The genetic basis of cytoplasmic male sterility and fertility restoration in wheat

Melonek et al.

	1	250	500	750	1,000	1,250	1,500	1,750	2,000	2,250	2,500	2,750	3,000 3,144
1. R0934F.300k_Assembly_Contig_78_candidate Rf3													
2. RFL29a													
3. RFL29b			ŀ	1	I								
4. RFL29b_captured		1 11 11	ŀ	1	1								
5. RFL29c				1 - 1									_

Supplementary Figure 1. Sequence alignment. Analysis of sequences present in RFL group 29. In this visualization, each vertical bar in the aligned sequences represents a mismatch compared to R0934F.300k_Assembly_Contig_78 sequence that was selected as *Rf3* candidate. An insertion in the putative 5' UTR of *RFL29b* is underlined in red and the open reading frames are shown in blue. Source data are provided as a Source Data file.



Supplementary Figure 2. Principal components analysis (PCA) of the read counts for each transcript from anthers collected at three developmental stages: early heading (A), early-mid anthesis (B) and late anthesis (C) in the fertile and sterile plants. n=3. The first principal component (pc1, 39% of variance) distinguishes the differentiation between the three stages in the fertile genotypes and the second principal component (pc2, 17% of variance) distinguishes strong differentiation between the fertile and sterile genotypes. *Rf1* corresponds to R0932E, *Rf3* to R0946E, *Rf1Rf3* to R0934F and *rf* to Fielder*CMS line. RNA-Seq counts and Python code used for plotting the figure are available from Dryad (https://doi.org/10.5061/dryad.6djh9w10d).

а



Supplementary Figure 3. Analysis of the fertility restoration in the *Rf1* and *Rf3* transformants compared to Fielder*CMS line. (a) Examples of the flowering phenotype of the transgenic plants. (b) Fertility restoration scores. The fertility scores were calculated by dividing the total number of seeds threshed from a spike by the number of counted spikelets. Centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. Number of spikelets analyzed per line: Fielder WT n=20777, Fielder*CMS n=2386, ZmUbiRFL79 n=5930, RFL79 n=1660, ZmUbiRFL29a n=3819, RFL29a n=3633, ZmUbiRFL29b n=1934, RFL29b n=809, *RFL29a*+*RFL79b* n=6943. Source data are provided as a Source Data file.



Supplementary Figure 4. Expression and processing of *orf256* in Fielder*CMS and *RFL79* and RFL29a transformants. (a) Schematic overview of the orf256 genomic structure. (b) RT-PCR analysis of the expression of *orf256* in different wheat genotypes. The experiment was performed three times and similar results were obtained. (c) Northern blot analysis of orf256 processing in mitochondria of several wheat accessions. The probe used to detect orf256 was prepared as described previously¹. The experiment was performed twice and similar results were obtained. (d) Survey of available T-CMS accessions in regard to orf256 processing. Several sterile lines including Alixan, Kalahari, Apache show processing of orf256 at cleavage site I that does not correlate with the restoration of fertility. Cleavage site II was detected only in T. timopheevii. Asterisks indicate CMS lines. M - RNA Markers. The experiment was performed twice and similar results were obtained. (e) Mapping the 5'-ends of orf256 RNA by 5'RACE analysis. Cleavage product I of *orf256* was amplified in CMS as well restorer lines. Cleavage product II was amplified only in T. timopheevii. The cleavage of orf256 does not correlate with the restoration phenotype observed for the RFL79 and RFL29a transformants. The amplification products were obtained with Gene Specific Primer 1 (GSP1) listed in Supplementary Table 6. M - 100 bp DNA ladder. The experiment was performed three times and similar results were obtained. Source data underlying Supplementary Figure 4b-e are provided as a Source Data file.

read-depth ratio CMS vs restored line



Supplementary Figure 5. Analysis of expression pattern of *orf124* between sterile and restored genotypes. (**a**) Ratio of strand-specific RNA-Seq coverage from Fielder*CMS lines (sterile) and restored (fertile) samples plotted across the *T. timopheevii* mitochondrial genome (NCBI accession number NC_022714.1) for forward and reverse strand is shown. The genomic region carrying *orf124* (reverse strand) is indicated at the top of the plots. (**b**) Normalised RNA-Seq coverage in the *orf124* region across the accessions. RNA-Seq counts and Python code used for plotting (**a**) and (**b**) are available from Dryad (<u>https://doi.org/10.5061/dryad.6djh9w10d</u>).



Supplementary Figure 6. Mapping cleavage sites of orf279 RNA by 5'-RACE approach. (a) 5'-RACE reaction products separated on 1.5% agarose gel and stained with ethidium bromide. For amplification the Gene Specific Primer 1 (GSP1) listed in Supplemental Table 8 was used. M: 100 bp NEB ladder. The experiment was performed three times and similar results were obtained. (b) Position of mapped cleavage sites within orf279 RNA identified by sequencing of the obtained 5'-RACE products. The sequence encoding the first 96 amino acid residues at the N-terminus of Orf279 corresponding to the ATP synthase subunit 8 encoded by the mitochondrial *atp8* gene is shown in bold. The blue and brown triangles point to cleavage sites identified in *RFL29a* and *RFL79* transformants, respectively. The numbers within the triangles indicate the number of clones sequenced to contain RNA starting at this position. The predicted binding sites for RFL29a and RFL79 are underlined in blue and brown, respectively. Source data underlying Supplementary Figure 6a are provided as a Source Data file.



