%)–ethanol (50%). Tissue embedding and sectioning was performed as described previously[37]. Immunohistochemistry was carried out using standard protocols[38], except an antigen-retrieval step was also included. Antigen retrieval was performed by microwaving the slides in 10 mM sodium citrate buffer (pH 6.0) for 10 min. Rabbit anti-GFP antibody ab6556 (Abcam) was used as the primary antibody and goat anti-rabbit alkaline phosphatase conjugate A9919 (Sigma) was used as the secondary antibody. For toluidine blue staining, after rehydration, sections crosslinked to glass slides were stained with 0.01% toluidine blue for 30 s.

Flow cytometry. Nuclei for fluorescence-activated cell sorting (FACS) analysis were isolated by a leaf-chopping method described previously[39]. The isolated nuclei were stained with propidium iodide at 40 μg ml^{-1} in Galbraith's buffer. FACS analysis and DNA-content estimation was carried out using a Becton Dickinson FACScan system using standard protocols[40,41]. DNA histograms were gated out for the initial debris.

Alexander staining of pollen grains. Stamens were collected just before anthesis. Anthers were put on a glass slide in a drop of Alexander's stain containing 40 μl of glacial acetic acid per millilitre of stain[42]. Anthers were covered with a coverslip and slides were heated at 55 °C on a heating block, until the visible staining of pollen was observed.

Library preparation and sequencing. PCR-free DNA libraries were prepared from a wild-type Kitaake control plant, the T_0 *S-Apo* line 1 mother plant, two T_1 and two T_2 progeny clones from *S-Apo* line 1 with 500 ng of input DNA, using NuGEN Celero DNA-Seq kit, following the manufacturer's instructions. Samples were multiplexed and six libraries per lane were run on Illumina HiSeq platforms at the University of California-Davis Genome Center.

Whole-genome DNA sequencing and statistical analysis. Adaptor removal and quality trimming of 150-bp paired-end reads was performed using Trimmomatic

0.38[43] resulting in 13–16 gigabases of sequence for each library. The reads were aligned to the *O. sativa* reference genome (Nipponbare, Release 7.0)[44] using *bwa mem*[45]. To discover variants that were heterozygous in the T_0 mother plant (line 1), the variant finder GATK4.0 HaplotypeCaller was used in single-sample mode[46] and selecting only for SNPs. Repeated elements of the genome were masked from analysis using annotated repeats from http://www.phytozome.org (Osativa_323_v7.0.repeatmasked_assembly_v7.0.gff3). Variants were retained for analysis after filtering on the basis of mapping quality (MQ = 60), QualByDepth (QD >2), StrandOddsRatio (SOR <1.8), unfiltered read depth ($10 \leq DP \leq 40$) and fraction of the alternate allele ($0.4 \leq DP \leq 0.6$), with the expectation that a truly heterozygous locus should show roughly equal numbers of read counts for each allele. To increase certainty that the set of loci included only true heterozygous SNPs, loci which were called heterozygous in the wild-type sample were also discarded. This strategy guards against instances in which incorrect read-mapping over multi-copy regions lead to spurious designation of loci as heterozygous, even though it is likely that we also discarded true heterozygous loci in the process. A final list of 60 high-quality heterozygous SNPs at 57 loci were analysed for segregation in the four progeny clones (T_1 clone A, T_1 clone B, T_2 clone 7 and T_2 clone 21). All SNPs were called heterozygous by HaplotypeCaller in all the progeny samples (Supplementary Table 1).

For statistical analysis of genetic ratios: Either a chi-square goodness-of-fit test or a two-tailed Fisher's exact test was carried out wherever applicable, and the result specified in the legend of the relevant figure or table.

RT–PCR and RT–qPCR. All the cDNAs were synthesized using the iScript cDNA synthesis kit (BioRad) according to the manufacturer's instructions. RT–PCRs were performed with MyTaq Red Mix (Bioline) and RT–qPCRs with iTaq universal SYBR Green supermix (BioRad) using CFX96 Touch real-time PCR system (BioRad). *UBIQUITIN5* (Os03g13170) was used as the internal control and fold changes in the relative abundance of transcripts were calculated as described previously[47]. For RT–qPCR, amplifications for each gene were performed in two biological replicates, and each biological replicate was repeated in three technical replicates for each sample. For *BBM1*, *BBM1* RT F 5′-TACTACCTTTCCGAGGGTTCG-3′ was used in combination with B1RNAi R 5′-GATATC CCAGACTGAGAACAGAGGC -3′ to detect endogenous transcript and with GR RT R 5′-TCTTGTGAGACTCCTGCAGTG-3′ to detect *BBM1-GR* transgenic transcript in RT–qPCR experiments. *BBM1*intronF 5′-GTGGCAGGAAACAAGGATCTG-3′ with B1RNAi R which spanned an intron was used in RT–PCR experiments. For other genes tested in this study, the following primer combinations were used: *LEC1A* F 5′-GACAGGTGATCGAGCTCGTC-3′, *LEC1A* R 5′-CTCTTTCGATGAAACGGTGGC-3′; *LEC1B* F 5′-ACAGC AGCAGAATGGCGATC-3′, *LEC1B* R 5′-CTCATCGATCACTACCTGAACG-3′; *GE* F 5′-CAGGAGCACAAGGCGAAGCG-3′, *GE* R 5′-CTTCGCCTGGATCT CCGGGTG-3′; *OSH1* F 5′-GAGATTGATGCACATGGTGTG-3′, *OSH1* R 5′-CGAGGGGTAAGGCCATTTGTA-3′; and *UBIQUITIN5* F 5′-ACCACTTCGA CCGCCACT-3′, *UBIQUITIN5* R 5′-ACGCCTAAGCCTGCTGGTT-3′.

SNP analysis. Detection of SNPs in *BBM1* transcripts from hybrid zygotes was performed by PCR of 2.5 HAP zygote cDNAs from reciprocally crossed rice *japonica* cultivar Kitaake and *indica* cultivar IR50, as described previously[11]. Primers B1RNAi F 5′-CCTCGAGCAACTATGGTTCGCAGC-3′ and B1RNAi R, which amplified a gene-specific fragment of about 600 bp of *BBM1*, contains 5 SNPs between Kitaake and IR50 (Extended Data Fig. 2a). The PCR amplicons were Sanger-sequenced[26] and chromatograms were analysed for SNPs. For detection of heterozygous SNPs present in the *S-Apo* mother plants and their progeny, 50 ng of input DNA was used for each PCR reaction. Sanger-sequenced[26] PCR chromatograms were analysed for the presence of SNPs. The primers for 11 SNPs analysed are: 1 Chr2 F 5′-TGGGTGCCA CGTTATCTAGG-3′, 1 Chr2 R 5′-GGATTTGGCTACCCTCAAGCT-3′; 2 Chr2 F 5′-GAATGGGCAACTAACAACCGTG-3′, 2 Chr2 R 5′-ACCGTG GAAAGGAACAGCTG-3′; 1 Chr3 F 5′-TGCTGAAGGTGACGTTGATCTG-3′, 1 Chr3 R 5′-CGACGCCAACGAGAAGGA-3′; 2 Chr3 F 5′-GCTCCAGTGCTA GAGAGACATC-3′, 2 Chr3 R 5′-AGCCACCCAGTAACCGTTG-3′; Chr4 F 5′-GATTGGCAAACCAGCTACTGC-3′, Chr4 R 5′-CTGATGGCAAG CTGTTGGC-3′; Chr5 F 5′-ATGATCTGCTGCTTGTTTCAATGC-3′, Chr5 R 5′-TATCCTTCAAGCACCACTGCC-3′; Chr6 F 5′-ACTAATGGGACCACT TGACAGC-3′, Chr6 R 5′-TCAGCCTGAGATGGCTTGG-3′; Chr8 F 5′-CAGACTGTGGGACGCTACATG-3′, Chr8 R 5′-AGAAGATCT GGGCAGCAGTC-3′; Chr9 F 5′-GCTGCACCTGTTAGCTATGTGA-3′, Chr9 R 5′-AGCATCCCAAAAGCACACATG-3′; Chr10 F 5′-TCAGCAGCCTAAGGTT GAAGG-3′, Chr10 R 5′-CTGCTGCTGCTTCATGATCAC-3′; and Chr11 F 5′-GCAGGAACTATTGCCTCTCATGA-3′, Chr11 R 5′-TCAGTCTCATAGCGCA CCAC-3′.

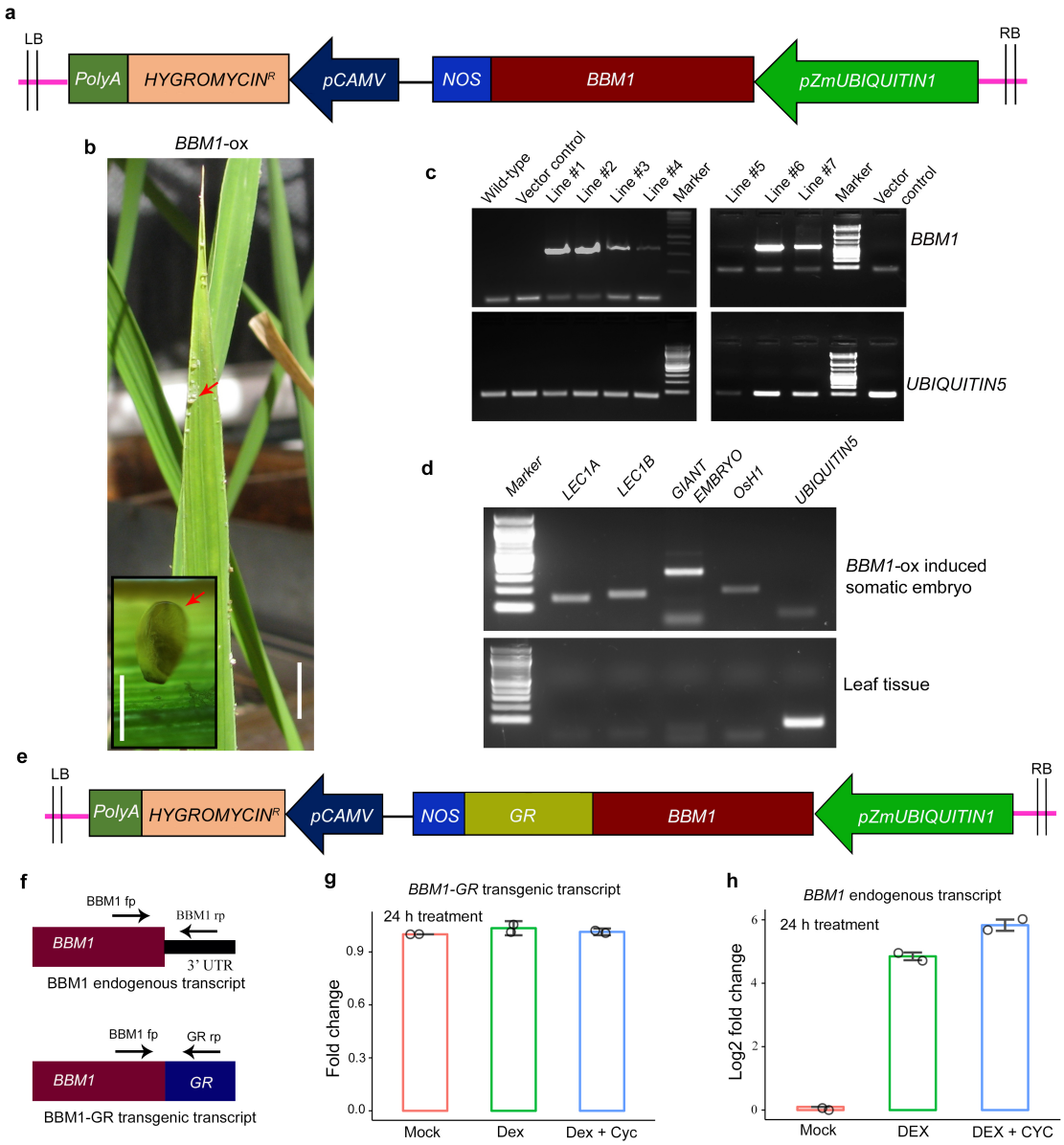
Code availability. Codes for the different analyses are available for non-commercial use from the corresponding author upon request.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

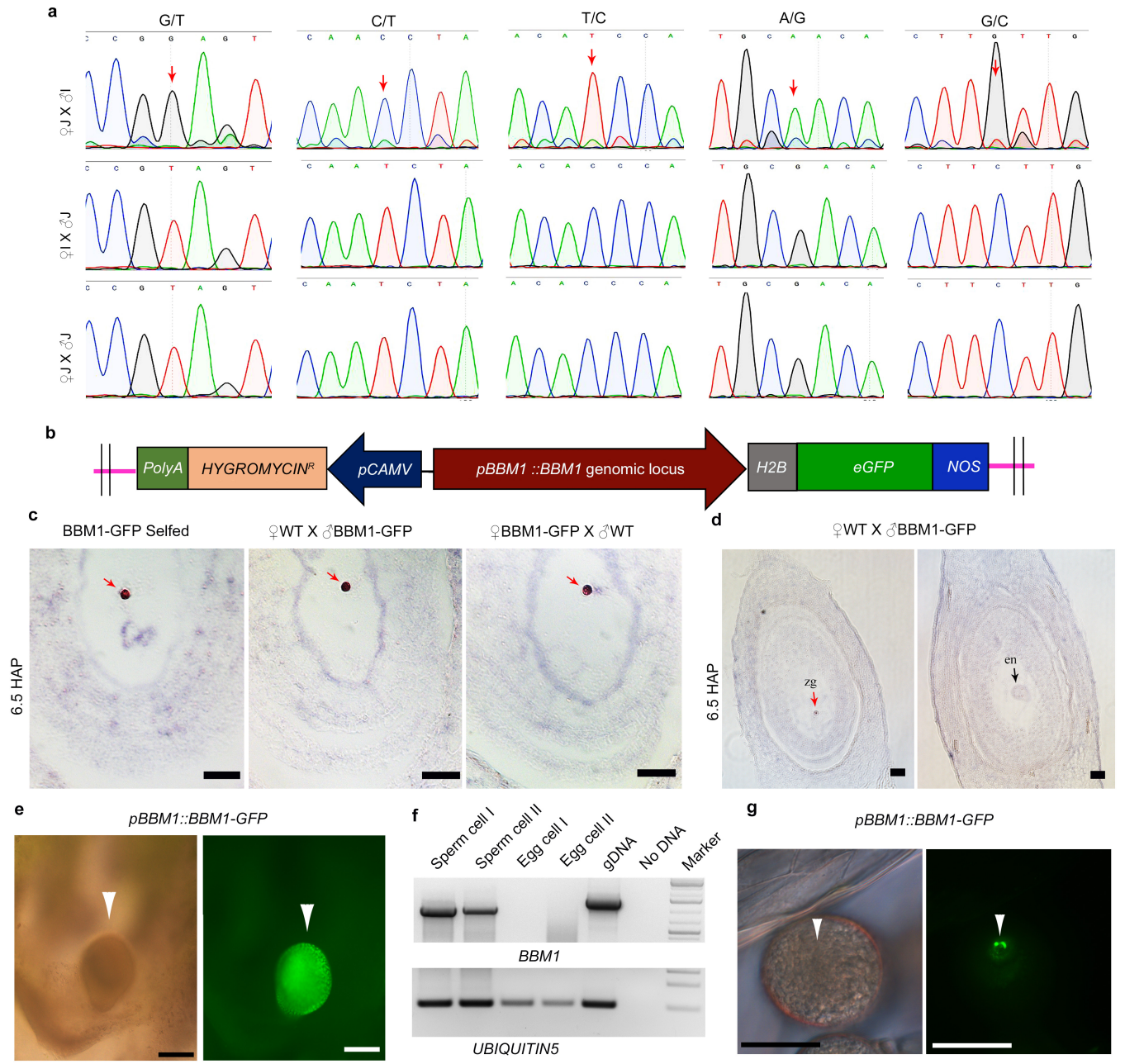
Data availability

Whole-genome DNA sequencing data for *S-Apo* line 1 mother plant, the four progeny clones from two generations, and the Kitaake wild-type control are available from National Center for Biotechnology Information (NCBI) BioProject number PRJNA496208. RNA sequencing data from previously published datasets[11,15] are available from the NCBI Short Read Archive as Project SRP119200 and from the NCBI Gene Expression Omnibus under accession number GSE50777.

