**Exploration of the *T. cacao* genome sequence to decipher the incompatibility system of *Theobroma cacao* and to identify diagnostic markers**

Claire Lanaud1†, Olivier Fouet1†, Thierry Legavre1†, Ulison Lopes2†, Olivier Sounigo1, Marie Claire Eyango4, Benoit Mermaz1, Marcos Ramos da Silva2, Rey Gaston Loor Solorzano3, Xavier Argout1, Gabor Gyapay6, Herman Ebai Ebaiarrey4, Kelly Colonges1, Christine Sanier1, Ronan Rivallan1, Géraldine Mastin3, Nicholas Cryer7, Michel Boccara1, Ives Bruno Efombagn Mousseni4, Karina Peres Gramacho2, Didier Clement1

† These authors contributed equally to this study.

1- Centre de Cooperation Internationale en Recherche Agronomique pour le Developpement (CIRAD), UMR AGAP, F-34398 Montpellier, France.

2-Centro de Pesquisas do Cacau (CEPEC), CEPLAC, Rod. Ilhéus-Itabuna, km 22, Ilhéus, BA 45605-350, Brazil

3-Centre de Cooperation Internationale en Recherche Agronomique pour le Developpement (CIRAD), UR Bioagresseurs, BP 2572 Elig-Essono, Yaoundé, Cameroon

4-Institut de Recherche Agricole pour le Developpement (IRAD) PO Box 2123, Yaoundé, Cameroon

5-Instituto Nacional de Investigaciones Agropecuarias (INIAP), EET-Pichilingue, CP 24 Km 5 via Quevedo El Empalme, Quevedo, Los Ríos, Ecuador

6-Commissariat à l’Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France

7-Mondelez UK R&D Limited, PO Box 12, Bournville Place, Bournville Lane, Birmingham, B30 2LU, United Kingdom

**Abstract**

We explored the *Theobroma cacao* genome sequence to progress in the knowledge of the *T. cacao* incompatibility system. Cocoa self-compatibility is an important yield factor and has been described as controlled by a late gameto-sporophytic system involving several locus, and resulting in gametic non-fusion. In this work, we identified two different mechanisms controlling the *T. cacao* self-incompatibility system at two separate loci, located on chromosome one and four (CH1 and CH4). Both loci are responsible for gametic selection, but only one (the CH4 locus) is involved in the main fruit drop. The CH1 locus acts prior to gamete fusion and independently of CH4 locus. Fine mapping and genome wide association studies focused analyses of restricted regions without recombinant plants where several candidate genes were identified. Their expression analysis showed differential expression during incompatible or compatible reactions for some of them. Highly polymorphic SSR diagnostic markers, designed in the CH4 region identified by fine mapping, allowed the development of efficient diagnostic markers predicting self-compatibility and fruit setting according to allele or genotype presence. SSR alleles specific to self-compatible Amelonado and Criollo varieties were also identified allowing screening for self-compatible plants in cocoa populations.

**Introduction**

One important *T. cacao* yield factor is the self-compatibility status of cocoa trees. Self-compatible (SC) young trees would produce an average of 66% more fruits than the self-incompatible (SI) trees (Cope, 1939a). A higher proportion of self-compatible trees was also observed among higher producing trees (Lachenaud et al., 2005). Various mechanisms of self-incompatibility (SI), preventing the self-fertilisation of plants, have been described in angiosperms (de Nettancourt, 1997; Takayama et al., 2005; Rea et al., 2008). These mechanisms act at the level of pollen growth inhibition for most species (gametophytic (GSI) and sporophytic (SSI) self-incompatibility systems) (McClure, 2008; Suzuki et al. 1999), or at the ovary level (late acting self-incompatibility system (LSI), (Gibbs, 2014)), leading to an embryo development failure. The SI mechanism in *T. cacao* has been described as a LSI system: in SC and SI reactions the pollen can reach the ovary with a similar rate of pollen tube growth 4h after pollination (Bouharmont, 1960), but the double fertilization is completed after 24h in SC reactions, whereas in SI reactions male nuclei are released into the embryo sac, but fail to achieve gamete fusion in part of the ovules, resulting in floral abscission (Cheesman, 1927; Cope, 1939b, 1940, 1958, 1962; Knight and Rogers 1955; Posnette, 1940; Bouharmont, 1960; Ford et al., 2012). The genetic control of *T. cacao* SI was studied by several authors (Cope, 1939b, 1940, 1958, 1962; Knight and Rogers,
who hypothesized the existence of a S-locus and several alleles with dominance relationships between them, including an amorphous Sf allele, present in the SC Amelonado variety and leading to self-compatibility when homozygous. In addition, two other loci involved in T. cacao self-incompatibility were also hypothesized (Cope, 1962), based on cytological studies showing 25%, 50% or 100% of ovules without gamete fusion, after selfing of SI cocoa trees. It was concluded that the genetic system of SI in cacao has both aspects of sporophytic and gametophytic controls (Cope, 1958).