Supplementary Figure 7. In vitro RNA-binding assays with recombinant RFL29a and RFL79 proteins. The RFL29a and RFL79 proteins were expressed as N-terminal fusions with Maltose Binding Protein (MBP) and Histidine-tag (His-tag). The recombinant proteins were purified on Ni-NTA columns (Bio-Rad) by gravity flow. (a) To analyse the state of the recombinant proteins (ratio of soluble fraction vs aggregates), the purified proteins were subjected to sizeexclusion chromatography (SEC) using an ENrich[™] SEC 650 column (Bio-Rad) on an NGC Chromatography System (Bio-Rad). The size of separated proteins detected by 280 nm absorbance was estimated using size exclusion standards (Bio-Rad): thyroglobulin Mr 670,000 (S1), bovine γ-globulin Mr 158,000 (S2), chicken ovalbumin Mr 44,000 (S3), equine myoglobin Mr 17,000 (S4). The approximate retention of the standards is indicated. The SEC analysis revealed the presence of high-molecular weight protein aggregates of RFL29a or RFL79 eluting first followed by protein fractions with estimated molecular weight similar to free MBP. The experiment was performed once. (b) The fractions collected during the run with a BioFrac fraction collector (Bio-Rad) were separated on a Mini-PROTEAN TGX Stain-Free gel (Bio-Rad). SDS-PAGE confirmed that the protein aggregates seen during the SAC analysis represent recombinant RFL79 protein. Input - Ni-NTA purified RFL79 protein fraction before SEC analysis, F4 - fraction containing high-molecular weight aggregates, F15 - fraction containing the low molecular weight protein. M - Molecular mass standard. The experiment was performed once. (c) SDS-PAGE of protein fractions used in all REMSA experiments. The experiment was performed once. (d) To study the specificity of RNA binding, the purified RFL29a and RFL79 were incubated with each other's predicted RNA targets. Both proteins bind to the other's RNA target as shown by REMSA assay. Serial protein dilutions ranging from 1.8 µM to 28.1 nM were used for RFL29a and RFL79. The final concentration of the RNA probes was 1 nM. B = bound - RNA-protein complex, U = unbound - free RNA probe. On each gel, the left lane acts as a marker for unbound probe. The experiment was performed three times and similar results were obtained. Source data are provided as a Source Data file.