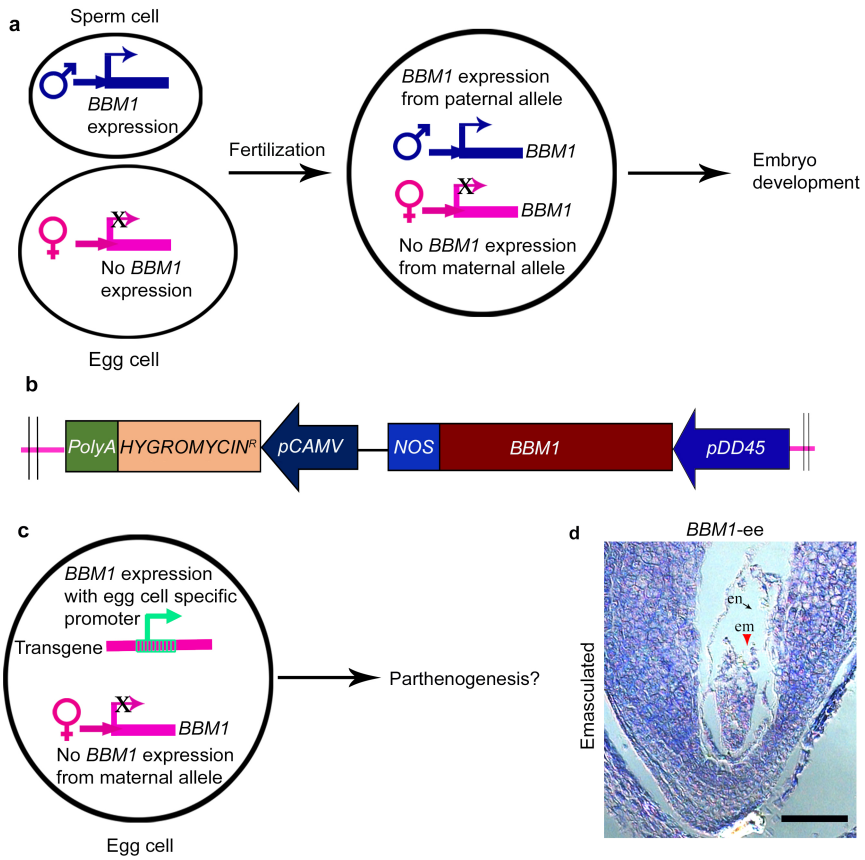
31. Murashige, T. & Skoog, F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* **15**, 473–497 (1962).
32. Khanday, I., Yadav, S. R. & Vijayraghavan, U. Rice *LHS1/OsMADS1* controls floret meristem specification by coordinated regulation of transcription factors and hormone signaling pathways. *Plant Physiol.* **161**, 1970–1983 (2013).
33. Aoyama, T. & Chua, N. H. A glucocorticoid-mediated transcriptional induction system in transgenic plants. *Plant J.* **11**, 605–612 (1997).
34. Xie, K., Zhang, J. & Yang, Y. Genome-wide prediction of highly specific guide RNA spacers for CRISPR–Cas9-mediated genome editing in model plants and major crops. *Mol. Plant* **7**, 923–926 (2014).
35. Zhou, H., Liu, B., Weeks, D. P., Spalding, M. H. & Yang, B. Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Res.* **42**, 10903–10914 (2014).
36. Hiei, Y. & Komari, T. Agrobacterium-mediated transformation of rice using immature embryos or calli induced from mature seed. *Nat. Protoc.* **3**, 824–834 (2008).
37. Javelle, M., Marco, C. F. & Timmermans, M. In situ hybridization for the precise localization of transcripts in plants. *J. Vis. Exp.* **57**, e3328 (2011).
38. Sessions, A. Immunohistochemistry on sections of plant tissues using alkaline-phosphatase-coupled secondary antibody. *Cold Spring Harb. Protoc.* https://doi.org/0.1101/pdb.prot4946 (2008).
39. Galbraith, D. W. et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**, 1049–1051 (1983).
40. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
41. Cousin, A., Heel, K., Cowling, W. A. & Nelson, M. N. An efficient high-throughput flow cytometric method for estimating DNA ploidy level in plants. *Cytometry A* **75A**, 1015–1019 (2009).
42. Alexander, M. P. Differential staining of aborted and nonaborted pollen. *Stain Technol.* **44**, 117–122 (1969).
43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
44. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y.)* **6**, 4 (2013).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1–11.10.33 (2013).
47. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).



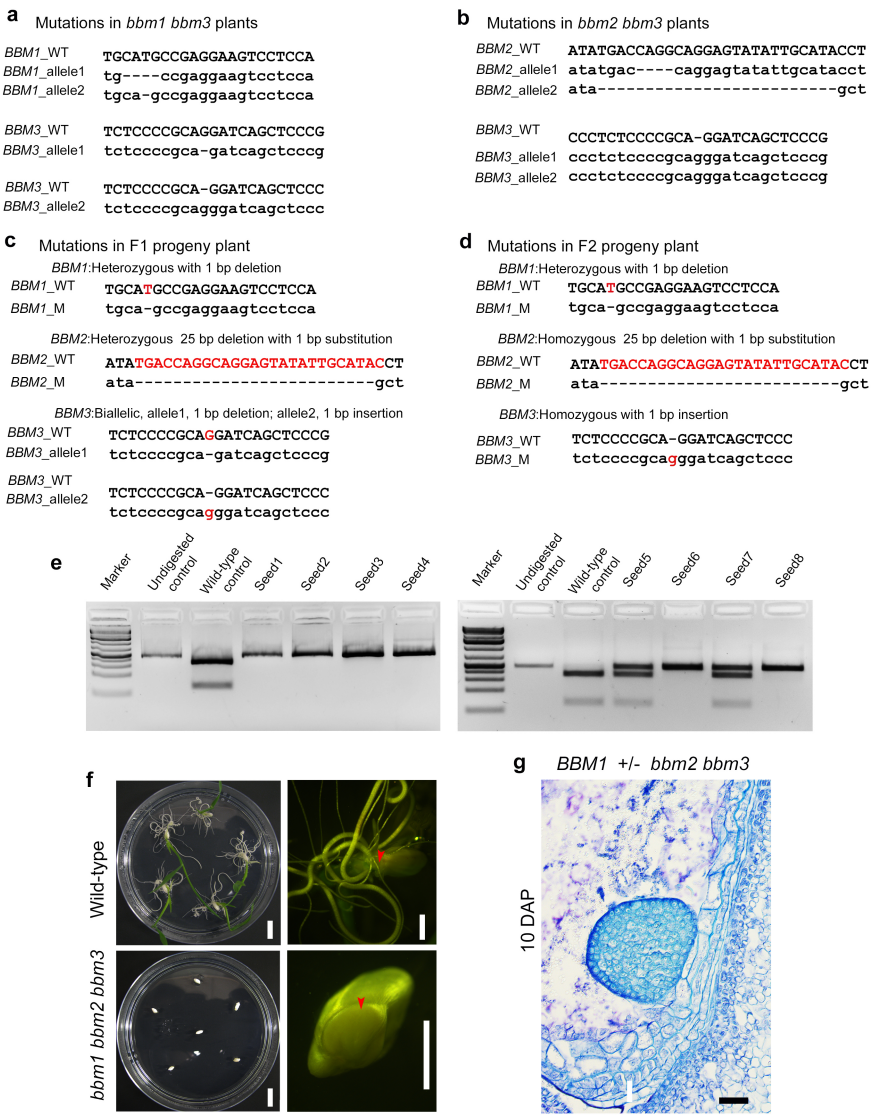
Extended Data Fig. 1 | *BBM1*-induced somatic embryogenesis and auto-activation. a, Schematic of binary construct between T-DNA borders used for ectopic expression (*BBM1*-ox). **b**, Somatic embryo-like structures induced by *BBM1* ectopic expression in rice leaves ($n = 14/20$ transgenic lines). Scale bar, 1 cm. Inset, magnified view of a somatic embryo; scale bar, 0.5 mm. Fourteen of the twenty transgenic plants raised showed the development of such embryo-like structures observed on adult seedlings from the fourth leaf onwards. **c**, Confirmation by RT–PCR of ectopic *BBM1* expression in leaf tissues of transgenic lines. *BBM1* is not expressed in wild-type leaves ($n = 2$ independent replicates). **d**, RT–PCR of embryo marker genes to confirm the embryo identity of somatic embryos induced by *BBM1* overexpression. *OsH1*, *O. sativa HOMEOBOX1*; *LEC1*, *LEAFY COTYLEDON1* ($n = 2$ independent biological replicates). **e**, Schematic of plasmid construct for DEX-inducible BBM1–GR expression system. **f**, Schematic showing primer combinations to distinguish between endogenous *BBM1* and *BBM1-GR* fusion transcripts. **g**, RT–qPCR for fold changes in *BBM1-GR* fusion transcript in samples treated for 24 h with the indicated reagents, showing essentially no differences between treatments. $n = 2$ independent biological replicates (see Methods), data are mean ± s.e.m. and each data point represents the average fold change from three replicates. **h**, Autoactivation of *BBM1* in samples treated with DEX for 24 h, detected by RT–qPCR. $n = 2$ independent biological replicates (see Methods), data are mean ± s.e.m. and each data point represents the average fold change (measured as $\log_2$(change in expression)) from three replicates.



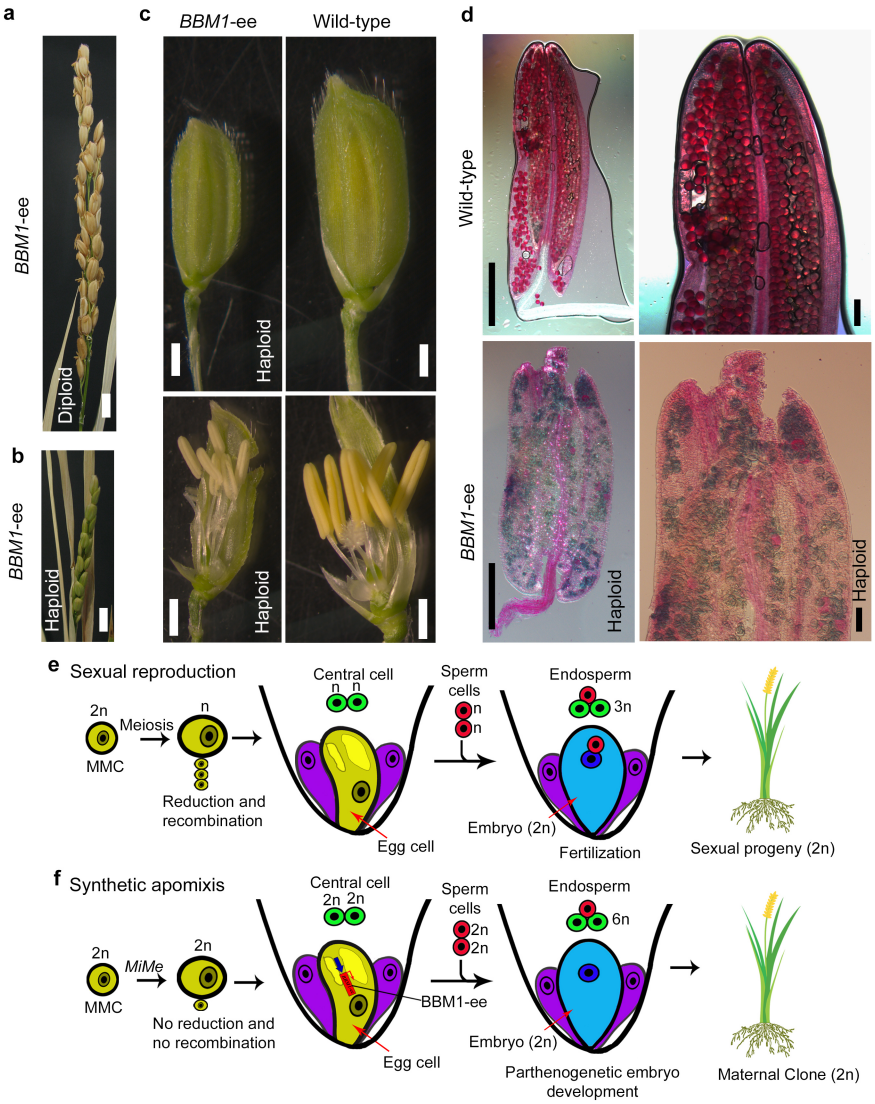
Extended Data Fig. 2 | *BBM1* expression in zygotes and gametes. **a**, Five SNPs sequenced after RT–PCR amplification (red arrows), showing expression only from the male allele in hybrid (J, *japonica*; I, *indica*) 2.5 HAP zygotes ($n = 2$ biological replicates). **b**, Schematic of the *BBM1-GFP* binary construct. **c**, Immunohistochemistry showing expression from both male and female *BBM1* alleles in isogenic 6.5 HAP zygote nuclei ($n = 20$), as compared to male-specific expression at 2.5 HAP (Fig. 1a). Scale bars, 25 μm. **d**, Holistic view of a 6.5 HAP embryo sac showing BBM1–GFP expression in the zygote nucleus (left), while in the same embryo sac expression is not detected in the dividing endosperm (right). zg, zygote. $n = 20$. Scale bar 100 μm. **e**, BBM1–GFP expression in globular-stage rice embryos (white arrowhead, $n = 30$). Differential interference contrast image (left); fluorescence image (right panel). Scale bars, 200 μm. **f**, RT–PCR showing *BBM1* expression in sperm cells; however, the transcript is not detected in egg cells ($n = 2$ independent biological replicates). Primers used for detecting *BBM1* transcript span an intron (see Methods). **g**, BBM1–GFP expression in sperm cells (white arrowhead points to sperm nuclei, $n = 20$). Differential interference contrast image (left) and fluorescent image (right) of a germinating pollen grain showing BBM1–GFP expression in the two sperm cell nuclei.



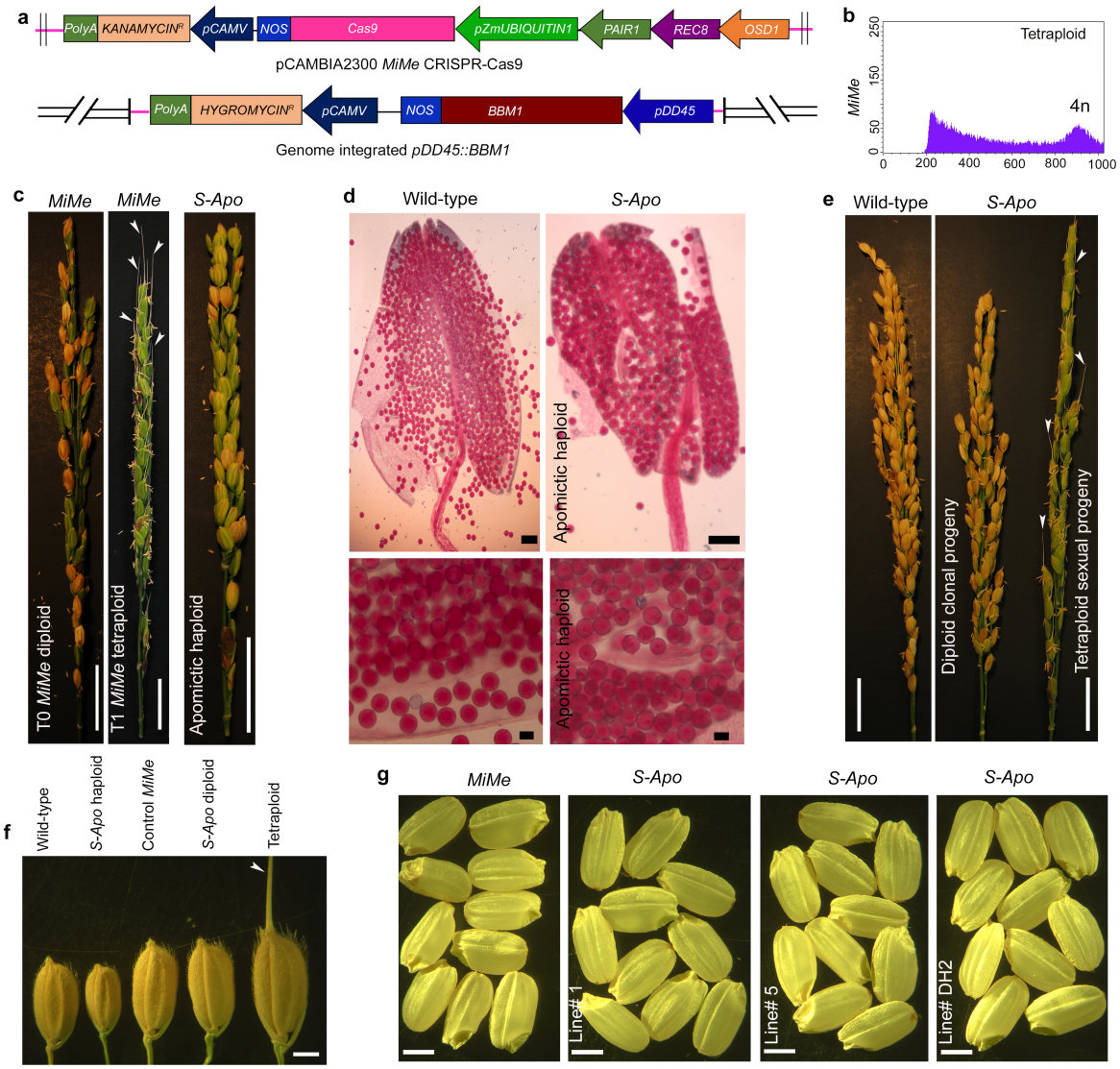
Extended Data Fig. 3 | Parthenogenesis induction by expression of *BBM1* in the egg cell. a, Schematic showing wild-type expression pattern of *BBM1*. **b**, Sketch of T-DNA region of the binary vector used for *BBM1* expression in the egg cell. **c**, Schematic representation of the hypothesis that the expression of *BBM1* in the egg cell can induce parthenogenesis. **d**, A degenerating parthenogenetic embryo (*BBM1*-ee) at 9 days after emasculation (red arrowhead). No endosperm development (black arrow) is observed in emasculated carpels, leading to the abortion of embryos ($n = 12/98$). Scale bar, 100 μm.