а

Ideath.	1 19	zņ	30	49	60	eņ	79	80	90	100	110	120	190	140	160	190
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	ATGCCCCGCTI	CTCCTCCACCA CTCCTCCACCA	CGCCAATGTCGC	CACCCCGCCT	CCGCCTCCGACT		CACTCCTO	CCTCCACCTCI	CATCCCTCACG	CATCIGGGAI	CCCCACGCCC	CCTTCGCCGC	GCGGCNCAGC GCGNCGCAGC	GGGCG CGCTC	IGGCACGCT IGGCACGCT	CACCACGG CACCACGG
Identity	110	145		290 	10	240		240	210	240	10	240	240	, .	12	240
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	AGGACGCACAC AGGACGCACAC 310	CACCTGTTTGA CACCTGTTTGA 340	GAATTGCTGCG GAATTGCTGCG 350 36	GCGGGGCAAT GCNGGGCAAT	CCTGTCCAGGAG CCTGTCCAGGAG	CGTCCCTTGA CGTCCCTTGA 390	TAANTT CTAACTT	CTGGCTGCCC CTGGCTGCCC	CTCGCCCGCGCGCG CTCGCCCGCGCGCG 10 420	CCCGCGTCCG CCCGCGTCCG 430	CATOCTGC CATOCTGC 440	CGATGGCCCC CGATGGCCCC 450	GCCCTGGCNGT GCCCTGGCOGT 440	CGCCCTCTTO CGCCCTCTTO 479	3G C C G T T T G 3G C C G T T T G 460	TCCCGAGA TCCCGAG
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	CGCCGGACGAC CGCCGGACGAC	CGGGTGGCGCAG CGGGTGGCGCAG 10 510	CCAAATGTCTTC CCAAATGTCTTC	ACCTATGGCG ACCTATGGCG 500	TCCTCATGGACT TCCTCATGGACT 540	GCTGCTGCCG GCTGCTGCCG	GCTTCCC GCCGCCC 80	CGCCACAGATC1 CGCCTGGATC1 579	PGG GGTCGCC1 PNG CGNTCGCC1 540	TCTT GCCCG	TCTCCTCAAC TCTCCTCAAC	ACGGGCCTGG ACGGGACTGG 610	AGGCAAACCAA AGGCAAACCAA 620	GTCGTCTTCA GTCATCTTC 630	CACCCTCC CACCCTCC	TCAAGGG TCAAGGGA 650
Identity			ور في المراجع													
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	CTOTGCCACAC CTOTGCCACGC	CAAAGCGGGCGG CAAAGCGCTCAG ¢70	ATGAGGCTOTGG ATGAGGCTOTGG	ACGTG TGCT ACGTG TGCT	TCACAGGATGCC TCACAGGATGCC	TGAGCTGGGC TGAGCTNGGC 710	TGCACCC TGCACCCC	CAAGTGGTG RAAGTGGTGGTG 730	GCCTATAACAC GCCTATACCAC 740	GGTCATCCA GGTCATCCAC 750	GGCTTCTT GGCTTCTT 760	AAGGAAGGCCA AAGGAAGGCCA 100	GTANGCAAGG NGTAGCAAGG 140	COTGCAATCT CATGCAATCT 790	ATTCCATGA ATTCCATGG	AATGGCGC AATGGCGC 810
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	AGCAGGGCGTI AGCAGGGCGTI 820	FAAGCCTAATGT FGCGCCTGATGT 830	GGTGACATATAA GGTGACATATAA 840	CTCNGTTATC CTCGGTTATC	CGATGCGCTGTGC CGATGCGCTGTGC 860 870	AAGGCCAGAG AAGGCCAGAG 890	CATGGAC	CAAGGCAGAGG CAAGGCAGAG 890	add TCCTTCGT add TCCTTCGT add 91	GANATGOTTG ANATGOTTG 8	ATCATGGTG1 ATMATGGTG1 0 8	NGGACCTGAT NGGACCTNAT NO 94	AATGTGACGTA AATGTGACNTA 940	TAGTAGCCTC TANTAGCCTC 960	ATCCATGGA ATCCATGGA 970	TATTCCTC
Identity													000000000000000000000000000000000000000			
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	TTCAGGCCAC TTCGGGCCAC 100	GGAAGGAGGCA AGAAGGAGGC 990 1.00	GTTAGGGTATIG GTTAGGGTGCTG 0 1,010	AAAGAGATGA AAAGA N ATGA 1,000	ICAAGOCGGAGGG ICAAGNCAGGGAA 1,050	TGA ACCAGA TGA ACCAGA 1,010	TG TCATT/ TG TCATT/ 1,050	ACCTGCACC 1,000	ACTOATGACCT	TTCCTTTGCAA TTCCTTTGCAA 1,000	GNATGGAAG GNATGGAAN 1,990	LAG CAA N G AAG LAG CAA <mark>G</mark> G AAG 1,100	CTGCAGCAATT CTGCAGNAATT 1.110	TTTGATACCA TTTGATCAA 1,120	IGGCCACCA IGGCCACCA 1,190	AGGGCCTG AGGGCCTG 1.40
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	AAACOTGACA AAACATGACG 1,160	CGTTTCATATG CGTTTCATATG 1,160	CTATTCTCCTTC CTATTCTCCTTC 1,179	ATGGGTATGC ATGGGTATGC 1,100	CNCCGAAGGATG 1,190	CTTGGTTGAT 1,200	ATGATTAJ ATGATTAJ 1,210	ATCTCTTCAA1 ATCTCTTCAA1 1,220	FTCCATGGAAAG FTCCATGGACAG 1,230	AGACTGTATI 1,240	CTACCTNACS 1,250	IGTCSTATCTT IGTCNTATCTT 1,260	CAACATACTGA CAACATACTGA 1.270	TTRATGCATA TTRATGCATA 1,200	IGCTAAATC IGCTAAATC 1,290	TGGGAAGC TGGGAAGC 1,300
Identity																
 R0934F-300k_Assembly_Contig_120_1 R197.300k_Assembly_Contig_120_1 	TTGATAAGGC1 1,310	TATGCTTATOTT TATGCTTATNTT 1,320	AGAGAGATGCA 1,330	GAAACAAGGA GAAACAAGGA 1,340	IG TG AG CCAGAT IG TG AG CCAGAT 1,350 1,3	GCASTCACAT GCASTCACAT 60 1.3	ATTCAACO	TAATACATO TAATACATO 1,800	CATTTTG AAG CATTTTG AAB	AAGGGTAGG1 AAGGGTOGG1 1,400 1	TGGACGATG0 410	TATGATAAAG CTATGATAAAG 1,420 1,	TTTAATCAGAT TTTAATCAGAT 430 1,4	GATTGATACA GGTTGATACA 0 1,45	GAGTACGA GAGTACGA	CAGGGCAC 460
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	AGCTGTTTAT AGCTGTTTAT 1,470	GTCCCCTAATC GTCCCCTAATC 1,490 1	CAGGGTTTTTGT CAGGGTTTTTGT 490 1.500	ACACA ACACA 50 1,510	ATTTGG TGAAAG ATTTGG TGAAAG 1,520	GAAAGGAAT GAAAGGAAT 1,530	GTTACTO GTTACTO 1,540	GAAATGATGAJ GAAATGATGAJ 1,550	AGAAAGGTATGC AGAAAGGTATAC 1,560	CTCCTCCTGA CTCCTCCTGA 1,570	ATTATGTTO ATTATGTTO 1,580	CTTC AG TTCAA CTTC CA TTCAA 1,590	TCATGCAGAAC TCATGCAGAAC 1,600	CTATG CACAG. CTATG CACAG. 1,610	AAGGAAGGG AAGGAAGGG 1,620	TA AC AGAA TA GT AGAA 1,6%
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	GCACGGGATAT GCACGGGATAT 1,44	CCTTGACTTGA CCTTGCCTTGA 1,460	TAG CACATAG TAG CACACATAG 1,660	GTATGAGGCC GTATGAGGCC 1,670	TAATG TTACA TAATG TT TG CAC 1,680	ATTTAATAT ATTTAATATA 1,600	CTGATCGG CTGATCGG 1,790	STGGATACTGO STGGATACTGO 1,710	CCTAGTCCGCAA CCTAGTCCGCAA 1,720	AGATGG CA GAT AGATGG AG GAT 1,730	GCATEGAAAA GCCTCAAAAA 1,740	TATTTGATGA TATTTCATGA 1,750	TATGGTGTCAT TATGNTGTCAT 1,760	ATGGTTTAGA ATGGTTTAGA 1,70	ACCTT <mark>G</mark> TAA ACCTT G TAA 1,780	ICTTACGT 1,790
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	ATGGTATACTI ATGGTATACTI 1,800	PATTAATGGCTA PATTAATGGCTA 1,810	TTGCAAAAA TTGCAAAAACAG 1,1420	AAGGATTGAT AAGGATTGAT 1,490	GACGGGCTOATT GACGGGCTGATT 1,840	CTGTTCAAAG CTGTTCAAAG	AGATGOTO ANATGOTO	GCACAAGGGAC GCGCAAGGGAC 1,470	CTTAAACCTACA CTTAAACCTACA 1,890	ACTTTTAATT ACTTTTAATT 1,890	ACAACGTCA ACAACATCA 1,900	PACTGGATGGA PACTGGATGGA 1,910	TTATTTCTGGC TTATTTCTGGC 1,920 1	TGGACAAACT TGGACGAACT \$90 1,	STTGCTGCA STTGCTGCA HO	AAAGAGAA AAGGAAAAA 1,950
Identity																
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	GTTTGATGAGA GTTTGATGAGA 1,990	ATGGTTGAATCT ATGGTTGAATCT 1,870	GGAGTAAGTG GGAGTAAGTNTG 1,980 1.88	TGCAT CAG TA TGCAT CAG TA 10 2.0	ICATACTCTATAA ICTACTCTATAG 80 2,019	TTCTTCGTGG TTCTTCGTGG 2,029	ACTTTGT/ ACTTTGT/ 24	AGAAATAGCTO AGAAATAATTO X00 2.0	5 CAG CAAG 3 AG CGAAG 40 2,050	CGATCACCT CCATCACGCT 2,002	ATTCCGGA ATTCCAGAC 2,070	ATTAAGCGCAA ATTAAGCGCAA 2,000	TG AATG TG AAA TGGATG TG AAA 2,090	TT CA ATATTA TT CA ATATTA 2,190	AATTGTCA AATTGTCA 2,110	ATATCAT <mark>S</mark> ATATCAT <mark>S</mark>
Identity			ندی وی ا													
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	ATTGGTGCCTT ATTGNTGCCTT 2,120 2,1	PATACAGGGTCG POTECAGGGTCC 30 2.140	AGAGAAACCAAG AGCGAAACCAAG 2,150	AGCTAAGGA ANGCTAAGGA 2,190	TTTGTTTGCTGC TTTGTTTGCTGC 2,170	ATCCAGCC ATACAGCC 2,180	AATGGGT1 AATGGGT1 2,190	rggTT <mark>C</mark> CTAA1 rggTT <mark>G</mark> CTAA1 2,200	IGOTGTTACCTA IGOTGTTACCTA 2,210	ACACCGTAATG ACAGCCTAATG 2,220	ATGACAAATO ATGACAAATO 2,230	CTTATAAAAGA CTTATAAAAAGA 2240	AGG TCAG TGG AGG CCAG TGG 2,250	AAGA <mark>A</mark> GCTGA AAGA <mark>G</mark> GCTGA 2,260	CACACTOTT CACACTOTT 2,270	TTATCCA TTATCCA 2,280
Identity																L L
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	TGGAGAAGAGC TGGAGAGAGC 2,290	CGGCTGTACTGC CGGCTGTACTC 2,309	GAACTCTCGTGCAT GAACTCGTGGAT 2,319	GTTAAATCAT GTTAAATCAT 2.320	ATOATCAGANGG ATOATCAGAGGG 2,330	TTACTGGAAA TTGCTGGAAA 2.240	LA AGG AG AG LA AGG AG AG 2,160	GATAGTCAAGO GATAGTCAAGO 2,800	GCTGGAAATTA1 GCTGGA GG TTA1 2,370	ATGTCTAAAG ATGTCTAAAG 2,299	TTGATGCAA TTGATGCGAI	AGAGCTACTCA AGAGCTACTCA	CTTGAAGCTAA CTTGAAGCTAA	AACTG TTTCG AACTG TTTCG	TGCTGATC	TOTCTOTT TATCTOTT
Identity							ĺ.									
1. R0934F.300k_Assembly_Contig_78_1 2. R197.300k_Assembly_Contig_120_1	TTCA <mark>R</mark> GGAAAG TTCA <mark>G</mark> GGAAAG	GGAAATA T AGA GGAAATA C AGA	GAACACAT C A N A GAACACAT N A G A	TTGCT CCTA	CAAAGTATCAGT CAAAGTATCAGT	TTCT <mark>C</mark> GAAGA	AGCAGCCI AGCAGCCI	ACAG TTGAAT ACAG TTGAAT	G GTTTGCTATAT	'G A						

b

)	1								10									20								:	io								4	,								60	, ,		
RFL79 RFL29a	M M	P I P I	R F R F	SS	SS	T T	T T	P P	M M	S F S F	P	R R	L L	R	L L	R R	L L	G Z	A R	H	SS	55	SS	T	SS	H I H I	P S P S	R	I	W	D	P P	H H	A I A I	A E A E	A	A A	A A	TA	00	R R	A A	R	5 G	T	L	T
RFL79 RFL200	Т	E	DA	H	H	60 L	Fp	D	E	ĿĮ	R	Q.	G	N	Pp	V	00	EH	P	Ļ	T	N	Fri	ερ L	A	A	LA	R	5 A D	Pp	A	S	яр А	F		PPR	G	P	A	L	A	100 V	A	LE	G	R	L
RFL258		Б	11	o PP	R		ŕ		35			120	G	14	E	,	~	5 1		10	130	K	ŕ	11	^	PP	R	. K	-	140	^	5	^	5			9	Ľ	150	7	^	35	<u> </u>	P	PR	5	
RFL79 RFL29a	S 160	R 📕	G A D V	G	R R	R R	V V	A A	0 0 1	P 1 P 1 70	I V	Eri Eri	T T	Ϋ́Υ	G G	V	L	M I M I 1	0 0	CC	C C	R R	A A	R C	R R	T	D L D L 19	v	L	A A	lin lin	Er fr	A G	R I R I		К К 20	PPR	GG	L L	E	A A	N N	00 T	7 1 7 V	F F 210	C N	T
RFL79 RFL29a	L L	L I	K G K G	L	СС	H H	A T 220	K	R 📕	S I A I) E E	A A	L L	D	V V	V L 230	L L	H F	R M	I P I P	EE	L L	GG	C	T 240	P I P I	35 N V N V	V	A A	Y Y	5 T N	T T	V V	I 1 I 1 150	H G	F	F	K	토토	GG	0 H	V V	G 1 S 1	K A K A	c c	N N	L L
RFL79 RFL29a	FF	H H	G M E M	1 A 1 A 270	0.0	0iQ	GG	VV	A K	2 1 2 1 2 1	S V V	V V	Т Т 250	¥ Ÿ	5 N N	S	V V	II) A) A	L L	СС	К К 290	A A	R	A A	M I M I	D K	AA	테더	Y V	F V 300	L L	R R	<mark>2</mark> 1 Q 1	M V M I	D	D	GG	V V	V G 310	P P	35 N D	N Y N Y	/ 1 / 1	Y Y Y	5 N S	SS
RFL79 RFL29a	L L	I I 320	H G H G	Y Y	52 52	22 22	S	GG	H H	D P W P 33	E E P	A A	V V	R	V	F	K K	E 1 E 1	1 I 1 I 39	RS	QR	GR	I	I T	P A	35 D 1 D 1	V H	Т Т 350	C Y	5 T N	S M	Life	M M	T	F I F I	C	Р К К 360	N	G	K R	C2 C2	ĸ	E i	A A A A	G	⊥ ⊥ 370	FF
RFL79 RFL29a	DD	ST	M A M A	T	K K	GG	L L	К К 3ș0	H P	D A	V	55	Y Y	A A	I	L L	L L 390	H (H (y Y	A	TA	E	GG	C	L L	R V 1 V 1 фо	D M D M	II	N N	L L	[Inc]Inc	N N	SS	M M 4	D R E F 10	D	СС	I	L	P P	35 N D	c 📕	H R 4	PPR I F I F 20	5 N N	I	L L
RFL79 RFL29a	I	Y N	A Y A Y	A A	К К 430	C2 C2	GG	K	L I	D B	A A	M M	L L	⊥ ⊥ 440	fing fing	R N	D E	M (M (0 K	010	G G	V V	S S 450	P P	35 D N	A A	F T	Y Y	<mark>5</mark> 53 53	T T	L V	I I 460	H	A A	F C F C	R K	K K	GG	R R	L L	D D 470	D D	A A	M I 7 I	K	Eta Eta	N N
RFL79 RFL29a	00	M M	V D I D	999) T) T	R G G	V V	R R	Q P	35 G D	T A A S	V V 490	ŶŶ	5 G R	S P	L	I	00	G I G I	r c	T T Sộc	HH	G G	D D	L	V V	K 📕	G K A K	E E	L ¥ 510	V V	T T	[1] [1]	M M	M M	N K	G	I M	P P 520	P P	P P	35 D D	I	M M	PPR F F	5 H S	SS	⊥ ⊥ 530
RFL79 RFL29a	M M	QQ	N L N L	C C C	T	E	GG	R R	V V 540	V E	A A	R	D D	I	L	G D	L L	I I șo	H	I	GG	M M	R	p, p,	35 N N	V I	<mark>С Т</mark> І І қо	F. Fr	5 N N	I	L L	Ī	G G	G G	Y C Y C s7	PPR	V V	R R	K K	M M	<mark>E</mark> A	D D	A A	5 K K 59	U U U	F	H D
RFL79 RFL29a	DD	M M	M S V S	Y Y	GG	L L 590	E	P P	35 S C	N N	T	Y Y	5 G G	Ĩ	L L 600	I	N N	G 1	c c c	K	N N	R R	R R	eio I	D D	D (D (G L G L	I	L L	file file	K	ЫE	M M 620	L I	R K	G	L L	K	P, P	35 T T	T T	F F 630	N N	ин У N У N	I	I	L L
RFL79 RFL29a	DD	G G	L F L F 64	L L	A A	GG	R Q	T T	V V	A A A A	K K	Е Е 650	K	Eri Eri	D D	E	M M	V 1 V 1	s s	G	olio ∆ ∆	SS	M V	35 C C	I	S D	T Y T Y	5 S S	I	V I e70	L L	R G	G G	L (L (C R	PR N	N S	CC	S S 80-	GS	E	A A	I	F L F L	F	Q Ř	TK
RFL79 RFL29a	L L 690	S S	A M A M	PPR 1 D 1 N	V V	ĸĸ	FF	35 N D	I I 7	R 1 T 1	V	5 N N	I	M I	I	D G	A A	F I	R R	V V	Q E	R R	K N	PR Q Q	[4] [4]	A 1 A 1	К D К D 72	, L	Ing Ing	A A	A A	I M	T P	A I A I	N G	L L 73	VV	A P	35 N N	V A	F V	T T	Y Y	R 5 5 1 7 V	M M 740	M M	ŢŢ
RFL79 RFL29a	N N	L L	I K	E	GG	SS	V V 750	E E	E	A D A D) T N	L	[in] fin	L L	SS	M M 760	E	M si	G G	СС	T T	SA	35 N N	SS	W C 770	M	L N L N	L	Ī	I	R R	G R	L L	L I L I	E K	G	EE	I	V V	K	A A	G G	C N nșo	Y N Y N 79	1 S 1 S 2	ĸ	V V
RFL79 RFL29a	D D	A I A	K S K S	Y Y	SS	L L	35 E E	A A	K K	T V T V	7 S 7 S	L L	L L	I	Y S	L L	FF	S 📕	K	G	K	¥ ¥	R R	EE	H H	I	R L K L	L L	PP	T T	K	Y Y	00	F 1 F 1	LE	E	A A	35 A A	T	V V	E	W	F	A I			