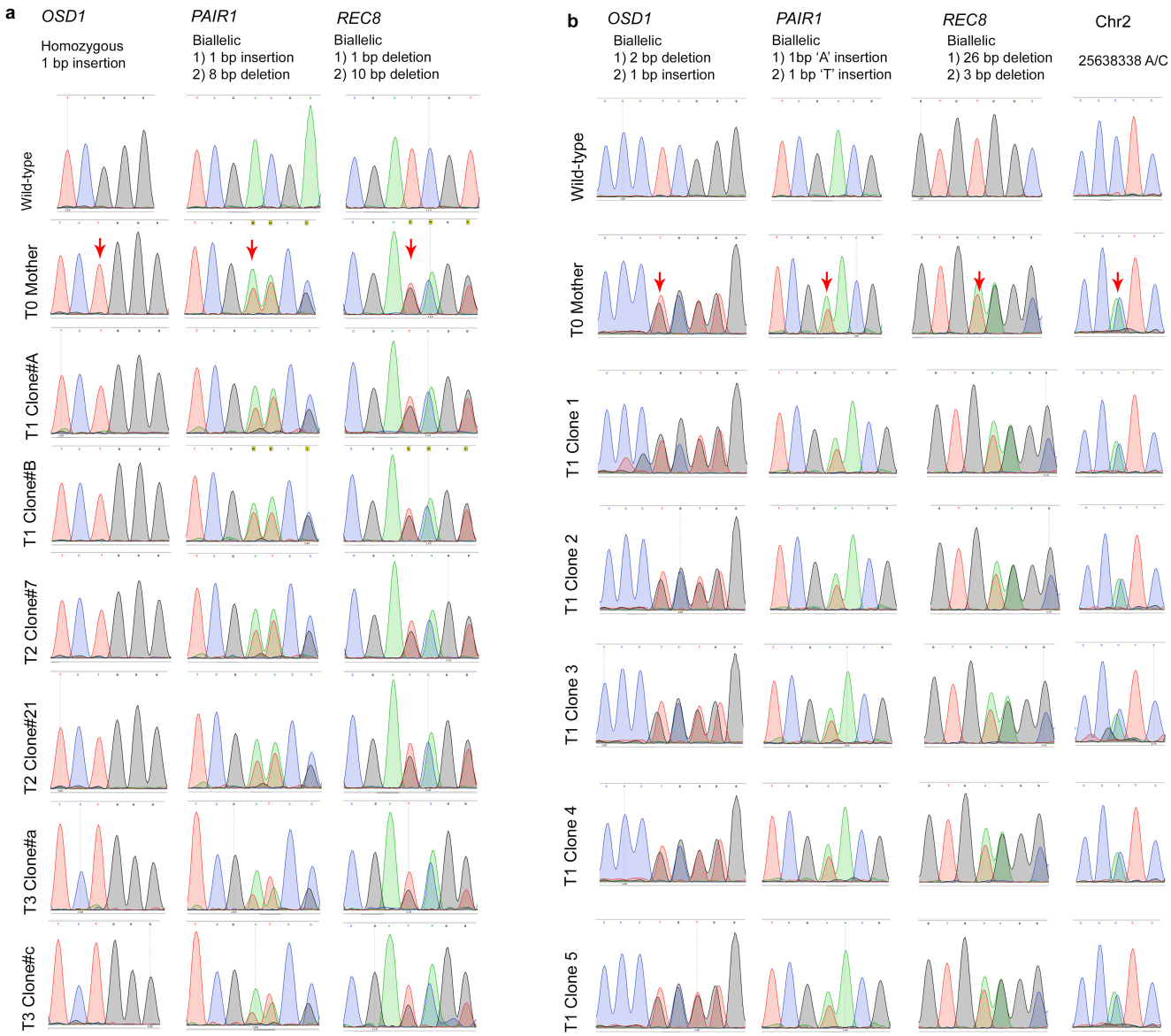
Extended Data Fig. 4 | CRISPR–Cas9 edited mutations in *BBM1*, *BBM2* and *BBM3* in rice. a, DNA sequences of mutations in *bbm1/bbm1 bbm3/bbm3* plants. **b**, DNA sequences of mutations in *bbm2/bbm2 bbm3/bbm3* plants. **a** and **b** were chosen as parents for crosses to generate the *bbm1 bbm2 bbm3* triple homozygous mutants shown in **c** and **d**. **c**, Mutations in the F_1 progeny plant. It is heterozygous for *BBM1* and *BBM2*, and biallelic for *BBM3*. **d**, Mutations in the F_2 progeny plant used for genetic analysis. The plant is heterozygous for *BBM1* with a 1-bp deletion. The *BBM2* locus has a homozygous 25-bp deletion and 1-bp substitution, and the *BBM3* locus is a homozygous mutant with 1-bp insertion. **e**, Genotyping of non-germinating seeds ($n = 8$). The 1-bp deletion mutation in *BBM1* results in disruption of an SphI restriction site. **f**, Seed lethality in *bbm1 bbm2 bbm3* triple homozygous plants. Top, germinating one-week-old wild-type seeds ($n = 30$). Scale bars, 1 cm. A magnified view is shown on the right. Bottom, non-germinating seeds of *bbm1 bbm2 bbm3* triple homozygous plants ($n = 70$). A zoomed-in image of a non-germinating *bbm1 bbm2 bbm3* seed, one week after plating, is shown on the bottom right. No seedling emerged from the embryo site (red arrowhead). **g**, Additional image of a *BBM1/bbm1* heterozygous *bbm2/bbm2 bbm3/bbm3* homozygous 10 DAP embryo ($n = 3/53$) showing no organ formation, similar to triple homozygote phenotype (see Fig. 2a). Scale bar, 100 μm.



Extended Data Fig. 5 | Haploid induction and synthetic apomixis. Haploids shown are derived from *BBM1*-ee diploids by parthenogenesis. **a**, A control diploid sibling panicle with fertile florets ($n = 442$ plants). Scale bar, 1 cm. **b**, A haploid panicle with infertile florets ($n = 113$ plants). Scale bar, 1 cm. **c**, Differences in floret and floral organ sizes between haploid and control diploid. Left, *BBM1*-ee haploid; right, wild-type control ($n = 20$). Scale bars, 1 mm. **d**, Pollen viability in haploids as assessed by Alexander staining. Top, control wild-type anther with viable pollen ($n = 10$). Bottom, *BBM1*-ee haploid anther with non-viable pollen ($n = 20$). Scale bars, 0.5 mm (left) and 200 μm (right). **e**, **f**, Sexual reproduction compared with asexual reproduction through seed (synthetic apomixis). **e**, Schematic representation of sexual reproduction. Gametes form by meiotic recombination and division; fertilization and gamete fusion give rise to diploid progeny. **f**, Synthetic apomixis. *MiMe* omits meiosis and gives an unrecombined and unreduced ($2n$) egg cell. The $2n$ egg cell is converted parthenogenetically into a clonal embryo by *BBM1*-ee. The endosperm forms in both pathways by fertilization of central cell (homodiploid in wild type, tetraploid in synthetic apomicts) by a sperm cell (haploid in wild type, diploid in synthetic apomicts). The maternal:paternal genome ratio of 2:1 is maintained in the endosperm in both the pathways, ensuring normal seed development.



Extended Data Fig. 6 | Asexual propagation through seed in rice. **a**, Top, schematic of the CRISPR–Cas9 plasmid construct used for genome editing of the three *MiMe* rice genes. Bottom, schematic of genome-integrated *pDD45::BBM1* in the *BBM1*-ee plants. **b**, DNA histogram of flow cytometric peak showing 4*n* ploidy in T_1 progeny ($n = 33/33$ tested) of a control T_0 *MiMe* plant. **c**, Left, panicle of a control T_0 diploid *MiMe* plant with fertile seeds. Middle, a tetraploid T_1 *MiMe* panicle, exhibiting complete infertility; that is, no seed filling, and larger flowers (note scale bars), with awns (white arrowhead). Awns are normally suppressed in most *japonica* rice cultivars including Kitaake. All T_1 *MiMe* progeny ($n = 139$) were scored for the phenotype of complete infertility and presence of awns, including 33 plants that were additionally confirmed in **b** by flow cytometry. Right, panicle of an *S-Apo* haploid plant showing fertile seeds ($n = 45$). Scale bars, 2 cm. **d**, Wild-type and *S-Apo* haploid anthers, showing viable pollen ($n = 15$). Scale bars, 0.2 mm (top) and 100 μm (bottom). **e**, Comparison of panicles from wild type (left), with diploid clonal progeny (57/381) and sexual tetraploid progeny ($n = 324/381$) from a diploid *S-Apo* plant (right). The white arrowheads show awns in tetraploid. Scale bars, 2 cm. **f**, Size comparison of progeny seeds from control wild type, a synthetic *S-Apo* haploid, a control *MiMe*, a synthetic *S-Apo* diploid clone, and an infrequent (3%) filled seed produced by the sexual tetraploid progeny of an *S-Apo* diploid ($n = 100$ for each genotype). Scale bar, 2 mm. **g**, Comparison of seed size between control *MiMe*, diploid *S-Apo* line 1, diploid *S-Apo* line 5 and double-haploid *S-Apo* line DH2 ($n = 100$ for each transgenic line). No noticeable variation in seed size is observed. Scale bars, 2 mm.



Extended Data Fig. 7 | *MiMe* mutations and confirmation of clonal progeny from *S-Apo* plants. a, Sequence chromatograms at mutation sites of *MiMe* genes in wild-type, T_0 diploid *S-Apo* mother plant and two diploid progeny from each of T_1, T_2 and T_3 generations of *S-Apo* line 1 ($n = 7$). Red arrows point to mutation sites. *PAIR1* and *REC8* are biallelic whereas *OSD1* is homozygous. **b**, Sequences of the T_0 *S-Apo* mother plant and five T_1 *S-Apo* diploid progeny at *MiMe* mutation sites and one heterozygous SNP in apomixis line 5 ($n = 6$). Red arrows show the mutation sites or SNP. All three *MiMe* mutations—*OSD1*, *PAIR1* and *REC8*—are biallelic. All progeny across different generations in both the *S-Apo* lines have same mutations as the T_0 mother plants, indicating absence of segregation and thus clonal propagation.

Extended Data Fig. 8 | Confirmation of SNPs by PCR. Sequence chromatograms of 11 SNPs are shown for wild-type, T_0 diploid *S-Apo* mother plant and two diploid *S-Apo* progeny from each of the T_1, T_2 and T_3 generations for line 1 ($n = 7$). All the 11 SNPs were found to be present in the T_0 mother plant and all the progeny across different generations, confirming that there is no segregation; thus clonal propagation. The red arrows show the location of the SNP. Chr, chromosome; the numbers indicate the position on the respective chromosome.

Extended Data Table 1 | Functional characterization of *BBM* genes in rice

a

Locus/Isoform	Gene	EC	SpC	Z2.5	Z5	Z9	Z2.5_JxI	Z2.5_IxJ
LOC_Os11g19060.1	*BBM1*	0	2.03	1.45	4.5	5.5	0.51	0.38
LOC_Os02g40070.1	*BBM2*	0	0	0.63	1.75	1.56	0.2	0.28
LOC_Os01g67410.1	*BBM3*	0	0.54	1.01	2.04	0.45	0.15	0.14
LOC_Os04g42570.1	*BBM4*	0	0.2	0	0	0	0.295	0

b

Number of seeds tested	Seeds germinated	Seeds that did not germinate	Percentage non-germinated	Genotypes of germinated seedlings: *BBM1/BBM1: BBM1/bbm1: bbm1/bbm1*
297	191	106	35.6	81:108:2*

c

Female *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* X Male *bbm1/BBM1 bbm2/bbm2 bbm3/bbm3*					
No. of Seeds	**Seeds germinated**	**Wild-type for *BBM1***	**Heterozygous for *BBM1***	**Seeds did not germinate**	**Non -germinating seeds genotyped**
149	121	59	62	28	23, all heterozygous**

d

Female *bbm1/BBM1 bbm2/bbm2 bbm3/bbm3* X Male *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3*					
No. of Seeds	**Seeds germinated**	**Wild-type for *BBM1***	**Heterozygous for *BBM1***	**Seeds did not germinate**	**Non-germinating seeds genotyped**
67	67	35	32	0	0

a, Expression of four *BBM*-like genes in rice gametes and zygotes from previous studies[11,15] presented as reads per million averaged from three replicates. Z2.5, Z5 and Z9 columns are from isogenic *japonica* zygotes at 2.5, 5 and 9 HAP, respectively. J×I and I×J columns are hybrid zygotes from crosses, the female parent is listed first. EC, egg cell; I, *indica*; J, *japonica*; SpC, sperm cell; Z, zygote.
b, Summary of seed viability in progeny of *BBM1/ bbm1 bbm2/bbm2 bbm3/bbm3* mutant plants. A loss of viability was observed, as around 36% (106/297) of seeds fail to germinate. Of the germinated seedlings, only 1% (2/191) were triple homozygotes, instead of the expected 25% if there is no effect of genotype on viability. **c**, **d**, Dependence of seed viability on paternal allele transmission of *BBM1*. **c**, When the *bbm1* allele is transmitted by the male parent, around 27% of the genotyped heterozygotes fail to germinate (23/(23 + 62)), despite a functional *BBM1* allele inherited from the female parent. **d**, All seeds germinate when the mutant *bbm1* allele is transmitted by the female parent ($n = 67$).
*The chi-square value for goodness-of-fit between the expected Mendelian 1:2:1 ratio and the observed data is 68.623; the corresponding right-tail P value is 1.714×10^{-15}.
**The two-tailed Fisher's exact test P value is 0.0001, for the genotyped non-germinating seeds to contain all heterozygotes and no wild types.

Extended Data Table 2 | Haploid induction and clonal propagation in rice

a

Transgenic line#	Generation	Number of plants tested	Number of haploids	% Haploid induction
1	T1	28	2	7.1
3	T1	25	2	8
4	T1	32	3	9.3
5	T1	34	2	5.8
8	T1	57	6	10.5
10	T1	27	2	7.4
11	T1	31	2	6.4
Haploid induction in homozygous T1 progeny line#8c				
8c	T2	185	54	29.2
	T3	40	13	32.5
	T4	33	9	27.2
	T5	18	5	27.7
	T6	21	6	28.5
	T7	24	7	29.1

b

Frequencies of haploid asexual progeny from haploid *S-Apo* plants					
Transgenic line#	**Generation**	**Number of plants tested**	**Number of haploids**	**Number of diploids**	**% Apomixis**
1	T1	19	5	14	26.3
	T2	31	8	23	25.8
2	T1	56	8	47	14.2
	T2	116	19	97	16.3
	T3	34	5	29	14.7
Frequencies of diploid asexual progeny from diploid *S-Apo* plants					
Transgenic line#	**Generation**	**Number of plants tested**	**Number of diploids**	**Number of tetraploids**	**% Apomixis**
1	T1	27	3	24	11.1
	T2	27	4	23	14.8
	T3	13	2	11	15.3
5	T1	41	12	29	29.2
DH#2	T2	121	14	107	11.5
	T3	123	18	105	14.6
	T4	29	4	25	13.7

a, Haploid induction in *BBM1*-ee (*pDD45::BBM1*) transgenic plants. The T_0 primary transformants were hemizygous for the *BBM1*-ee transgene. One diploid T_1 plant 8c from transformant 8 was maintained as a haploid inducer line up to the T_7 generation. **b**, Identification of synthetic haploid and diploid apomictic progeny from *S-Apo* (*MiMe* + *BBM1*-ee) plants of transformant line numbers 1 and 2 (haploids), and line numbers 1 and 5 (diploids). For T_2 and subsequent generations, propagation was performed by selecting from each generation, haploid and diploid progeny respectively. DH#2 refers to a doubled haploid derived from self-pollination of T_1 plants of the haploid apomixis line 2.

nature research

Corresponding author(s): Venkatesan Sundaresan

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Microscopic images were acquired with Zeiss AxioVision 4.8.2. Flow cytometry data were obtained with Becton Dickinson CellQuest. Real time PCR data were collected with Bio-Rad CFX Manager 3.1 software from BioRad. All the sequence chromatograms were acquired with SnapGene.
Data analysis	Trimmomatic 0.38, bwa mem, GATK4.0 HaplotypeCaller, Microsoft excel, BD CellQuest, GraphPad Prism 7.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the data are available from corresponding author upon suitable request .

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were considered to be sufficient based on significance and effect size, no special tests were employed. In most cases, care was taken to use sample sizes that exceed the limits for normal distributions, i.e., >30. Exceptions were made only when the experimental tests were particularly labour intensive, e.g., antibody staining of sections from individual ovules. In such cases, the statistical significance test was used as the sole measure of sufficient sample size.
Data exclusions	No data were excluded.
Replication	All experimental findings were repeated or replicated at least once, in some cases by conducting a second independent analysis in parallel, e.g. the whole genome sequencing was performed on two clonal progeny from each generation. All attempts at repetition or replication were successful.
Randomization	For this type of study, randomization techniques are not applicable. However, where possible, care was taken to select plants or seeds randomly for analysis.
Blinding	Not performed, because of inapplicability to this study

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	All unique materials used are readily available from the authors.