Supplementary Figure 8. Comparison between *RFL79* (*Rf1* candidate) and *RFL29a* (*Rf3*) candidate DNA sequences (**a**) and encoded protein sequences (**b**). The sequence of mitochondrial targeting peptide is highlighted by yellow box, PPR motifs by green boxes, and amino acid residues at position 5 and 35 of each PPR motif by blue boxes. Source data are provided as a Source Data file.



Supplementary Figure 9. *Orf279* as genetic basis of cytoplasmic male sterility in wheat T-CMS plants. (a) Alignment of the promoter and 5' UTR regions of *T. aestivum atp8-1* gene, *T. timopheevii* full-length *atp8* gene, *orf279*, and *T. aestivum atp8-2* gene (GenBank accessions numbers: AP013106/ NC_022714 and AP008982). (b) Phylogenetic relationships between chimeric *atp8*-orfs related to CMS in plants. Alignment was generated with Geneious (https://www.geneious.com) and tree with FastTree. The GenBank accession numbers for the CMS-genes are as follow: *orf138*² (P68513.1), *orf125*³ (AB015327), *orfB*⁴ (AB033490), *orf224* and *orf222*⁵ (U10423 and U10428), *orfH522*⁶ (CAA37614.1), *orf279* (this manuscript), *atp8* (NC_022714). Source data are provided as a Source Data file.

Supplementary Table 1. Markers identified in genetic screen as associated with fertility restoration conferred either by *Rf1* or *Rf3* locus in wheat.

Marker ID	Restorer	SNP	location	strand	START	STOP	Sequence
cfn0522096	Rf1	G/C	chr1A	[-]	14505133	14505203	ATGCAAAGTAGTACTCGTAGAGAGTTAACACAGAC[G/C]AGTGATTTATTGGGTGGTATTCTACTTGATATTTG
cfn0527067	Rf1	A/G	chr1A	[-]	15346055	15346425	GACAATATGATTCACCCTAGATCCTTCACCTTACA[A/G]TTCGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
cfn1249269	Rf3	A/G	chr1B	[+]	17780363	17780422	CGTTTAAAAGAACACAAATGTGGCCCTAGTGATCA[A/G]GTACACATATTTGTCACCTCTTTGAATCTTACTTA
BS00090770	Rf3	T/C	chr1B	[-]	20589170	20589229	TAGCCGTAGGTCGTAGCACATAGCCGTTTA[T/C]GTAATGCATAGTTGTCCGAAGGAATGTTTC

Interval	# RFL	Gene name	Location	START	STOP	Strand	Encoded protein	Protein length (aa)	Predicted location ⁷	Comment
	RFL77	TraesCS1A02G031600.1	chr1A	14577582	14579933	[+]	Pentatricopeptide repeat-containing protein (PPR)	784	m (0.853, mTP 20)	full-length
	RFL105*	TraesCS1A02G031700.1	chr1A	14584912	14587137	[+]	Pentatricopeptide repeat-containing protein (PPR)	741	m (0.845, mTP 23)	truncated
		TraesCS1A02G031800.1	chr1A	14589565	14594115	[-]	Protein kinase, putative			
		TraesCS1A02G031900.1	chr1A	14603494	14615692	[-]	Protein kinase, putative			
		TraesCS1A02G032000.1	chr1A	14744093	14748608	[-]	O-methyltransferase			
		TraesCS1A02G032100.1	chr1A	14790721	14791908	[-]	Chalcone synthase			
<i>Rf1</i> (14.5-15.3 Mbp)		TraesCS1A02G032200.1	chr1A	14799110	14800511	[-]	O-methyltransferase			
1		TraesCS1A02G032300.1	chr1A	14838308	14839886	[-]	Chalcone synthase			
		TraesCS1A02G032400.1	chr1A	14857039	14858282	[-]	O-methyltransferase			
		TraesCS1A02G032500.1	chr1A	14988541	14988944	[+]	alpha/beta-Hydrolases superfamily protein			
		TraesCS1A02G032600.1	chr1A	15186553	15188366	[+]	Anthocyanidin synthase			
		TraesCS1A02G032600.2	chr1A	15177031	15177339	[+]	Anthocyanidin synthase			
		TraesCS1A02G032700.1	chr1A	15190979	15195959	[+]	Replication protein A 70 kDa DNA- binding subunit			
	RFL60	TraesCS1B02G038200.1	chr1B	17894195	17896564	[-]	Pentatricopeptide repeat-containing protein (PPR)	790	m (0.910, mTP 25)	full-length
	RFL164	TraesCS1B02G038300.1	chr1B	17964601	17966241	[-]	Pentatricopeptide repeat-containing protein (PPR)	546	m (0.883, mTP 14) (pseudogene)	truncated
	RFL396	TraesCS1B02G038400.1	chr1B	18091427	18093253	[-]	Pentatricopeptide repeat-containing protein (PPR)	608	none (pseudogene)	truncated
	RFL29b	TraesCS1B02G038500.1	chr1B	18116277	18118649	[-]	Pentatricopeptide repeat-containing protein (PPR)	790	m (0.899, mTP 25)	full-length
<i>Rf3</i> (17.8-20.6 Mbp)	RFL252	TraesCS1B02G038600.1	chr1B	18363378	18364565	[-]	Pentatricopeptide repeat-containing protein (PPR)	396	none (pseudogene)	truncated
1 /		TraesCS1B02G038700.1	chr1B	18385559	18387800	[-]	Nitrate transporter 1.1			
		TraesCS1B02G038800.1	chr1B	18419377	18423130	[-]	Protein kinase			
		TraesCS1B02G038900.1	chr1B	18570322	18571755	[-]	Cysteine protease			
		TraesCS1B02G039000.1	chr1B	18578571	18579939	[+]	Glycosyltransferase			
	RFL89	TraesCS1B02G039100.1	chr1B	18683381	18685732	[+]	Pentatricopeptide repeat-containing protein (PPR)	783	m (0.936, mTP 18)	full-length

Supplementary Table 2. RFL genes identified in the *Rf1* and *Rf3* genomic regions in the IWGSC RefSeqv1.0 genome of Chinese Spring.

RFL58	TraesCS1B02G039200.1	chr1B	18867715	18870069	[+]	Pentatricopeptide repeat-containing protein (PPR)	784	m (0.874, mTP 19)	full-length
	TraesCS1B02G039300.1	chr1B	19018793	19020813	[+]	Oligopeptidase A			
	TraesCS1B02G039400.1	chr1B	19,025,038	19025620	[+]	Leucine-rich repeat protein kinase family protein			
	TraesCS1B02G039500.1	chr1B	19038922	19043415	[+]	Receptor protein kinase, putative			
	TraesCS1B02G039600.1	chr1B	19047105	19051706	[+]	Protein kinase, putative			
RFL97f*	chr1B:R:19063406- 19062072 TraesCS1B02G039625	chr1B	19061029	19063406	[-]	Pentatricopeptide repeat-containing protein (PPR)	445+363	m (0.946, mTP 16) (pseudogene)	truncated
RFL97	TraesCS1B02G039625	chr1B	19073841	19076117	[-]	Pentatricopeptide repeat-containing protein (PPR)	759	m (0.715, mTP 22)	full-length
	TraesCS1B02G039800.1	chr1B	19217861	19232577	[-]	Receptor-like protein kinase			
	TraesCS1B02G039900.1	chr1B	19314295	19314714	[-]	Defensin			
	TraesCS1B02G040000.1	chr1B	19718626	19720046	[-]	Chalcone synthase			
	TraesCS1B02G040100.1	chr1B	19779389	19780576	[-]	Chalcone synthase			
	TraesCS1B02G040200.1	chr1B	19890445	19891967	[-]	Chalcone synthase			
	TraesCS1B02G040300.1	chr1B	19950229	19951416	[-]	Chalcone synthase			
	TraesCS1B02G040400.1	chr1B	20015264	20017794	[-]	O-methyltransferase-like protein			
	TraesCS1B02G040500.1	chr1B	20046035	20047787	[-]	Chalcone synthase			
	TraesCS1B02G040600.1	chr1B	20059838	20061082	[-]	O-methyltransferase			
	TraesCS1B02G040700.1	chr1B	20167177	20169022	[+]	Anthocyanidin synthase			
	TraesCS1B02G040800.1	chr1B	20172786	20177655	[+]	Replication protein A 70 kDa DNA- binding subunit			

No.	Tribe	Species	Cultivar	Website	References	Identified RF-like PPR proteins
1		Aegilops tauschii (1)	AL8/78	http://plants.ensembl.org/Aegilops_tauschii/Info/Index	8	46
2		Aegilops tauschii (2)		http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	8	31
3		Aegilops speltoides		http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	9	20
4		Aegilops sharonensis		http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	9	32
5		Triticum aestivum		http://plants.ensembl.org/Triticum_aestivum/Info/Index	9	98
6		Triticum durum	cv. Cappeli	http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	9	30
7		Triticum durum	cv. Strongfiled	http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	9	30
8	Tritiana	Triticum monococcum		http://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies	10	12
9	Triticeae	Triticum turgidum		http://www.ncbi.nlm.nih.gov/bioproject/PRJNA191054	11	25
10		Triticum urartu		http://archive.gramene.org/Triticum_urartu/Info/Annotation/	12	24
11		Hordeum vulgare	Barke	http://pgsb.helmholtz-muenchen.de/plant/barley/index.jsp	13	13
12		Hordeum vulgare	Morex	http://pgsb.helmholtz-muenchen.de/plant/barley/index.jsp	13	12
13		Hordeum vulgare	Bowman	http://pgsb.helmholtz-muenchen.de/plant/barley/index.jsp	13	13
14		Hordeum vulgare	var. distichum	http://plants.ensembl.org/Hordeum_vulgare/Info/Index	13	0
15		Lolium perenne	P226/135/1	http://www.ncbi.nlm.nih.gov/nuccore/GAYX0000000	15	2
16		Secale cereale		http://pgsb.helmholtz-muenchen.de/plant/rye/index.jsp	15	0
17		Oryza sativa		http://phytozome.jgi.doe.gov/pz/portal.html#!bulk?org=Org_Osativa	16	14
18		Oryza sativa japonica		http://plants.ensembl.org/Oryza_sativa/Info/Index	17	14
19		Oryza indica	Nipponbare	http://plants.ensembl.org/Oryza_indica/Info/Index	18	18
20		Oryza sativa Nipponbare		http://rapdb.dna.affrc.go.jp/index.html	17	14
21		rice cultivar 'Kasalath'	Kasalath	http://rapdb.dna.affrc.go.jp/index.html	18	11
22		Oryza barthii		http://plants.ensembl.org/Oryza_barthii/Info/Index	19	13
23	Oryzeae	Oryza brachyantha		http://plants.ensembl.org/Oryza_brachyantha/Info/Index	20	4
24		Oryza glaberrima		http://plants.ensembl.org/Oryza_glaberrima/Info/Index	21	2
25		Oryza glumaepatula		http://plants.ensembl.org/Oryza_glumaepatula/Info/Index	19	12
26		Oryza meridionalis		http://plants.ensembl.org/Oryza_meridionalis/Info/Index	19	15
27		Oryza nivara		http://plants.ensembl.org/Oryza_nivara/Info/Index	19	13
28		Oryza punctata		http://plants.ensembl.org/Oryza_punctata/Info/Index	19	7
29		Oryza rufipogon		http://plants.ensembl.org/Oryza_rufipogon/Info/Index	19	12
30	Brachypodieae	Brachypodium distachyon		http://phytozome.jgi.doe.gov/pz/portal.html#!bulk?org=Org_Bdistachyon	22	11