Antibodies

Antibodies used	Rabbit anti-GFP antibody (ab6556, Abcam, Cambridge, UK) was used as primary antibody for immunohistochemistry and goat anti-rabbit AP conjugate (A9919, Sigma-Aldrich, USA) as secondary antibody.
Validation	Validation: Primary rabbit anti-GFP antibody (Abcam, ab6556) was used at a dilution of 1:1000. Secondary goat anti-rabbit AP conjugate (Sigma-Aldrich, A9919) antibody was used at a dilution of 1:200. Anti-GFP antibody has been used for immunohistochemistry previously and shown to react with GFP epitopes in several plant species (Li et al., Nature Communications. 2018; Figueroa-Yañez et al., PLoS One, 2016), Drosophila (Hamp et al., Journal of Cell Science, 2016), mice (Tang et al., The Journal of Neuroscience, 2015) etc (for details https://www.abcam.com/gfp-antibody-ab6556.html). Goat anti-rabbit secondary antibody has been tested previously in plants (Reichelt, et al., The Plant Journal, 1999), Xenopus (Almeida et al., Neural Development 2010) and other species (https://www.sigmaaldrich.com/catalog/product/sigma/a9919?lang=en&region=US).

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Leaf tissue were chopped with a razor blade.
Instrument	Becton Dickinson FACScan system
Software	Becton Dickinson CellQuest
Cell population abundance	All the leaf cells used for flow cytometry are same.
Gating strategy	Only the initial debris was gated out.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

PERSPECTIVE

The Development of C_4 Rice: Current Progress and Future Challenges

Susanne von Caemmerer,[1]* W. Paul Quick,[2] Robert T. Furbank[3]

Another "green revolution" is needed for crop yields to meet demands for food. The international C_4 Rice Consortium is working toward introducing a higher-capacity photosynthetic mechanism—the C_4 pathway—into rice to increase yield. The goal is to identify the genes necessary to install C_4 photosynthesis in rice through different approaches, including genomic and transcriptional sequence comparisons and mutant screening.

As the world population races toward 10 billion, agricultural scientists are realizing that another "green revolution" is needed for crop yields to meet demands for food. In rice, yield potential is limited by the photosynthetic capacity of leaves that, as carbohydrate factories, are unable to fill the larger number of florets of modern rice plants. One potential solution is to introduce a higher-capacity photosynthetic mechanism—the C_4 pathway—into rice. This is the goal of researchers in the international C_4 Rice Consortium: to identify and engineer the genes necessary to install C_4 photosynthesis in rice (*1*).

Rubisco, the primary CO_2-fixing enzyme in rice, is a poor catalyst of CO_2 at current atmospheric conditions. It has a tendency of confusing its substrate CO_2 with the more abundant O_2 as well as being a very slow catalyst of CO_2, turning over only once or twice per second. Rubisco's oxygenase activity requires the recycling of phosphoglycolate in the photorespiratory pathway, resulting in an energy cost and loss of previously fixed CO_2. Many photosynthetic organisms, including cyanobacteria, algae, and land plants, have developed active CO_2-concentrating mechanisms to overcome Rubisco's inefficiencies (*2*). Among land plants, this led to the development of C_4 photosynthesis, a biochemical CO_2-concentrating mechanism. C_4 photosynthesis arose multiple times in the past 60 million years in warm semi-arid regions, with early occurrences coinciding with low atmospheric CO_2 in the late Oligocene (*3*). During C_4 photosynthesis, CO_2 is fixed within specialized leaf tissues known as mesophyll cells to produce C_4 acids, which diffuse to and are decarboxylated in another type of specialized tissue, the bundle sheath cells. This process elevates the CO_2 concentration in the bundle sheath and inhibits Rubisco oxygenase activity, allowing Rubisco to operate close to its maximal rate (Fig. 1). In comparison with C_3 crops such as rice, C_4 crops (such as maize and sorghum) have higher yields and increased water- and nitrogen-use efficiency (*1*, *4*).

Building the C_4 Machinery

In an evolutionary context, the transition from C_3 to C_4 photosynthesis has occurred independently in more than 60 different plant taxa (*3*). Genomic and transcriptional sequence comparisons of cell-specific and leaf-developmental gradient transcription profiles between closely related C_3 and C_4 species are being used to identify C_4-specific regulatory genes (*4*). Combining this information in parallel with screens of mutagenized C_4 *Sorghum bicolor* and *Setaria viridis* along with activation-tagged rice populations hopefully will

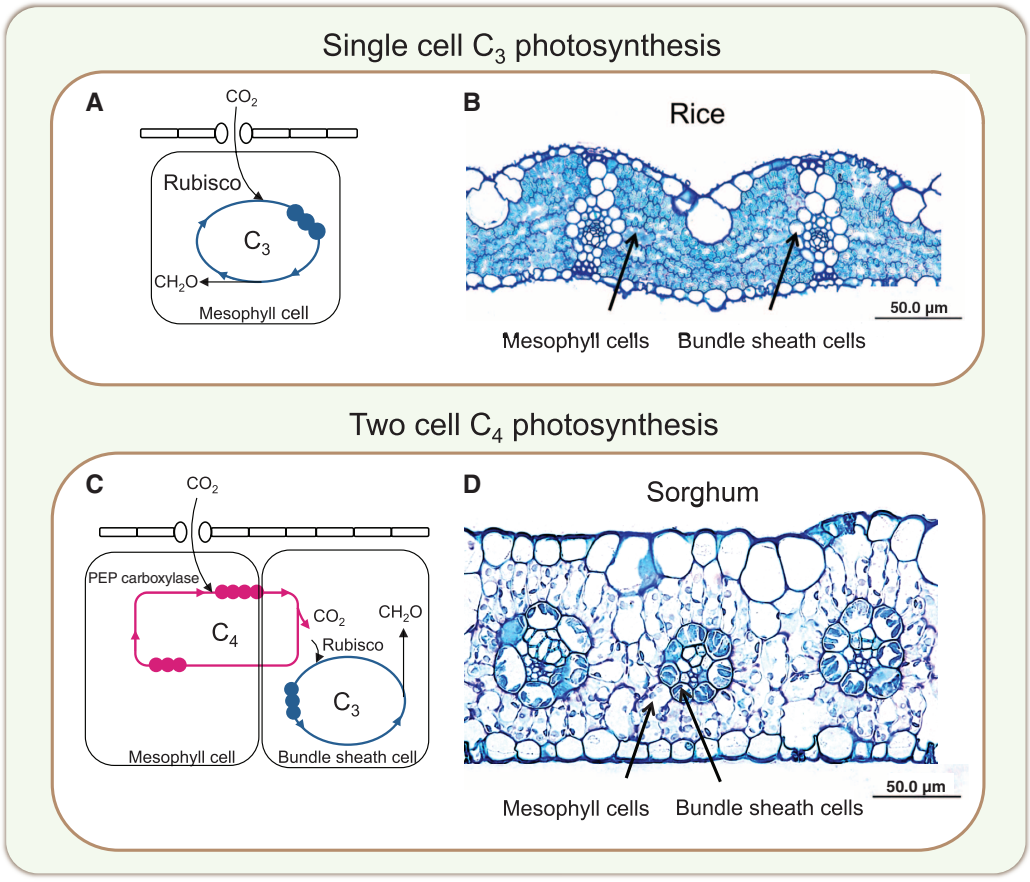


Fig. 1. (**A**) C_3 photosynthesis fixes atmospheric CO_2 into C_3 acids with Rubisco in single cells. (**C**) Two-cell C_4 photosynthesis requires spatial separation of fixation of atmospheric CO_2 into C_4 acids and the donation of CO_2 from these C_4 acids to Rubisco. Also shown are light microscopy images of transverse sections of leaves of (**B**) rice, a C_3 plant, and (**D**) sorghum, a C_4 plant. The rice section shows vascular bundles with few chloroplasts and large numbers of mesophyll cells between the vascular bundles typical for C_3 species. The sorghum leaf section shows chloroplasts in bundle sheath and only two or three mesophyll cells in between the vascular tissue typical of a C_4 species.

[1]Research School of Biology, Australian National University, Canberra, ACT 0200, Australia. [2]International Rice Research Institute, Los Banos, Philippines, and University of Sheffield, Sheffield S10 2TN, UK. [3]High Resolution Plant Phenomics Centre, Commonwealth Scientific and Industrial Research Organization (CSIRO) Plant Industry, Canberra, ACT 2601, Australia.

*To whom correspondence should be addressed. E-mail: susanne.caemmerer@anu.edu.au

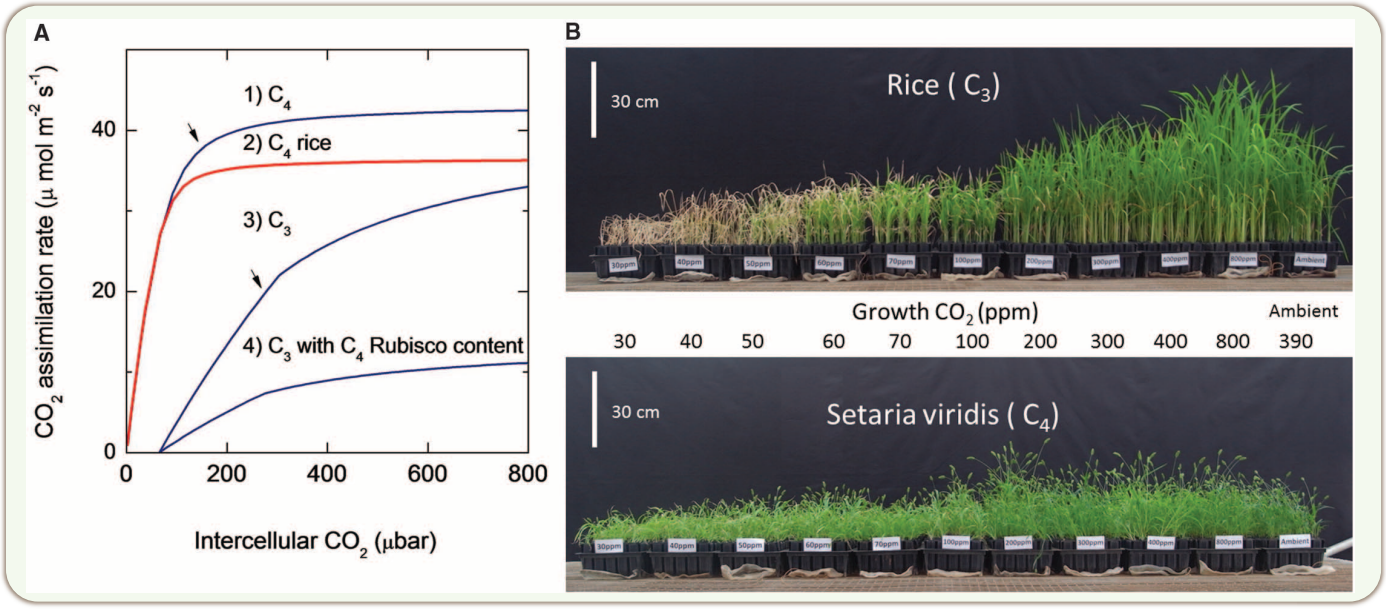


Fig. 2. (**A**) Modeled changes in CO_2 assimilation rate in response to changes in leaf intercellular CO_2 partial pressure for C_3 and C_4 photosynthesis and for a hypothetical C_4 rice. Curves 1, 2, and 4 have Rubisco levels typically found in a C_4 leaf (10 μmol m^{-2} catalytic Rubisco sites). Curve 3 shows a typical response for C_3 leaves with three times the Rubisco level of C_4 leaves. Curve 1 shows the response of a C_4 leaf with C_4 Rubisco kinetic properties. Curve 2 models how a C_4 leaf with C_3 Rubisco kinetic properties would respond (a hypothetical C_4 rice with C_3 Rubisco kinetics). The comparison of these two curves shows the increase in CO_2 assimilation rate achieved with C_4 compared with C_3 Rubisco kinetic properties within a functional C_4 mechanism. Arrows to curves 1 and 3 show intercellular CO_2 partial pressures typical at current ambient CO_2 partial pressures for C_4 and C_3 photosynthesis. To generate the curves, model equations were taken from (*11*) and comparative Rubisco kinetic constants from (*12*). (**B**) Growth of 21-day-old rice and *S. viridis* seedlings at different ambient CO_2 concentrations ranging from 30 to 800 parts per million.

reveal candidate genes in the C_3-to-C_4 switch that can be tested in transgenic rice and *S. viridis* (*5*). Because C_4 plants can carry out net CO_2 assimilation at very low CO_2 levels whereas C_3 plants cannot (Fig. 2), we can use growth screens to identify gain of function in activation-tagged rice mutants and loss of function in *S. viridis* mutants (Fig. 2). We are also using the fact that C_4 photosynthesis imparts a distinct carbon isotope signature on dry matter (*6*) in a loss-of-function screen for C_4 mutants.

A subset of genes required for the major biochemical components and metabolite transporters involved in the C_4 pathway have been cloned and coupled to suitable promoters to give cell-specific expression in rice (*7*). Attempts to install C_4 photosynthesis in plants lacking the appropriate anatomy show that a biochemical approach alone will not be enough (*8*). Bundle sheath cells in rice are smaller than in C_4 plants and have less chloroplasts, and there are a large number of mesophyll cells between vascular bundles (Fig. 1) (*4*). Promising mutants have been identified in rice that show reduced vein spacing. Combined with studies of sorghum, we are optimistic that we will be able to identify the genes controlling this aspect of anatomy (*4*, *7*).

Lessons Learned and Future Challenges

Although C_4 leaves have close veins and high rates of photosynthesis, C_4 photosynthesis is also naturally supported around widely spaced veins in maize husk tissue, albeit at lower rates (*6*). Thus, a prototype C_4 rice may be achievable with a subset of C_4 genes, but a "good" C_4 rice will require substantial fine tuning of biochemistry and anatomy. Particularly intriguing is the need for additional metabolite transport across membranes of organelles in C_4 photosynthesis (*4*). A functional C_4-concentrating mechanism in rice would allow for an approximately two-thirds reduction in Rubisco levels, relative to wild-type rice, but Rubisco would be sequestered in bundle sheath cells and ideally have a greater catalytic turnover rate (Fig. 2) (*2*). Antisense gene suppression of key photosynthetic enzymes has illuminated C_4 metabolism and engineering strategies, including the surprising find that phosphorylation of phosphoenolpyruvate (PEP) carboxylase by the regulatory enzyme PEP carboxylase phosphokinase is not needed for C_4 function (*9*). With the adoption of the C_4 model plant *S. viridis*—with its short life cycle, small stature, and genome size—along with advances in efficient transformation, we anticipate that much more will soon be learned (*5*). We expect to have a C_4 rice prototype within 3 years. However, we estimate that another 15 years of research are required for optimization of the phenotype and field testing for C_4 rice to become ready for cultivation in farmers' fields.

Norman Borlaug's green revolution was based on just a handful of genes (*10*). However, the need for even greater food plant production looms. The promise of C_4 rice has resulted in one of the largest consortia of plant biologists pursuing a common goal. We optimistically take on this challenge, anticipating that advances in our understanding of plant metabolism, and C_3 and C_4 photosynthesis in particular, will better serve humanity in years to come.

References and Notes

1. J. M. Hibberd, J. E. Sheehy, J. A. Langdale, *Curr. Opin. Plant Biol.* **11**, 228 (2008).
2. M. R. Badger *et al.*, *Can. J. Bot.* **76**, 1052 (1998).
3. R. F. Sage, P. A. Christin, E. J. Edwards, *J. Exp. Bot.* **62**, 3155 (2011).
4. J. A. Langdale, *Plant Cell* **23**, 3879 (2011).
5. T. P. Brutnell *et al.*, *Plant Cell* **22**, 2537 (2010).
6. J. J. L. Pengelly *et al.*, *Plant Physiol.* **156**, 503 (2011).
7. K. Kajala *et al.*, *J. Exp. Bot.* **62**, 3001 (2011).
8. M. Miyao, C. Masumoto, S. Miyazawa, H. Fukayama, *J. Exp. Bot.* **62**, 3021 (2011).
9. T. Furumoto, K. Izui, V. Quinn, R. T. Furbank, S. von Caemmerer, *Plant Physiol.* **144**, 1936 (2007).
10. N. Borlaug, *Science* **318**, 359 (2007).
11. S. von Caemmerer, *Biochemical Models of Leaf Photosynthesis*, vol. 2, *Techniques in Plant Sciences* (CSIRO Publishing, Collingwood, Australia, 2000).
12. A. B. Cousins, O. Ghannoum, S. von Caemmerer, M. R. Badger, *Plant Cell Environ.* **33**, 444 (2010).

Acknowledgments: This work was supported by the Bill and Melinda Gates Foundation. We are thankful for the scientific contributions of all the members of the C_4 Rice Consortium.

10.1126/science.1220177

The Development of C_4 Rice: Current Progress and Future Challenges

Susanne von Caemmerer, W. Paul Quick and Robert T. Furbank

Science **336** (6089), 1671-1672.
DOI: 10.1126/science.1220177

ORGANIC CHEMISTRY

An "ideal lignin" facilitates full biomass utilization

Yanding Li[1,2], Li Shuai[1,3], Hoon Kim[1,4], Ali Hussain Motagamwala[1,5], Justin K. Mobley[1], Fengxia Yue[1,4], Yuki Tobimatsu[1,4,6], Daphna Havkin-Frenkel[7,8], Fang Chen[9,10], Richard A. Dixon[9,10], Jeremy S. Luterbacher[11], James A. Dumesic[1,5], John Ralph[1,2,4]*

Lignin, a major component of lignocellulosic biomass, is crucial to plant growth and development but is a major impediment to efficient biomass utilization in various processes. Valorizing lignin is increasingly realized as being essential. However, rapid condensation of lignin during acidic extraction leads to the formation of recalcitrant condensed units that, along with similar units and structural heterogeneity in native lignin, drastically limits product yield and selectivity. Catechyl lignin (C-lignin), which is essentially a benzodioxane homopolymer without condensed units, might represent an ideal lignin for valorization, as it circumvents these issues. We discovered that C-lignin is highly acid-resistant. Hydrogenolysis of C-lignin resulted in the cleavage of all benzodioxane structures to produce catechyl-type monomers in near-quantitative yield with a selectivity of 90% to a single monomer.

INTRODUCTION

Lignin is a polymeric material composed of phenylpropanoid subunits and is one of the largest sources of naturally produced aromatics on the planet. Because of its aromatic nature, lignin has a higher energy density than polysaccharide polymers, as well as a higher potential commercial value (*1*). However, because of lignin's complexity, its efficient utilization, either as a polymer or from its derivable small-molecule products, is currently problematic (*1*–*3*).

Although mild depolymerization methods, such as oxidative (*4*, *5*) and hydrogenolytic (*6*–*8*) procedures, have produced encouraging results in laboratory-scale experiments, their applicability in industrial processes has been limited. Direct hydrogenolysis, that is, the hydrogenation of unprocessed solid biomass by a heterogeneous metal catalyst, remains one of the most promising methods for cleaving lignin's ether bonds and producing aromatic monomers in high yields (*8*–*10*). However, hydrogenolysis still suffers from product complexity issues. In most wild-type biomass, the lignin polymer is composed of three phenylpropanoid subunits—*p*-hydroxyphenyl (**H**), guaiacyl (**G**), and syringyl (**S**)—derived by combinatorial radical coupling from the three main monolignols (*p*-coumaryl, coniferyl, and sinapyl alcohols). Although **H** units are typically at low-levels, this results in at least three different types of monomers (**H**, **G**, and **S**), each with a selection of side chains, as the primary hydrogenolysis products, which makes monomer separation and utilization difficult. Lignin's principal alkyl-aryl-ether units with their β–O–4 interunit bonds (45 to 85%) can be selectively cleaved, but other linkages including β–5 (1 to 12%), β–β (5 to 12%), 5–5 (1 to 9%), 4–O–5 (~2%), and β–1 (1 to 2%), which are also present in lignins, remain largely uncleaved (*8*); carbon-carbon (C–C) and diaryl ether (4–O–5) units typically result from dimeric or higher oligomeric products.

The use of extracted lignins rather than whole biomass has the advantage that the material can be fully dissolved in organic solvents, facilitating catalyst recovery and continuous processing. However, acidic industrial lignin fractionation is known to cause some β-ether cleavage and condensation between units via the electrophilic substitution of acid-generated benzylic carbocation intermediates on the electron-rich aromatic rings (*7*, *11*), limiting depolymerization yields (Scheme 1A) (*12*–*14*). There are some elegant solutions focusing on suppressing the condensation reaction, either using a capping agent (*7*, *15*) or using two-step strategies (*4*, *5*, *11*). However, extra chemicals or catalysts are needed to achieve this goal.

Bioengineered biomass could be used to achieve higher hydrogenolysis yields and simpler product mixtures. For example, the recent use of formaldehyde protection during lignin extraction from a high-**S** poplar lignin (*7*, *16*) that has up to 98% syringyl **S** units and ~90% β–O–4 linkages [from nuclear magnetic resonance (NMR) estimates] prevented condensation reactions and allowed an unprecedentedly high monomer yield (78%) under hydrogenolytic conditions (*7*). However, even in this high-**S** lignin, some 10% of the linkages are C–C bonds that do not cleave. The use of formaldehyde to protect the lignin from condensation reactions also resulted in some formaldehyde addition to the ring, complicating the hydrogenolysis products with methyl-substituted aromatics. Without formaldehyde, the lignin extracted under acidic conditions had significant condensation, thwarting the production of monomers and resulting in a hydrogenolysis monomer yield of only 26% (*7*). Although new methods for displacing formaldehyde for the protection from acid-catalyzed condensation reactions, retaining much of the yield (70%) and producing a simpler monomer mix, have recently been revealed (*17*), extra protection chemicals remain necessary during the lignin extraction.

RESULTS

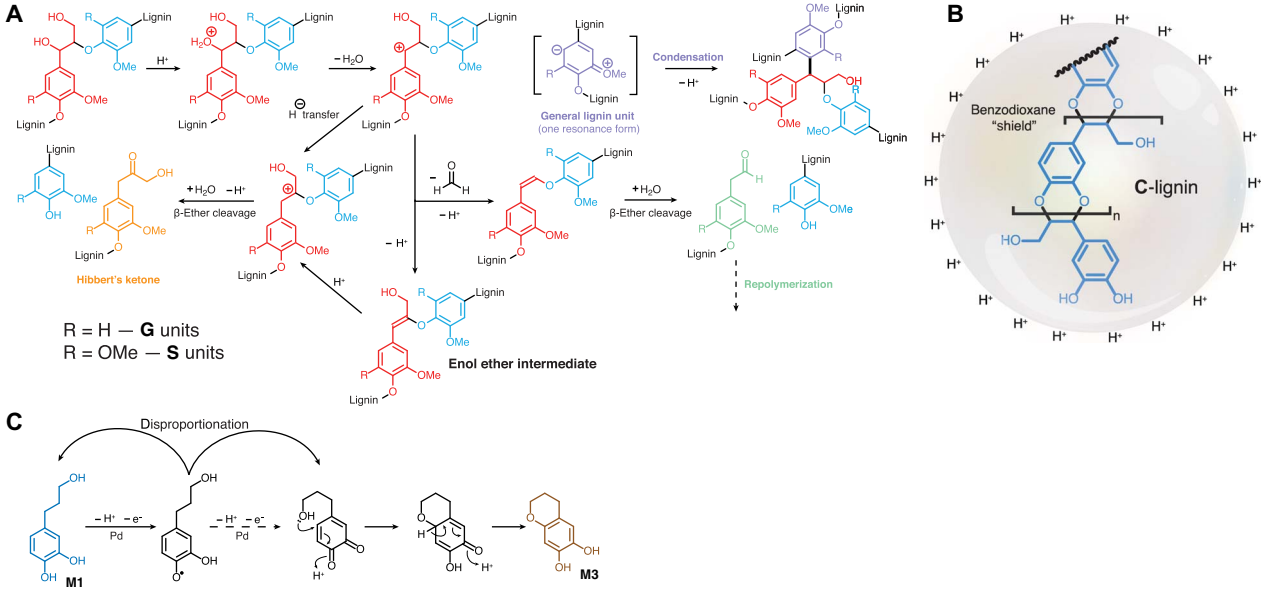
An "ideal lignin" archetype

On the basis of the plethora of information stemming from the lignin biosynthetic research community over the last decade, and with the revelations regarding lignins' structural malleability from studies on lignin pathway mutants and transgenics as well as on various "natural" plants discovered to have unusual lignins, researchers have been able to

[1]U.S. Department of Energy Great Lakes Bioenergy Research Center, and Wisconsin Energy Institute, University of Wisconsin–Madison, Madison, WI 53726, USA. [2]Department of Biological Systems Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA. [3]Department of Sustainable Biomaterials, Virginia Tech, Blacksburg, VA 24061, USA. [4]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI 53706, USA. [5]Department of Chemical and Biological Engineering, University of Wisconsin–Madison, Madison, WI 53706, USA. [6]Research Institute for Sustainable Humanosphere, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [7]Department of Plant Biology and Pathology, Rutgers, State University of New Jersey, New Brunswick, NJ 08901, USA. [8]Bakto Flavors LLC, 772 Cranbury Crossroad, North Brunswick, NJ 08092, USA. [9]BioDiscovery Institute and Department of Biological Sciences, University of North Texas, Denton, TX 76203, USA. [10]Center of Bioenergy Innovation, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [11]Laboratory of Sustainable and Catalytic Processing, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.
*Corresponding author. Email: jralph@wisc.edu

contemplate designing lignins for improved utilization (*18*). It is now a realistic juncture to posit the characteristics for an ideal lignin archetype for biomass processing. For the depolymerization of the polymer to monomers, lignin should have at least the following three characteristics. First, if acidic pretreatment is used, then it should be stable under acidic conditions to prevent condensation and the generation of undesired new C–C bonds. Second, it should contain only ether (C–O) interunit linkages in its backbone so that it can be fully depolymerized. Finally, it should be generated *in planta* from a single phenylpropanoid monomer to allow the production of the simplest array of compounds.

We have reported the discovery of an unusual catechyl lignin (**C**-lignin) present in the seed coats of vanilla (*Vanilla planifolia*) (*19*) and various members of the Cactaceae of the genus *Melocactus* (*20*). In this special case, the lack of *O*-methyltransferase (OMT) activity for conversion from catechyl **C** to guaiacyl **G** and, subsequently, on to syringyl **S**, aromatic-level precursors, results in 100% **C** units in the cell wall (CW). This **C**-lignin was, somewhat surprisingly, found to be essentially a homopolymer synthesized almost purely by β–O–4 coupling of caffeyl alcohol with the growing polymer chain, producing benzodioxanes as the dominant unit in the polymer (Fig. 1A). If it has particular stability toward biomass pretreatment conditions, then this



Scheme 1. Mechanisms for lignin condensation, C-lignin structure, and monomer M3 formation. (**A**) Mechanism of lignin acidolysis and condensation routes. (**B**) The benzodioxane structure acts as a "shield" that can protect **C**-lignin from unwanted acidolysis and condensation reactions. (**C**) Proposed mechanism for the cyclization reaction of **M1** to **M3**.

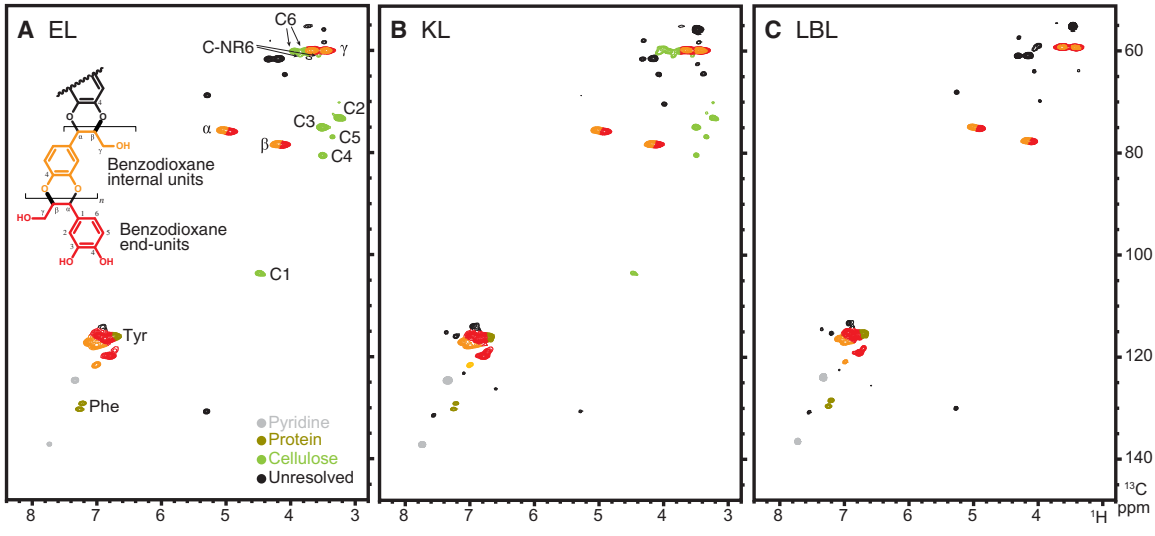


Fig. 1. NMR spectra. Partial 2D HSQC NMR spectra of (**A**) EL, (**B**) KL, and (**C**) LBL from vanilla (*V. planifolia*) seed coat. There are no obvious lignin structural changes after the acidic lignin extraction processes. Cellulose was labeled following the conventional monosaccharide nomenclature; NR is the nonreducing end of the cellulose. Protein residuals were labeled by the aromatic amino acid. Tyr, L-tyrosine; Phe, L-phenylalanine; ppm, parts per million.

C-lignin might therefore represent an example of such an ideal lignin that can, in principle, be depolymerized to a single product by hydrogenolysis. Furthermore, this substrate has the potential to produce valuable catechol monomers, whereas the large majority of monomers produced from lignin have been **S** or **G** derivatives (*1*, *2*). Expanding the arsenal of lignin-derived platform molecules could play an important role in the successful use of this fraction within future biorefineries. Here, we describe the ideal nature of this lignin via a revised compositional characterization of the vanilla seed coat fiber, new features of the **C**-lignin's reactivity and stability, and our successful attempts at converting it to monomers in near-quantitative yields.

Acid stability of C-lignin

Because of the lack of an accessible and eliminable benzylic hydroxyl group in **C**-lignin units (Scheme 1B), condensation reactions due to the formation of benzyl cations might be mitigated under acidic conditions. We therefore examined the acid stability of the polymer to determine whether acidolytic methods could be used to purify the lignin. Comparison of the two-dimensional (2D) heteronuclear single-quantum coherence (HSQC) NMR spectra from the enzyme lignin (EL) (derived by removing polysaccharides via crude cellulases treatment) (*21*) and Klason lignin (KL) showed no significant differences in the lignin structure (Fig. 1, A and B) (*22*). The **C**-lignin survives even the harshest of acidic pretreatment methods—the KL isolation procedure includes a 1-hour treatment in 72% (w/w) sulfuric acid, followed by dilution to 4% (w/w) sulfuric acid and autoclaving at 121°C for 1 hour—while retaining its original lignin structure. An efficient acidic lithium bromide (LiBr) pretreatment method was also used to purify the lignin. This treatment method is known for its quick and near-quantitative removal of the polysaccharides to give an LiBr lignin (LBL) (Fig. 1C) (*23*). The molecular weight of the LBL was shown to be similar to that of the EL (Fig. 2). The **C**-lignin polymer appeared to survive this pretreatment based on the retention of its key lignin structural features in its NMR spectra and little change in its molecular weight distribution. On the basis of these results, we can conclude that, unlike normal **S**-**G** lignins, polysaccharides can be removed via acid pretreatment from **C**-lignin without its suffering from unwanted condensation reactions. After removing the polysaccharides, the resulting lignins (EL, KL, and LBL) were completely soluble in various organic solvents [for example, acetone, dioxane, or tetrahydrofuran (THF)] mixed with water to match lignin solubility parameters (*24*). Efficient lignin solubilization should greatly facilitate continuous processing in an industrial setting.

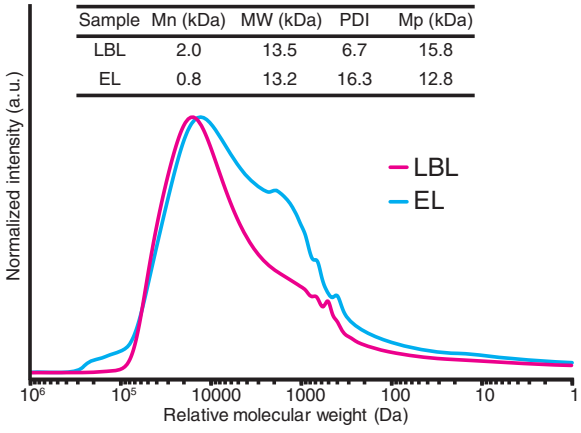


Fig. 2. Molecular weight profiles. Molecular weight profiles of EL (cyan) and LBL (magenta) from *V. planifolia* seed coat measured by gel-permeation chromatography (GPC). The *x* axis indicates the apparent molecular weight of individual lignin polymers and is shown as a log scale. The *y* axis shows the response of a UV-light detector (at 280 nm) normalized to the most abundant signal in each chromatogram. The most abundant signal in the each of the two samples corresponds to a molecular weight of ~13,000 Da (determined via polystyrene standards); comparison shows that there was no obvious lignin polymer degradation during the acid pretreatment. PDI is the polydispersity index. a.u., arbitrary units. MW, molecular weight.

Response of C-lignin to traditional degradative methods

To investigate the potential for **C**-lignin depolymerization, we applied two traditional lignin degradative analytical methods, alkaline nitrobenzene oxidation (NBO) and thioacidolysis, to a **C**-lignin model compound, the caffeyl alcohol dimer **D1** (**C**-dimer), and to the vanilla bean seed coat CW (Fig. 3 and fig. S3). Although relatively low yields of the corresponding monomeric products (30 to 60%) were obtained from the dimeric compound, the use of the CW gave monomeric products in extremely low yields (<1%). As discussed widely in the past, both thioacidolysis and alkaline oxidation need the involvement of a free benzylic hydroxyl group on the lignin side chain (*25*, *26*). It was therefore concluded that, because of the stability of the 1,4-benzodioxane structure, especially under the tested acidic and alkaline oxidative conditions, traditional lignin chemical degradation methods are ineffective for the depolymerization of **C**-lignin. A computational approach to evaluate the bond dissociation energy (BDE) of **C**-lignin using density functional theory models suggested that depolymerization of **C**-lignin is theoretically possible (*27*). Although the benzodioxane β–O–4 bond calculates to have a slightly higher BDE value than a conventional β–O–4 bond, it is still much lower than the BDEs of lignin's C–C bonds (*28*).