Supplementary Table 3. Summary of plant genomes used in the study.

31	Eragrostideae	Eragrostis tef		http://www.tef-research.org/index.html	23	33
32	Poniceae	Setaria italica		http://phytozome.jgi.doe.gov/pz/portal.html#!bulk?org=Org_Sitalica	24	15
33	Andropogoneae	Sorghum bicolor	BTx623	http://plants.ensembl.org/Sorghum_bicolor/Info/Index	25	20
34	Andropogoneae	Sorghum bicolor	BTx623	http://www.ncbi.nlm.nih.gov/	25	11
35	Andropogoneae	Zea mays	B73	http://plants.ensembl.org/Zea_mays/Info/Index	26	6
					Total	633
				Reference RFLs	27	49
				Sorghum WGS data sets	28	517
					Total	1199

#	Restoration status	Accession name	Number of assembled contigs composed of > 100 reads	Number of identified RFL ORFs >210 aa*	Number of orthologous groups with at least one RFL from the accession	Number of RFL ORFs > 350 aa assigned to orthologous groups
	weak Rf3					
1	restorer	Chinese Spring	204	216	205	161
2	maintainer	Anapurna	211	221	202	156
3	maintainer	Fielder	231	223	212	138
4	Rf1	R197	219	241	219	174
5	Rf1	R0932E	221	245	221	183
6	Rf3	R0946E	239	262	237	171
7	Rf3+RF1	R0934F	223	237	215	174
8	Rf3	Primepi	226	234	215	162
9	Rf1	Triticum timopheevii	138	143	129	114
					397 (non-	
	Total		1912	2022	redundant)	1433

Supplementary Table 4. Summary of the RFL capture experiment.

Supplementary Table 5. Selection of candidate RFL groups based on restoring status of analysed wheat accessions and location within a given interval.

		Cono	Gene		•		Restoring	g genotype	~		
		located	located within <i>Rf1</i>	Maintainer		Rf1 1	restorer	<i>Rf</i> 3 re	estorer	Rf1 + Rf3	<i>Rf1</i> restorer
RFL gene	Protein size (aa)	interval (<i>in silico</i> mapping with tblastn)	mapping interval (genetic mapping (nb of markers))	Chinese Spring	Anapurna	R197	R0932E	R0946E	Primepi	R0934F	Triticum timopheevii
RFL1	988	no	n.a.	0	0	1*	1	0	0	0	0
RFL56	804	no	n.a.	0	0	1	1*	0	0	0	1
RFL59	813	no	n.a.	0	0	1	2	0	0	0	1
RFL73	813	no	n.a.	0	0	1	1	0	0	0	1
RFL74	813	no	n.a.	0	0	1	1	0	0	1	0
RFL79	808	no	yes (4)	0	0	1	1	0	0	1	1
RFL93	775	no	no	0	0	1	1	0	0	0	0
RFL104	757	yes	yes (4)	0	0	1	1	0	0	1	1
RFL129*	693	no	no	0	0	1	1	0	0	1	0
RFL185*	524	yes	yes (4)	0	0	1	1	0	0	1	1
RFL268*	382	yes	yes (4)	0	0	1	1	0	0	1	1

Selection of candidate RFL groups based on restoring status of analysed wheat accessions and location within the *Rf1* interval.

* - non-functional sequence truncated or disrupted by a frameshift.

		Cono	Gene				R	estoring g	enotype			
RFL	Protein	located within <i>Rf3</i>	located within <i>Rf3</i> mapping	weak <i>Rf3</i> restorer line	Maintainer	Rf1 i	restorer	<i>Rf3</i> re	storer	Rf1 + Rf3	<i>Rf1</i> restorer	Maintainer
gene	size (aa)	(<i>in silico</i> mapping with tblastn)	interval (genetic mapping (nb of markers))	Chinese Spring	Anapurna	R197	R0932E	R0946E	Primepi	R0934F	Triticum timopheevii	Fielder
RFL67	820	yes	yes	0	0	0	0	1	1	1	0	0
RFL89	801	no	yes (7)	1	0	0	0	1	1	1	0	1
RFL140*	637	no	yes (4)	0	0	0	0	1	1	1	0	0
RFL164*	391	yes	yes (6)	1	0	0	0	1	1	1	0	0
RFL166*	568	no	yes (2)	0	0	0	0	1*	1*	1*	0	0
RFL252*	396	no	yes (9)	1	0	0	0	1	1	1	0	0
RFL28	824	no	no	1*	1*	1*	1*	1	1	1	0	0
RFL29	790	yes	yes (7)	1	0	0	0	1	1	1	0	1*
RFL170	560	no	no	1*	1*	1*	0	1	1	1	0	1

Selection of candidate RFL groups based on restoring status of analysed wheat accessions and location within the *Rf3* interval.

* - non-functional sequence truncated or disrupted by a frameshift

Supplementary Table 6. List of oligos used in the study.