Catalytic hydrogenolysis of C-lignin

We reasoned that hydrogenolysis had the potential to more efficiently depolymerize **C**-lignin. We first sought efficient methods for cleaving dimeric model **D1**, rationalizing that, although the corollary is not necessarily true, any reaction conditions that did not produce high yields from **D1** would have little chance of being effective on the polymer. When hydrogenolysis was applied to the **C**-dimer **D1** and vanilla seed coat CW, analysis by gas chromatography with flame-ionization detection (GC-FID) showed that the products were rather simple with dominant products **M1** (catechylpropanol), **M2** (catechylpropane), and **M3** (chroman-6,7-diol), together with some minor products (Fig. 3). The major products, **M1** and **M2**, were identified by comparison with authentic synthetic standards. The initially puzzling minor product **M3**, which is a cyclization product from **M1**, was separated from the product mixture by silica-gel chromatography, characterized, and structurally identified by NMR and high-resolution mass spectrometry (MS). Because it was not obvious whether the chromane ring oxygen originated from the lignin γ-OH or from water, the hydrogenolysis reaction was run in ^{18}O-labeled $H_2{}^{18}O$. No ^{18}O was detected in the product **M3**, so the cyclization mechanism was concluded to involve the γ-OH via a radical disproportionation reaction (Scheme 1C) (*29*). This is the first report of this lignin hydrogenolysis product. The minor impurity peaks displayed in the chromatograms from the CW materials (fig. S3) were derived from the solvent, polysaccharide, and fatty acid products, which were identified via GC-MS.

Monomer production data under different conditions are shown in Fig. 4. Yields are normalized to the total molar concentration of caffeyl alcohol in **C**-lignin determined from quantitative ^{13}C NMR (table S2). Not surprisingly, the monomer distributions were heavily affected by

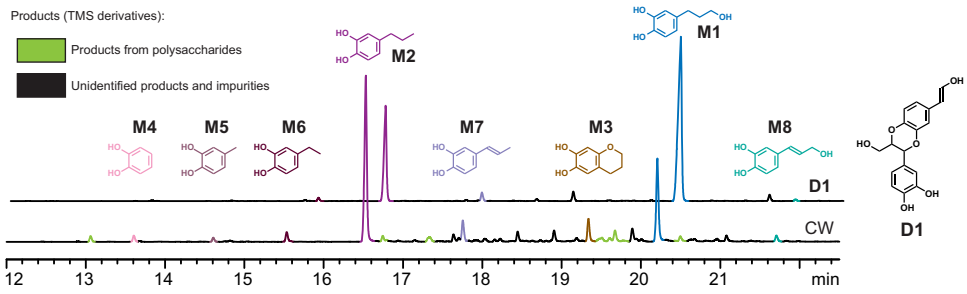


Fig. 3. GC-FID spectra of hydrogenolysis products from dimeric compound D1 and from CW. Hydrogenolysis condition: Pt/C, 200°C, 40-bar H_2, 15 hours. Coloring of peaks matches that of the structures for monomers **M1** to **M8**. Products from polysaccharide in the CW are colored light green, and unidentified products from other non-lignin compounds are left in black. TMS, trimethylsilyl. Note that the upper **D1** product chromatogram is offset by ~0.3 min.

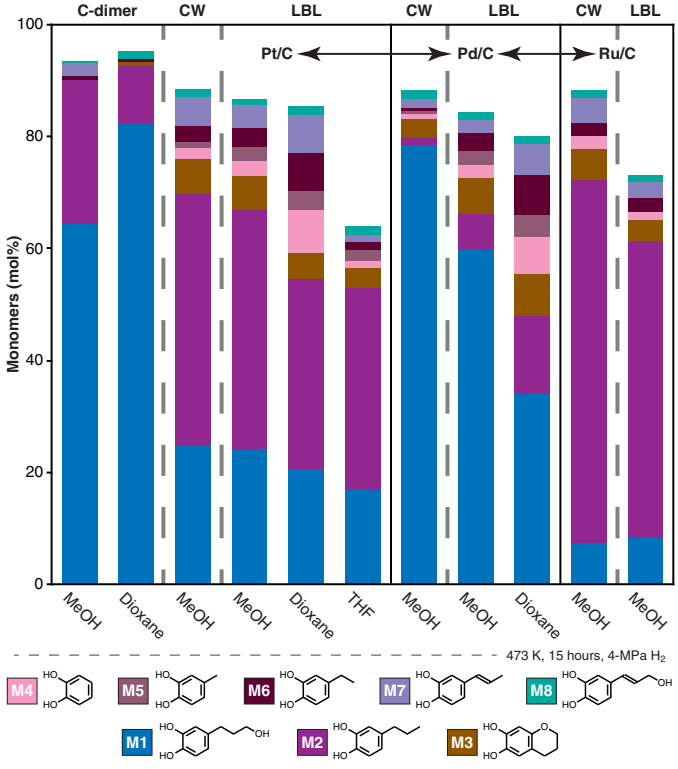


Fig. 4. Hydrogenolysis monomer yields from different catalyst and solvent combinations. Yields are on a **C**-lignin molar basis (see also table S3, from left to right: entries 1, 2, 3, 5, 7, 9, 10, 12, 14, 19, and 21).

the choice of catalyst and solvent (*17*, *30*, *31*). Here, we illustrate that Pt/C showed a slightly higher reactivity, whereas Pd/C and Ru/C showed a much better product selectivity. More side chain truncation products were obtained from **C**-LBL compared to that from vanilla seed coat CW, suggesting that a significant degree of side chain truncation occurred during the acid pretreatment stage or that the isolated lignin was more accessible to the catalyst. In terms of solvent effects, methanol produced a slightly higher monomer yield compared to dioxane, whereas THF gave a substantially lower yield. Both monomer yield and reaction selectivity were maximized using Pd/C or Ru/C as catalyst and methanol as the solvent. Retaining or losing the hydroxy group on the side chain can be controlled by simply changing the catalyst to satisfy the different intended purposes for using the catechyl monomers.

Thus, Pd/C produced the catechylpropanol monomer **M1** with 89% selectivity, whereas Ru/C produced the catechylpropane monomer **M2** with 74% selectivity. Increasing the hydrogenolysis reaction time from 3 to 15 hours (table S3, entries 16 to 19) led to an ~10% increase in lignin conversion and monomer yield. The resulting product oil mixture after vacuum drying was completely soluble in methanol, ethanol (EtOH), dioxane, pyridine, and other solvents but only partially soluble in acetone, ethyl acetate, and dichloromethane (DCM) due to the presence of products from degraded polysaccharides and other non-lignin components. The mass balance and total organic carbon (TOC) analyses (Table 1) indicated that volatile products were minimal or insignificant.

A 2D HSQC NMR spectrum of the total hydrogenolysis product, which was completely soluble in dimethyl sulfoxide (DMSO)/pyridine (4:1, v/v), demonstrated that the **C**-lignin had been completely depolymerized, that is, no detectable residual benzodioxane structures remained (fig. S4A). The major products were fully authenticated by comparison with synthetic compounds **M1** and **M2** and with authenticated isolated **M3**. No detectable products from side reactions or recondensation were detectable. The GPC molecular weight profile of the hydrogenolysis products mixture from **C**-LBL before and after the hydrogenolysis reactions showed a dominant monomer peak (fig. S4B). The high–molecular weight fractions were separated from monomer fractions, and the fractions were characterized by HSQC NMR (fig. S7). The data revealed that only traces of the original benzodioxane structures from the **C**-lignin remained in the product and that the high–molecular weight fractions contained only nonaromatic components present in the original sample and were therefore not from the lignin proper. It can therefore be safely concluded that essentially all of the **C**-lignin in the samples was depolymerized to monomeric compounds during hydrogenolysis. The non-lignin components in the lignin stream were nonextractable oils, waxes, or the other (difficult to remove) components in the sample that are not necessarily associated directly with the phenylpropanoid polymer.

DISCUSSION

Prospects for C-lignin and its derived catechylpropanoid monomers

Catechols in nature are remarkably biochemically active; because of the interaction of the vicinal phenolic hydroxyl groups, catechols play a vital role in both biomedical and biomimetic functional materials (*32*). Their synthesis is challenging because of the difficulty of transforming phenols

Table 1. Mass balance and TOC on hydrogenolysis of C-LBL and its resulting product oil.

Feed	CW	Dissolved C-LBL
Solid recovery*	55–74%	~100%
Oil recovery†	23–35%	50–60%
TOC of **C**-LBL	—	62.66 ± 0.23%
TOC of product oil	—	61.44 ± 0.34%

*Solid includes recovered CW material and catalyst. †Oil yield on a CW and **C**-LBL mass basis.

to catechols; although researchers have recently developed several catechol synthetic methods (*33*), applying those methods at scale remains complicated. There is little reference to high-yielding biomass conversion to catechols, although catechols were reported as hydrogenolysis products from organosolv lignin of candlenut shells (using Cu-doped porous metal oxides) in which the cleaving of the aromatic methoxyl groups during the reaction was claimed (*34*). Catechols act as important intermediates for the conversion of lignin-derived monomers to value-added platform chemicals via the bacterial β-ketoadipate pathway (*35*). We therefore contend that it would be beneficial, and more energy-efficient for aromatic metabolism/catabolism, if high yields of catechols could be obtained directly from lignin.

Our study provides a new perspective for the production of catechols from a renewable biomass source rather than petroleum. Compounds **M1** and **M2** are currently not available in bulk, so their commercial value is not obvious. However, an enriched diversity of the raw materials from the catechol family would likely provide significant value. A LiBr pretreatment method is able to convert a fraction of β–O–4 units in **S**-**G**–type lignins into benzodioxanes (*36*). Large amounts of catechols are potentially producible if we could produce benzodioxane-type lignins in energy crops. We do not yet know if genetically engineered plants, including plantation trees such as pines and poplars, in which the production of lignin is on a large scale, will tolerate **C**-lignins in stem tissues; **C**-polymers have been evidenced in a gymnosperm tracheary element system, in which OMT activity was down-regulated (*37*), but have not yet been found in OMT–down-regulated dicots, and we suspect that additional activities will need to be suppressed for the synthesis and deposition of the **C**-lignin polymer. Given the unique acid-resistant property of **C**-lignin, the potential value of the monomeric products, the homogeneous nature of **C**-lignins that is already known to aid lignin fiber production (*38*), and the high conversion to catechol monomers by hydrogenolysis reported here, we suggest that continuing to pursue the means to produce **C**-lignins in planta is decidedly worthwhile.

C-lignin therefore has numerous compelling features for a biorefinery operation aimed at delivering value from its lignin component. It maintains its native structure after treatment under even strongly acidic conditions; acid pretreatment can therefore be applied to vanilla seed coats to recover the polysaccharide while retaining the native **C**-lignin structure. After sufficient pulverization followed by acid pretreatment, **C**-lignin could be dissolved in organic solvents, enabling both detailed NMR analysis and continuous processing schemes. **C**-lignin can be completely depolymerized by a hydrogenolytic method to produce simple monomeric catechols near-quantitatively and, by selecting the catalyst, with a single monomer accounting for 90% of the monomer product. The yield and selectivity for a single monomer are higher than for any other lignin or biomass to date (fig. S6). There is therefore considerable potential for economic hydrogenolysis of **C**-lignin–rich waste biomass resources only now being structurally characterized, such as Jatropha (*Jatropha curcas*) seed coats (*39*) and candlenut (*Aleurites moluccanus*) shells (*40*), and via genetic engineering if high levels of **C**-lignin could be expressed in traditional biomass sources. Such an approach toward significantly valorizing lignins and biomass in biorefining processes would aid process economics.

MATERIALS AND METHODS

C-lignin sample pretreatment

Processing of seed coat material

Vanilla seed and pod were received as a mixture from a natural vanilla processing plant (Bakto Flavors LLC). The mixture was sifted, and the lower-density remaining pod powder was blown away using a heat gun (set on cold). Preparation of vanilla seed coat NMR samples was via methods described previously (*22*). Briefly, isolated vanilla seed coats (4 × 300 mg) were ball-milled (30 × 10 min, 5-min cooling cycle) using a Retsch PM100 ball mill vibrating at 600 rpm with ZrO_2 vessels containing ZrO_2 ball bearings. Preground seed coat was extracted using a modified Bligh and Dyer extraction (*41*) to remove oils and extractives.

Modified Bligh and Dyer extraction

Vanilla seed material (100 g in total) was shaker-milled (MM400, Retsch) at 3600 rpm for 5 min using a 50-ml stainless steel jar and a single 20-mm ball bearing. The milled sample was transferred to a 1-liter volumetric flask, and a magnetic stir bar was added. Deionized (DI) water (80 ml), chloroform (100 ml), and methanol (200 ml) were added, and the mixture was stirred at 50°C for 30 min. To the mixture was then added 100 ml more of chloroform, and then, after another 30 min, 100 ml of DI water was added. The stirring was continued at 50°C for 24 hours, and the insoluble material was removed by centrifugation (3800 rpm for 15 min), retaining the solids by decanting off the solvent and keeping the filtrate as well. The residue was extracted again by the same method. The filtrates were combined, and the solvents were removed by rotary evaporation to produce the extractives fraction for analysis.

EL from vanilla seed coat

The ball-milled extract-free vanilla seed coat material (1 g) was placed in centrifuge tubes and digested at 35°C with crude cellulases [CELLULYSIN cellulases, *Trichoderma viride*; sample (50 mg/g) in acetate buffer (pH 5.0); two times over 3 days; fresh buffer and enzyme were added each time; catalog no. D00074989, Calbiochem], leaving all of the phenolic polymers and residual polysaccharides totaling 859 mg (85.9%) (table S1).

Acidic LiBr pretreatment of C-lignin from vanilla seed coat

C-LBL was prepared using the acidic LiBr trihydrate method described previously (*23*). Briefly, ball-milled extract-free vanilla seed coat material (1 g) was added into a 40-ml glass vial with a polytetrafluoroethylene (PTFE) lined cap, together with 4.50 ml of acidic 60 weight % (wt %) LiBr solution containing 0.04 M HCl. The vial was immersed into an oil bath preheated at 110°C under magnetic stirring. The mixture was filtered under vacuum and washed with water. The residues were dried at 40°C under reduced pressure (yield, 72.4%; table S1).

Compositional analysis

KL analysis was performed by the two-stage sulfuric acid hydrolysis following the National Renewable Energy Laboratory's standard protocol

(*42*). Briefly, 0.3 g of biomass (weighed to the nearest 0.1 mg) was treated in 72% (w/w) H_2SO_4 at room temperature for 60 min. The slurry was diluted to 4% (w/w) H_2SO_4 and autoclaved at 121°C for 60 min. After filtration, the acid-insoluble lignin (AIL = KL) and the acid-soluble lignin were quantitated gravimetrically and spectrophotometrically, respectively (table S1). Monosaccharides in the KL filtrates (hydrolysates) were quantitated using high-performance ion-chromatography on a Dionex ICS-3000 system equipped with an integrated amperometric detector and a CarboPac PA1 column (4 × 250 mm) at 30°C. DI water was used as an eluent at a flow rate of 0.7 ml/min according to the following gradient: 0 to 25 min, 100% water; 25.1 to 35 min, 30% water and 70% 0.1 M NaOH; and 35.1 to 42 min, 100% water. The post-run eluent of 0.5 M NaOH at a flow rate of 0.3 ml/min was used to purge remaining materials from the column to ensure baseline stability and detector sensitivity (*23*). Crude protein content was determined from the nitrogen (N) content using a 6.25 N-to-protein factor (table S1). The total N was determined using an elemental combustion system (model 4010, Costech Analytical Technologies). Samples (approximately 10 mg) were accurately weighed into tin combustion cups using a microbalance. After complete combustion, total N was measured as N_2 gas. The compositional analysis results are shown in table S1.

C-lignin characterization and quantification

Lignin characterization by 2D NMR spectroscopy

NMR spectra were acquired on a Bruker Biospin AVANCE III 700 MHz spectrometer fitted with a cryogenically cooled 5-mm QCI $^1H/^{31}P/^{13}C/^{15}N$ gradient probe with inverse geometry (proton coils closest to the sample), and spectral processing used Bruker's TopSpin 3.5pl6 (Mac) software. For NMR experiments, ball-milled whole vanilla seed coat material was swelled in DMSO-d_6/pyridine-d_5, isolated lignins and **C**–DHP (dehydrogenation polymer) were dissolved in 4:1 v/v DMSO-d_6/pyridine-d_5, and model compounds were dissolved in acetone-d_6. The central solvent peaks were used as the internal references (δ_C/δ_H: DMSO, 39.5/2.49; acetone, 29.84/2.05 ppm). Standard Bruker implementations of the traditional suite of 1D and 2D [gradient-selected and ^1H-detected; for example, correlation spectroscopy (COSY), ^1H–^{13}C HSQC (Fig. 1), and heteronuclear multiple-bond correlation (HMBC)] NMR experiments were used for structural elucidation and assignment authentication for monomers and dimers. Adiabatic 2D HSQC ("hsqcetgpsisp2.2") experiments for ball-milled seed coat material in a gel state were carried out using the parameters described previously (*22*). Processing used typical matched Gaussian apodization in F2 (LB = −0.5; GB = 0.001) and squared cosine-bell apodization in F1.

The characterization of the vanilla seed coat **C**-lignin was initially consistent with the previous report (*19*) but belied some issues. For both the CW and its EL (derived by removing polysaccharides via crude cellulases treatment) (*21*), characterization revealed each lignin to be an almost 100% benzodioxane polymer with only a trace level of the resinol (β–β) structure. Although not as high as we previously reported (~80%) (*19*), the seed coat sample had a very high KL value (~65%). However, the 2D HSQC NMR of the so-purified lignins contained many peaks in the aliphatic region that were not from the lignin itself (fig. S1). An alternative method (below) was therefore required for lignin quantification in these materials.

C-lignin quantification by ^{13}C NMR

Samples for quantitative ^{13}C NMR analysis were prepared by accurately weighing predried **C**-LBL samples (100 mg) dissolved in 1-ml internal standard [1,3,5-trioxane, DMSO-d_6 (3.12 mg/ml)] solution. The **C**-LBL concentration was also 100.0 mg/ml. Relaxation reagent chromium(III) acetylacetonate [$Cr(acac)_3$; ~2 mg] was added to the samples to facilitate the relaxation of the magnetization. Quantitative ^{13}C NMR spectroscopy was performed as previously described (*43*). The NMR spectra were acquired on the 700-MHz spectrometer described above. Relaxation delays were set to be ~5 times the longest T1 values of carbon signals (for inverse-gated proton decoupled ^{13}C NMR spectra); in our case, d1 = 12.5 s was used to fully relax of all of the carbons with the aid of the relaxation reagent. For the inverse-gated proton-decoupled ^{13}C spectrum, at least 38 hours (10K scans) were required. Spectral processing used both Bruker's TopSpin 3.5pl6 (Mac) and MestreNova 11.0 (Mac) software. The acquired FIDs were processed typically with a 5-Hz line broadening. The central solvent peaks were used as the internal references (δ_C/δ_H: DMSO, 39.5/2.49 ppm). Baseline was corrected manually over the 50- to 100-ppm region using TopSpin.

^{13}C NMR is mostly used to quantify low–molecular weight technical lignins (such as kraft lignin and organosolv lignin) or milled wood lignins (*43*, *44*). It is difficult to quantify native lignin with ^{13}C NMR for two reasons. One is the poor solubility of lignin, and the other is the overlapping peaks from the lignin side chain with polysaccharide peaks. However, **C**-LBL is a perfect sample for ^{13}C NMR analysis. First, the lignin structure is simple; there is only one type of structure in the lignin backbone—the benzodioxane derived from β–O–4-coupling. The chemical shifts of the benzodioxane carbons are unique (75 to 80 ppm), so that there is little chance of signal overlap with other components. Second, **C**-lignin is acid-resistant. Unlike the **S**-**G**–type lignins, harsh acid pretreatment can be applied to **C**-lignin without destroying the benzodioxane structure. Thus, we can easily remove the polysaccharides by acid pretreatment, further minimizing the signal overlap problem. According to the 2D HSQC spectrum of **C**-lignin (fig. S1), Cα and Cβ have the potential to allow ^{13}C NMR quantification of the phenylpropanoid unit derived from caffeyl alcohol in the **C**-lignin (fig. S2). Cγ cannot be used for the quantification because of the signal overlap with the unknown peaks (δ_H, 4.00 to 4.35 ppm; δ_C, 60.0 to 62.5 ppm). The aromatic region of **C**-lignin cannot be used for the quantification because of the overlap with signals from protein residues (tyrosine and phenylalanine) (*45*). Cα and Cβ may seem equally good for the **C**-lignin quantification; however, when looking at the HSQC spectrum at a lower contour level, peaks from polysaccharide residues cannot be completely ignored even after the acidic LiBr pretreatment; the residual C_3 and C_5 of the cellulose overlap with the Cβ of the **C**-lignin. Because the relaxation reagent $Cr(acac)_3$ was added to reduce the experiment time, the line broadening caused by the relaxation reagent made the overlap between Cβ and the cellulose residues even worse. As a result, Cα was chosen for the quantification as it had minimal peak overlap issues. Assuming that **C**-lignin is derived from pure caffeyl alcohol, the detailed calculation was as shown below (table S2)

$$c_{C\beta} = \frac{c_{IS} \times 3}{A_{IS}} \times A_{C\beta}$$

$$Y_{CA} = \frac{c_{C\beta}}{\rho_{LBL}}$$

$$W_{lignin(LBL)} = Y_{CA} \times Mw_{CA} \times 100\%$$

$$W_{lignin(CW)} = \frac{W_{lignin(LBL)}}{LBL\%}$$

In the equations, c_{IS} (mmol/ml) is the molar concentration of internal standard (IS; 1,3,5-trioxane), A_{IS} is the peak integral of internal standard in the quantitative ^{13}C NMR spectrum, $c_{C\beta}$ (mmol/ml) is the molar concentration of caffeyl alcohol unit in the **C**-lignin polymer, $A_{C\beta}$ is the peak integral of Cβ in the quantitative ^{13}C NMR spectrum, ρ_{LBL} (mg/ml) is the

mass concentration of **C**-LBL sample, Y_{CA} (mmol/mg) is the mole amount of caffeyl alcohol (CA) per milligram of **C**-LBL, Mw_{CA} (mg/mmol) is the molecular weight of caffeyl alcohol, $W_{lignin(LBL)}$ is the weight percentage of **C**-lignin in **C**-LBL, LBL% is the weight percentage of **C**-LBL obtained from whole CW, and $W_{lignin(CW)}$ is the weight percentage of **C**-lignin in whole CW.

Lignin depolymerization methods

Alkaline NBO

NBO was performed as previously described (*46*). Dimeric model compound **D1** (5 mg) or extracted vanilla seed coat (40 mg) was mixed with nitrobenzene (0.4 ml) and 2 M NaOH (7 ml) in a 10-ml stainless steel reactor vessel (Taiatsu Techno Co.) and heated at 170°C for 2 hours in an oil bath. The reactor was then cooled in ice water, and 1 ml of freshly prepared ethyl vanillin (3-ethoxy-4-hydroxybenzaldehyde; 5 mg/ml) in 0.1 M NaOH solution was added to the reaction mixture as an internal standard. The mixture was transferred to a 100-ml separatory funnel and washed three times with 15 ml of DCM. The remaining aqueous layer was acidified with 2 M HCl until the pH was below 3.0 and extracted with 2 × 20 ml of DCM and 20 ml of diethyl ether. The combined organic layers were washed with DI water (20 ml) and dried over $MgSO_4$. After filtration, the filtrate was collected in a 100-ml pear-shaped flask and dried under reduced pressure. For the TMS derivatization step, NBO products were transferred with pyridine (3 × 200 μl) into a GC vial, and *N*,*O*-bis(trimethylsilyl)trifluoroacetamide [BSTFA; 100 μl] was added. The mixture was heated to 50°C for 30 min. The silylated NBO products were analyzed by GC-MS and quantified by GC-FID using calibration curves (fig. S3, A and B).

Thioacidolysis followed by Raney nickel desulfurization

Thioacidolysis was performed as previously described (*47*). The thioacidolysis reagent was prepared freshly by adding 2.5 ml of EtSH and 0.7 ml of BF_3 etherate to a 25-ml volumetric flask containing 20 ml of distilled 1,4-dioxane and then complemented with dioxane to exactly 25 ml. Freshly made thioacidolysis reagent (4.0 ml) was added to a 5-ml screw-cap reaction vial containing extractive-free CW (40 mg) or model compound (15 mg) and a magnetic stir bar. The vial cap was screwed on tightly, and the vial was kept in an oil bath containing a heating block at 100°C for 4 hours with stirring. After the reaction, the vial was cooled in an ice-water bath for 2 min. A solution of 4,4′-ethylidenebisphenol in dioxane was prepared and used as an internal standard. The product mixture was transferred to a separatory funnel and 10 ml of saturated $NaHCO_3$ solution, along with internal standard solution, was added. Then, 5 ml of 1 M HCl solution was added to adjust the pH of the solution to below 3. The aqueous layer was extracted three times with 20 ml of DCM, and the combined organic phase was washed with saturated NH_4Cl, dried over anhydrous $MgSO_4$, and evaporated under reduced pressure at 40°C. The resulting products were desulfurized via Raney nickel. Briefly, the thioacidolysis products were dissolved in 3 ml of distilled dioxane with 1 ml of Raney nickel 3202 (Sigma-Aldrich) slurry. The mixture was heated at 80°C for 2 hours. After the reaction, nickel powder was removed using a magnet, and the reaction mixture was transferred quantitatively with DCM into a separatory funnel charged with 10 ml of NH_4Cl and 10 ml of DCM. Then, 5 ml of 1 M HCl solution was added to adjust the pH of the solution to below 3. The aqueous layer was extracted twice with 10 ml of DCM, and the combined organic phase was washed with brine, dried over anhydrous $MgSO_4$, and evaporated under reduced pressure at 40°C. For the TMS derivatization step, products were transferred with pyridine (3 × 200 μl) into a GC vial, and BSTFA (100 μl) was added. The mixture was heated to 50°C for 30 min. The silylated thioacidolysis products were analyzed by GC-MS and quantified by GC-FID using calibration curves (fig. S3, C and D).

Hydrogenolysis

Hydrogenolysis was performed as previously described (*7*). In cases in which isolated **C**-LBL was used as a feedstock, 200 mg of **C**-LBL was dissolved in 30 ml of methanol or dioxane/water (9:1, v/v) or THF/water (96:4, v/v) in a 100-ml high-pressure Parr reactor along with 100 mg of catalyst (5 wt % Pt/C, Pd/C, or Ru/C). The reactor was stirred with a mechanical propeller and heated via a high-temperature heating jacket. Once closed, the reactor was purged three times and then pressurized with H_2 (40 bar, 4 MPa). The reactor was heated to the desired temperature and then held at that temperature for the specified residence time. After the reaction was completed, the reactor was cooled in a water bath to room temperature. The resulting liquid was filtered through a nylon membrane filter (0.8 μm, 47 mm; Whatman) and washed with EtOH. The solvent was removed under reduced pressure at 40°C with a rotary evaporator. The crude products were dissolved in EtOH and made up to 10 ml in a volumetric flask. A 1 ml of aliquot was transferred into three 5-ml vials and then dried under reduced pressure. The dried samples were used for GC, GPC, and NMR analyses. For GC sample preparation, the sample was dissolved in 0.9 ml of pyridine and 0.1 ml of BSTFA, incubated at 50°C for 30 min, and then subjected to GC-FID and GC-MS. For NMR sample preparation (fig. S4A), the sample was dissolved in 0.6 ml of DMSO-d_6/pyridine-d_5 (4:1, v/v) and then transferred to a 5-mm NMR tube for NMR. For GPC sample preparation (fig. S4B), the sample was dissolved in 1 ml of dimethylformamide (DMF) containing 0.1 M LiBr.

For the cases in which hydrogenolysis was performed directly on the CW material, 200 mg of preextracted vanilla seed coat was mixed with 30 ml of methanol and 100 mg of the catalyst (5 wt % Pt/C, Pd/C, or Ru/C). The remaining procedure was performed as described above.

For the cases in which the lignin model compound was used as the feedstock, a solution of 50 mg of dimer **D1** in 30 ml of methanol or dioxane/water (9:1, v/v) was mixed with 50 mg of the catalyst (5 wt % Pt/C). The remaining procedure was performed as described above.

Analytical methods

GC-MS qualitative analysis of low–molecular weight products

Samples were dissolved in pyridine, and BSTFA was added for TMS derivatization. The mixture was heated to 50°C for 30 min. An aliquot of the sample (1 μl) was injected by an autosampler into a GC-MS (GC2010/PARVUM2, IC-1 column, Shimadzu Co.) equipped with a fused silica capillary column (30-m × 0.25-μm film; SHR5XLB capillary column, Shimadzu Co.) operating in split mode (split ratio of 20:1) to identify the products. The products were identified by comparison with the peak retention times and mass spectra of the authentic compounds and (or) by comparing with entries in the National Institute of Standards and Technology mass spectral library (fig. S5).

GC-FID quantitative analysis of low–molecular weight products

The identified major products were quantified by GC-FID (GC-2014, Shimadzu Co.) using calibration curves derived from authentic synthetic compounds (table S3). The yields of major hydrogenolysis products catechylpropanol **M1** and catechylpropane **M2** were quantified by using the calibration curves generated from their authentic synthetic standards. The yields of minor products without a primary hydroxy group [chroman-6,7-diol **M3**, catechol **M4**, 4-methylcatechol **M5**, 4-ethylcatechol **M6**, and 4-(1-propenyl)catechol **M7**] were calculated by the effective carbon number (ECN) method based on the yield of catechylpropane **M2**,

whereas the minor product with a primary hydroxy group (caffeyl alcohol **M8**) was calculated on the basis of the yield of catechylpropanol **M1**. The theoretical ECN of TMS-derivatized catechol **M4** (10.0), 4-methylcatechol **M5** (11.0), 4-ethylcatechol **M6** (12.0), catechylpropane **M2** (13.0), 4-(1-propenyl)catechol **M7** (12.9), chroman-6,7-diol **M3** (12.0), catechylpropanol **M1** (15.5), and caffeyl alcohol **M8** (15.4) was used for the calculation. The ECN contribution of aliphatic carbon 1.0, aromatic carbon 1.0, olefinic carbon 0.95, primary alcohol −0.5, and TMS 3.0 was used as described (*7*, *17*, *48*). The detailed calculation was as follows

$$n_{\text{monomer}} = \frac{A_{\text{monomer}}}{A_{\text{M1 or M2}}} \times n_{\text{M1 or M2}} \times \frac{\text{ECN}_{\text{M1 or M2}}}{\text{ECN}_{\text{monomer}}}$$

$$n_{\text{CA}} = Y_{\text{CA}} \times m_{\text{LBL}}$$

$$Y_{\text{monomer}} = \frac{n_{\text{monomer}}}{n_{\text{CA}}} \times 100\%$$

In the equations, n_{monomer} (mmol) is the molar amount of monomer in each analyzed sample, A_{monomer} is the peak area of monomer in the GC-FID chromatogram, $n_{\text{M1 or M2}}$ (mmol) is the molar amount of **M1** or **M2** in each analyzed sample based on its calibration curve, $A_{\text{M1 or M2}}$ is the peak area of **M1** or **M2** in the GC-FID chromatogram, $\text{ECN}_{\text{monomer}}$ is the effective carbon number of monomer, $\text{ECN}_{\text{M1 or M2}}$ is the effective carbon number of **M1** or **M2**, n_{CA} (mmol) is the molar amount of caffeyl alcohol in the feedstock, Y_{CA} (mmol/mg) is the mole amount of caffeyl alcohol per milligram of **C**-LBL from the quantitative ^{13}C NMR analysis (table S2), m_{LBL} (in milligrams) is the weight of **C**-LBL in the feedstock, and Y_{monomer} is the yield of monomer based on the molar amount of caffeyl alcohol in the feedstock.

Analytical GPC

Molecular weight distributions of lignins were determined by GPC using a Shimadzu LC20-AD LC pump equipped with a Shimadzu SPD-M20A UV-vis detector set at 280 nm and a Polymer Standard Services GPC column and guard column [PSS PolarSil analytical Linear S, 8-mm inner diameter (ID) × 5 cm and 5-μm particle size → PSS PolarSil analytical Linear S, 8-mm ID × 30 cm and 5-μm particle size]. The samples and column compartment were held at 40°C during analysis. The mobile phase was DMF with 0.1 M LiBr, and the flow rate was 1 ml/min. Molecular weight distributions were determined using Wyatt ASTRA 7 software (Wyatt Technology Corporation) via a conventional calibration curve using a ReadyCal polystyrene kit from Sigma-Aldrich [catalog no. 76552, M(p) 250-70000].

GPC fractionation of hydrogenolysis product mixtures

Using LBL as a hydrogenolysis feedstock and dioxane/water as the solvent, the product mixture was dried in vacuo, redissolved in pure dioxane with sonication, filtered through a PTFE membrane (0.2 μm), and then subjected to GPC. The GPC conditions here were slightly different from those in the analytical GPC method. For the fractionation, dioxane was used as the mobile phase instead of 0.1 M DMF/LiBr solution at a slower flow rate (0.3 ml/min) to achieve better fractionation. Four fractions were separated and collected (fig. S7A). The ultraviolet (UV) absorption contour map showed that different molecular weight fractions had completely different UV absorption properties. Because of peak overlap, each fraction was characterized by using its 2D HSQC NMR spectra and subtracting the overlapped fractions' spectra (fig. S7, B to E; note that f2 was characterized by subtracting f1 and f3 from f2, f3 was characterized by subtracting f2 and f4 from f3, and f4 was characterized by subtracting f3 from f4]. As seen in the NMR spectra, peaks from some nonaromatic components appear in all fractions (f1 to f4). The molecular weight of these nonaromatic components cannot be measured accurately because of the low GPC resolution and peak tailing, and/or the possibility that these nonaromatic components have a wide molecular weight distribution. Fractions f1 and f2 were almost identical to each other and contained only traces of aromatic peaks. The highest molecular weight component(s) in the product mixture was therefore not from lignin but from other components in the seeds. Fraction f3 contained the major hydrogenolysis products **M1** and **M3**. This fraction exhibited the strongest UV absorption in the UV contour map, which means that it was the dominant aromatic-containing mixture in the product. Fraction f4 was the other major hydrogenolysis product **M2**, which has a slightly lower molecular weight compared with **M1** and **M3**. It is inferred that there was a large amount of high–molecular weight products (the products in f1 and f2), which are distributed from f1 to f4 because of peak overlap. As these products lack aromatic rings, they are not from the caffeyl alcohol–derived phenylpropanoid polymer. Thus, they must be produced from other components existing in the seed, such as waxes, fatty acids, etc. These observations support our conclusion that the lignin content of vanilla seed coats is not determined accurately by KL and other traditional lignin analytical methods because of these nonextractable, nonaromatic components.