Oligo Name	Experiment	Sequence 5'->3'	Reference
Ta_Actin_F	qRT-PCR	GCCACACTGTTCCAATCTATGA	
Ta_Actin_R	qRT-PCR	TGATGGAATTGTATGTCGCTTC	
P1 (<i>orf256</i> + <i>coxI</i>)	qRT-PCR	ATGACAAATATGGTTCGATGGC	
P2 (orf256)	qRT-PCR	GCTTGGGGATCCTGAATC	
P3 (<i>coxI</i>)	qRT-PCR	GCTGTCACTAGAACGGACC	
orf256_GSP1	5'-RACE	GATTACGCCAAGCTTAAAACAGATCCTCCCTCATTCTCCCGCAGA	
orf279_GSP1	5'-RACE	GATTACGCCAAGCTTCATTACGTCCTCGACAACTCCACC	
RFL29a_attB1_f	REMSA	GGGGACAAGTTTGTACAAAAAGCAGGCTTCATGCCCCGCTTCTCCTCCAC	
RFL29a_attB2_r	REMSA	GGGGACCACTTTGTACAAGAAAGCTGGGTCCTATTCAACTGTGGCTGCTTCTTCCAGA	
RFL79a_attB1_f	REMSA	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCATATAGCAAACCATTCAACTGTGGCTG C	
RFL79a_attB2_r	REMSA	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCATGTTCTGCAGCGATGGCCCT	
BS RFL29a	REMSA	CUUCACCUGUGCACUUAUUUAUGUAU	
BS RFL79	REMSA	CUGUAAAUCAAGAAUUCCUCGAAGA	
BS orf256	REMSA	CCACUAGCAGGUUUACUGCUUUCU	
WORF256_211_806_for	Nothern Blot	ATCCCCAAGCTCTAGCTCATTTAG	1
WORF256_211_806_rev	Nothern Blot	GGGGGCTGGAAGAAAAGAAT	1
RFL capture_qPCR_for	To confirm untargeted sequence depletion (chloroplas t genome sequence)	TTTGGTTTCAAAGCCCTACG	
RFL capture_qPCR_rev	To confirm untargeted sequence depletion	AACTTGGATACCATGAGGCG	

	(chloroplas		
	t genome		
	sequence)		
ShBar_for	Genotypin	GCACCATCGTCAACCACTACA	
	g of		
	transgenic		
	plants		
ShBar_rev	Genotypin	GTCCACTCCTGCGGTTCCT	
	g of		
	transgenic		
	plants		
ShBar_TaqMan probe	Genotypin	FAM-CACGGTCAACTTCCGTAC-MGB-NFQ	
	g of		
	transgenic		
	plants		
GaMyb_for	Genotypin	GATCCGAATAGCTGGCTCAAGTAT	
	g of		
	transgenic		
	plants		
GaMyb_rev	Genotypin	GGAGACTGCAGGTAGGGATCAAC	
	g of		
	transgenic		
	plants		
GaMyb_TaqMan probe	Genotypin	5'VIC-CGTGGCTCCTGCGATGCAGC-TAMRA	
	g of		
	transgenic		
	plants		

Supplementary references

- 1. Song, J. S. & Hedgcoth, C. Influence of nuclear background on transcription of a chimeric gene (*orf256*) and *coxI* in fertile and cytoplasmic male-sterile wheats. *Genome* **37**, 203-209 (1994).
- 2. Bonhomme, S., Budar, F., Ferault, M. & Pelletier, G. A 2.5 kb NcoI fragment of Ogura radish mitochondrial DNA is correlated with cytoplasmic male-sterility in Brassica cybrids. *Curr Genet* **19**, 121-127 (1991).
- 3. Iwabuchi, M., Koizuka, N., Fujimoto, H., Sakai, T. & Imamura, J. Identification and expression of the kosena radish (*Raphanus sativus* cv. Kosena) homologue of the ogura radish CMS-associated gene, orf138. *Plant Molecular Biology* **39**, 183-188 (1999).
- 4. Nakajima, Y., Yamamoto, T., Muranaka, T. & Oeda, K. A novel orfB-related gene of carrot mitochondrial genomes that is associated with homeotic cytoplasmic male sterility (CMS). *Plant Molecular Biology* **46**, 99-107 (2001).
- 5. LHomme, Y., Stahl, R. J., Li, X. Q., Hameed, A. & Brown, G. G. Brassica nap cytoplasmic male sterility is associated with expression of a mtDNA region containing a chimeric gene similar to the pol CMS-associated orf224 gene. *Curr Genet* **31**, 325-335 (1997).
- 6. Kohler, R. H., Horn, R., Lossl, A. & Zetsche, K. Cytoplasmic male-sterility in sunflower is correlated with the co-transcription of a new open reading frame with the *Atpa* Gene. *Mol Gen Genet* **227**, 369-376 (1991).
- 7. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* **300**, 1005-1016 (2000).
- 8. Jia, J. *et al.* Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91-95 (2013).
- 9. Mayer, K. F. X. *et al.* A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345** (2014).
- 10. Brenchley, R. *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**, 705-710 (2012).
- 11. Krasileva, K. V. *et al.* Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol* **14**, R66 (2013).
- 12. Ling, H. Q. *et al.* Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87-90 (2013).
- 13. Mayer, K. F. X. *et al.* A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**, 711-716 (2012).
- 14. Farrell, J. D., Byrne, S., Paina, C. & Asp, T. *De novo* assembly of the perennial ryegrass transcriptome using an RNA-Seq strategy. *PLoS ONE* **9**, e103567 (2014).
- 15. Martis, M. M. *et al.* Reticulate evolution of the rye genome. *Plant Cell* **25**, 3685-3698 (2013).
- 16. Ouyang, S. *et al.* The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res* **35**, D883-D887 (2007).
- 17. Kawahara, Y. *et al.* Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**,4 (2013).
- 18. Sakai, H. *et al.* Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* **54**, e6 (2013).
- 19. Jacquemin, J., Bhatia, D., Singh, K. & Wing, R. A. The International Oryza Map Alignment Project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Current Opinion in Plant Biology* **16**, 147-156 (2013).

- 20. Chen, J. F. *et al.* Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nature Commun* **4**, 1595 (2013).
- 21. Wang, M. *et al.* The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat Genet* **46**, 982-988 (2014).
- 22. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).
- 23. Cannarozzi, G. *et al.* Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* **15**, doi:10.1186/1471-2164-15-581 (2014).
- 24. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant Setaria. *Nat Biotechnol* **30**, 555–561 (2012).
- 25. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556 (2009).
- 26. Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115 (2009).
- 27. Fujii, S., Bond, C. S. & Small, I. D. Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *P Natl Acad Sci USA* **108**, 1723-1728 (2011).
- 28. Mace, E. S. *et al.* Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun* **4**, 2320 (2013).