TOC analysis

A TOC analyzer (TOC-VCPH, Shimadzu Co.) with a solid sample module (SSM-5000A, Shimadzu Co.) was used to determine the total carbon content of the vanilla seed coat material and its hydrogenolysis products, its fractions, and the nonvolatile products. The hydrogenolysis products were dried at 50°C for 30 min to remove EtOH and then dried at 50°C in a vacuum oven for 30 min to completely remove water and EtOH. The dried solid samples (20.00 ± 1.00 mg) and hydrogenolysis products were measured as solids.

Using LBL as a hydrogenolysis feedstock and dioxane/water as the solvent, there was no significant change in the carbon content before (62.66 ± 0.23 wt %) and after (61.64 ± 0.34 wt %) the reaction (±SD, n = 2). Solvent degradation products (for example, ethylene glycol, diethylene glycol, etc.) were detected in the product mixture and identified by GC-MS when dioxane was used as solvent. It is still possible that some components in the **C**-LBL can either become volatile or attach to the catalyst. However, considering that the volatile products (for example, methane, ethane, and hexane) have much higher carbon contents (~75 to 85 wt %) compared with the solvent degradation products (~35 to 45 wt %), the loss of volatile products while introducing solvent degradation products should cause a significant decrease of carbon content. In our experiment, we did not observe any carbon content decrease nor did we observe any weight increase of the catalyst. This result suggested that the loss of volatile products during work-up and the effect of the solvent degradation products were negligible and also implied that most of the carbon-containing compounds were retained in the product mixture.

Synthetic model compounds and compound authentication

Synthetic methods are fully described in the Supplementary Materials.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/9/eaau2968/DC1

Synthetic model compounds and compound authentication.

Calibration curves and NMR spectra.

Fig. S1. 2D HSQC NMR.

Fig. S2. Quantitative ^{13}C NMR spectrum of C-LBL.
Fig. S3. NBO and thioacidolysis products.
Fig. S4. 2D HSQC NMR and molecular weight distributions.
Fig. S5. GC-MS total-ion chromatograms of hydrogenolysis monomer products.
Fig. S6. Yield and selectivity data.
Fig. S7. GPC fractionation of hydrogenolysis products from LBL.
Table S1. Compositional analysis of vanilla seed coat CWs.
Table S2. Quantitative ^{13}C NMR analysis of **C**-lignin content in the **C**-LBL and CW.
Table S3. Monomer yields from hydrogenolysis.

REFERENCES AND NOTES

1. R. Rinaldi, R. Jastrzebshi, M. T. Clough, J. Ralph, M. Kennema, P. C. A. Bruijnincx, B. M. Weckhuysen, Paving the way for lignin valorisation: Recent advances in bioengineering, biorefining and catalysis. *Angew. Chem. Int. Ed. Engl.* **55**, 8164–8215 (2016).
2. Z. Sun, B. Fridrich, A. de Santi, S. Elangovan, K. Barta, Bright side of lignin depolymerization: Toward new platform chemicals. *Chem. Rev.* **118**, 614–678 (2018).
3. W. Schutyser, T. Renders, S. Van den Bosch, S.-F. Koelewijn, G. T. Beckham, B. F. Sels, Chemicals from lignin: An interplay of lignocellulose fractionation, depolymerisation, and upgrading. *Chem. Soc. Rev.* **47**, 852–908 (2018).
4. A. Rahimi, A. Ulbrich, J. J. Coon, S. S. Stahl, Formic-acid-induced depolymerization of oxidized lignin to aromatics. *Nature* **515**, 249–252 (2014).
5. C. S. Lancefield, O. S. Ojo, F. Tran, N. J. Westwood, Isolation of functionalized phenolic monomers through selective oxidation and C–O bond cleavage of the β-O-4 linkages in lignin. *Angew. Chem. Int. Ed. Engl.* **54**, 258–262 (2015).
6. E. Feghali, G. Carrot, P. Thuéry, C. Genre, T. Cantat, Convergent reductive depolymerization of wood lignin to isolated phenol derivatives by metal-free catalytic hydrosilylation. *Energ. Environ. Sci.* **8**, 2734–2743 (2015).
7. L. Shuai, M. T. Amiri, Y. M. Questell-Santiago, F. Héroguel, Y. Li, H. Kim, R. Meilan, C. Chapple, J. Ralph, J. S. Luterbacher, Formaldehyde stabilization facilitates lignin monomer production during biomass depolymerization. *Science* **354**, 329–333 (2016).
8. S. Van den Bosch, W. Schutyser, R. Vanholme, T. Driessen, S.-F. Koelewijn, T. Renders, B. De Meester, W. J. J. Huijgen, W. Dehaen, C. M. Courtin, B. Lagrain, W. Boerjan, B. F. Sels, Reductive lignocellulose fractionation into soluble lignin-derived phenolic monomers and dimers and processable carbohydrate pulps. *Energ. Environ. Sci.* **8**, 1748–1763 (2015).
9. Y. Shao, Q. N. Xia, L. Dong, X. Liu, X. Han, S. F. Parker, Y. Cheng, L. L. Daemen, A. J. Ramirez-Cuesta, S. H. Yang, Y. Q. Wang, Selective production of arenes via direct lignin upgrading over a niobium-based catalyst. *Nat. Commun.* **8**, 16104 (2017).
10. L.-P. Xiao, S. Wang, H. Li, Z. Li, Z.-J. Shi, L. Xiao, R.-C. Sun, Y. Fang, G. Song, Catalytic hydrogenolysis of lignins into phenolic compounds over carbon nanotube supported molybdenum oxide. *ACS Catal.* **7**, 7535–7542 (2017).
11. C. Zhang, H. Li, J. Lu, X. Zhang, K. E. MacArthur, M. Heggen, F. Wang, Promoting lignin depolymerization and restraining the condensation via an oxidation–hydrogenation strategy. *ACS Catal.* **7**, 3419–3429 (2017).
12. D. S. Argyropoulos, H. I. Bolker, Condensation of lignin in dioxane-water-HCl. *J. Wood Chem. Technol.* **7**, 1–23 (1987).
13. T. Yokoyama, Revisiting the mechanism of β-O-4 bond cleavage during acidolysis of lignin. Part 6: A review. *J. Wood Chem. Technol.* **35**, 27–42 (2015).
14. P. J. Deuss, M. Scott, F. Tran, N. J. Westwood, J. G. de Vries, K. Barta, Aromatic monomers by in situ conversion of reactive intermediates in the acid-catalyzed depolymerization of lignin. *J. Am. Chem. Soc.* **137**, 7456–7467 (2015).
15. V. M. Roberts, V. Stein, T. Reiner, A. Lemonidou, X. Li, J. A. Lercher, Towards quantitative catalytic lignin depolymerization. *Chemistry* **17**, 5939–5948 (2011).
16. R. Franke, C. M. McMichael, K. Meyer, A. M. Shirley, J. C. Cusumano, C. Chapple, Modified lignin in tobacco and poplar plants over-expressing the Arabidopsis gene encoding ferulate 5-hydroxylase. *Plant J.* **22**, 223–234 (2000).
17. W. Lan, M. Talebi Amiri, C. M. Hunston, J. S. Luterbacher, Protection group effects during α,γ-diol lignin stabilization promote high-selectivity monomer production. *Angew. Chem. Int. Ed.* **57**, 1356–1360 (2018).
18. Y. Mottiar, R. Vanholme, W. Boerjan, J. Ralph, S. D. Mansfield, Designer lignins: Harnessing the plasticity of lignification. *Curr. Opin. Biotechnol.* **37**, 190–200 (2016).
19. F. Chen, Y. Tobimatsu, D. Havkin-Frenkel, R. A. Dixon, J. Ralph, A polymer of caffeyl alcohol in plant seeds. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1772–1777 (2012).
20. F. Chen, Y. Tobimatsu, L. Jackson, J. Nakashima, J. Ralph, R. A. Dixon, Novel seed coat lignins in the Cactaceae: Structure, distribution and implications for the evolution of lignin diversity. *Plant J.* **73**, 201–211 (2013).
21. H.-M. Chang, E. B. Cowling, W. Brown, E. Adler, G. Miksche, Comparative studies on cellulolytic enzyme lignin and milled wood lignin of sweetgum and spruce. *Holzforschung* **29**, 153–159 (1975).
22. H. Kim, J. Ralph, Solution-state 2D NMR of ball-milled plant cell wall gels in DMSO-d_6/pyridine-d_5. *Org. Biomol. Chem.* **8**, 576–591 (2010).
23. N. Li, X. J. Pan, J. Alexander, A facile and fast method for quantitating lignin in lignocellulosic biomass using acidic lithium bromide trihydrate (ALBTH). *Green Chem.* **18**, 5367–5376 (2016).
24. Z. M. Xue, X. Zhao, R. C. Sun, T. C. Mu, Biomass-derived γ-valerolactone-based solvent systems for highly efficient dissolution of various lignins: Dissolution behavior and mechanism study. *ACS Sustainable Chem. Eng.* **4**, 3864–3870 (2016).
25. V. E. Tarabanko, D. V. Petukhov, G. E. Selyutin, New mechanism for the catalytic oxidation of lignin to vanillin. *Kinetics Catal.* **45**, 569–577 (2004).
26. C. Lapierre, B. Monties, C. Rolando, Preparative thioacidolysis of spruce lignin: Isolation and identification of main monomeric products. *Holzforschung* **40**, 47–50 (1986).
27. L. Berstis, T. Elder, M. Crowley, G. T. Beckham, Radical nature of C-lignin. *ACS Sustainable Chem. Eng.* **4**, 5327–5335 (2016).
28. S. Kim, S. C. Chmely, M. R. Nimos, Y. J. Bomble, T. D. Foust, R. S. Paton, G. T. Beckham, Computational study of bond dissociation enthalpies for a large range of native and modified lignins. *J. Phys. Chem. Lett.* **2**, 2846–2852 (2011).
29. J. Ralph, P. F. Schatz, F. Lu, H. Kim, T. Akiyama, S. F. Nelsen, in *Quinone Methides*, S. Rokita, Ed. (Wiley-Blackwell, 2009), vol. 1, pp. 385–420.
30. X. Y. Wang, R. Rinaldi, Solvent effects on the hydrogenolysis of diphenyl ether with Raney nickel and their implications for the conversion of lignin. *ChemSusChem* **5**, 1455–1466 (2012).
31. W. Schutyser, S. Van den Bosch, T. Renders, T. De Boe, S.-F. Koelewijn, A. Dewaele, T. Ennaert, O. Verkinderen, B. Goderis, C. M. Courtin, B. F. Sels, Influence of bio-based solvents on the catalytic reductive fractionation of birch wood. *Green Chem.* **17**, 5035–5045 (2015).
32. J. Sedó, J. Saiz-Poseu, F. Busqué, D. Ruiz-Molina, Catechol-based biomimetic functional materials. *Adv. Mater.* **25**, 653–701 (2013).
33. Q. Wu, D. Yan, Y. Chen, T. Wang, F. Xiong, W. Wei, Y. Lu, W.-Y. Sun, J. J. Li, J. Zhao, A redox-neutral catechol synthesis. *Nat. Commun.* **8**, 14227 (2017).
34. K. Barta, G. R. Warner, E. S. Beach, P. T. Anastas, Depolymerization of organosolv lignin to aromatic compounds over Cu-doped porous metal oxides. *Green Chem.* **16**, 191–196 (2014).
35. W. Wu, F. Liu, S. Singh, Toward engineering *E. coli* with an autoregulatory system for lignin valorization. *Proc. Natl. Acad. Sci U.S.A.* **115**, 2970–2975 (2018).
36. N. Li, Y. Li, C. G. Yoo, X. Yang, X. Lin, J. Ralph, X. Pan, An uncondensed lignin depolymerized in the solid state and isolated from lignocellulosic biomass: A mechanistic study. *Green Chem.* 10.1039/C2018GC00953H (2018).
37. A. Wagner, Y. Tobimatsu, L. Phillips, H. Flint, K. Torr, L. Donaldson, L. Piers, J. Ralph, *CCoAOMT* suppression modifies lignin composition in *Pinus radiata*. *Plant J.* **67**, 119–129 (2011).
38. R. Dixon, N. D'souza, F. Chen, M. Nar, U.S. Patent US9890480B2 (2018).
39. Y. Tobimatsu, F. Chen, J. Nakashima, L. Jackson, L. L. Escamilla-Treviño, R. A. Dixon, J. Ralph, Coexistence but independent biosynthesis of catechyl and guaiacyl/syringyl lignins in plant seeds. *Plant Cell* **25**, 2587–2600 (2013).
40. M. Montazeri, M. J. Eckelman, Life cycle assessment of catechols from lignin depolymerization. *ACS Sustainable Chem. Eng.* **4**, 708–718 (2016).
41. E. G. Bligh, W. J. Dyer, A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **37**, 911–917 (1959).
42. A. Sluiter, B. Hames, R. Ruiz, C. Scarlata, J. Sluiter, D. Templeton, D. Crocker, "Determination of structural carbohydrates and lignin in biomass" (Technical Report, NREL/TP-510-42618, National Renewable Energy Laboratory, 2012).
43. Z. Xia, L. G. Akim, D. S. Argyropoulos, Quantitative ^{13}C NMR analysis of lignins with internal standards. *J. Agric. Food Chem.* **49**, 3573–3578 (2001).
44. K. M. Holtman, H.-M. Chang, H. Jameel, J. F. Kadla, Quantitative ^{13}C NMR characterization of milled wood lignins isolated by different milling techniques. *J. Wood Chem. Technol.* **26**, 21–34 (2006).
45. H. Kim, D. Padmakshan, Y. Li, J. Rencoret, R. D. Hatfield, J. Ralph, Characterization and elimination of undesirable protein residues in plant cell wall materials for enhancing lignin analysis by solution-state NMR. *Biomacromolecules* **18**, 4184–4195 (2017).
46. Y. Li, T. Akiyama, T. Yokoyama, Y. Matsumoto, NMR assignment for diaryl ether structures (4–O–5 structures) in pine wood lignin. *Biomacromolecules* **17**, 1921–1929 (2016).
47. C. Lapierre, B. Pollet, B. Monties, C. Rolando, Thioacidolysis of spruce lignin: Gas chromatography-mass spectroscopy analysis of the main dimers recovered after Raney nickel desulfurization. *Holzforschung* **45**, 61–68 (1991).
48. J. T. Scanion, D. E. Willis, Calculation of flame ionization detector relative response factors using the effective carbon number concept. *J. Chromatogr. Sci.* **23**, 333–340 (1985).

Acknowledgments
Funding: Funding was provided by the U.S. Department of Energy (DOE) Great Lakes Bioenergy Research Center (DOE Biological and Environmental Research Office of Science

DE-FC02-07ER64494 and DE-SC0018409), the DOE Center of Bioenergy Innovation (DE-AC05-000R22725) and the Swiss Competence Center for Energy Research: Biomass for a Swiss Energy Future, through the Swiss Commission for Technology and Innovation grant KTI.2014.0116. We are also grateful to N. Li and X. Pan (University Wisconsin–Madison) for help with LiBr solubilization methods and to Y. Mottiar and S. Mansfield (University of British Columbia, Vancouver, Canada) and R. Vanholme and W. Boerjan (Vlaams Instituut voor Biotechnologie, Gent, Belgium) for discussions on designing lignins that resulted in recent papers on this topic. We are also grateful to valuable *Science Advances'* reviewer comments. **Author contributions:** J.R., J.S.L., H.K., Y.L., L.S., and J.A.D. were responsible for the conception, planning, and organization of the experiments. D.H.-F. provided the vanilla seed coat material. Y.T., F.C., and R.A.D. provided information and were involved in discussions relating to it. Y.L. performed the synthesis of dimer **D1** and the synthetic **C**-lignin and isolated lignins, performed the hydrogenolysis experiments with L.S., and isolated, quantified, and identified the products. Y.L. also performed NBO reactions and analysis and sugars analysis. L.S. and A.H.M. aided in the hydrogenolysis experiments. F.Y. performed the thioacidolysis analysis and helped with the model compound synthesis. H.K. helped Y.L. perform the quantitative ^{13}C NMR. J.K.M. helped with the catalyst considerations and carried out the GPC and molecular weight analysis. Y.L. and J.R. were responsible for NMR and MS data and interpretation. The manuscript was primarily written by Y.L. and J.R. with critical input from all coauthors. Figures were prepared by Y.L. and J.R. with support from H.K. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 26 May 2018
Accepted 22 August 2018
Published 28 September 2018
10.1126/sciadv.aau2968

Citation: Y. Li, L. Shuai, H. Kim, A. H. Motagamwala, J. K. Mobley, F. Yue, Y. Tobimatsu, D. Havkin-Frenkel, F. Chen, R. A. Dixon, J. S. Luterbacher, J. A. Dumesic, J. Ralph, An "ideal lignin" facilitates full biomass utilization. *Sci. Adv.* **4**, eaau2968 (2018).

Science Advances

An "ideal lignin" facilitates full biomass utilization

Yanding Li, Li Shuai, Hoon Kim, Ali Hussain Motagamwala, Justin K. Mobley, Fengxia Yue, Yuki Tobimatsu, Daphna Havkin-Frenkel, Fang Chen, Richard A. Dixon, Jeremy S. Luterbacher, James A. Dumesic and John Ralph

Sci Adv **4** (9), eaau2968.
DOI: 10.1126/sciadv.aau2968