Two loci putatively involved in self-incompatibility observed by the % of fruit setting, were localised by QTL analysis at the top of chromosome 4 (Crouzillat et al., 1996; Royaert et al., 2010; Yamada et al., 2010), and in chromosome 7 (Yamada et al., 2010).

The gamete fusion failure, observed in part of the ovules and associated to fruit abscission, was observed only at the embryo sac level (Cope, 1939b; Bouharmont, 1960). However, fruits can reach maturity after self-pollination of a SI tree when compatible pollen is mixed with pollen from the SI trees (“mentor pollen” effect) (Opeke and Jacob, 1967; Bartley, 1969, 2005; Lanaud et al., 1987; Glendinning, 1960), allowing the development of fertilized ovules. With the help of molecular markers, the observations, among these progenies, of skewed segregations potentially linked to SI is then possible.

Here, we report the analysis of a large F2 progeny produced from a SI tree using the “mentor pollen” effect. A first molecular analysis, performed on a subsample of this F2 progeny, revealed a skewed segregation at the level of CH1 and CH4 regions (Allegre et al., 2012). In this work, we carried out fine mapping of these genome regions where candidate genes were identified and their expression characterised during SC and SI reactions. In order to refine the existing predictive model for self-compatibility (Da Silva et al., 2016), multi-allelic diagnostic SSR markers were searched in the CH4 region identified by fine mapping.

**Materials and Methods**

**Material**

**Mapping population used for fine mapping**: A large progeny was produced by CEPLAC, by selfing the clone TSH516, a hybrid of ICS 1 and Scavina 6 (SCA 6). Until 877 individuals from this progeny were analysed. Self seeds were produced with mixed pollinations using Herrania mariae pollen as mentor pollen, followed by pollination with TSH516 self-pollen.

**Analyses of potential skewed segregations in 2 other progenies**

- A progeny of 550 individuals was created at IRAD (Barombi-kang, Cameroon) from IMC 60 (SI) using a mixture of pollens of IMC 60 and of Catongo as mentor pollen. Self progenies were identified using molecular markers.
- A progeny of 96 plants from the cross UF 676 x ICS 95, planted in French Guyana, was analysed to observe the segregations in the CH1 locus.

**Cocoa populations used for genome wide association study (GWAS) and to analyse prediction level of diagnostic markers**

A population of 710 individuals evaluated for self-incompatibility, and from different origins (a farm and a breeding populations from IRAD (Cameroun) breeding populations from CEPLAC (Brazil) and INIAP (Ecuador), collections from INIAP, CRC (Trinidad and Tobago) and CIRAD (France)) was used to assess the efficiency of prediction based on SSR genotypes and alleles. A subset of 570 individuals from these populations was analysed by GWAS using SNP markers revealed by GBS (genotyping by sequencing).

**Samples for gene expression analyses**

Pollinations were carried out on the SCA 6 SI clone, present in the greenhouse of CIRAD/Montpellier, using pollen from ICS 1 (SC) or from SCA 6 (SI), and ovaries with their pistil (named “ovaries” in the text) were collected at different times after pollination. QPCR experiments were conducted on three biological repetitions of each RNA extract from these samples to analyse gene expression.

**Methods**

**Genotyping**

**F2 progeny**: new markers were defined in the CH1 and CH4 regions:

SSR markers were defined using the whole sequence of the Criollo genome V1 (http://cocoaegdb.cirad.fr/gbrowse/cgi-bin/gbrowse/theobroma/) (Argout et al., 2011), and the tool « Search for SSR » integrated in the ESTTIK database (http://esttik.cirad.fr/cgi-bin/SSR_server.cgi). Deletions/insertions were identified in the CH4 region between the Criollo genome and the Amelonado genome (http://www.cacaogenomedb.org/) after alignment of both sequence fragments.

**GWAS population**: The GWAS population was genotyped by sequencing (GBS) using the DArTseq technology after DNA restriction with PvuI and MseI (Killian et al., 2012).

**Genome wide association studies**: 

---

2
GWAS was conducted using Tassel 5.2.31 software (Glaubitz et al., 2014) on 570 individuals assessed for SC/SI coded as 0 (SI) and 1 (SC) and on a subsample of 388 individuals, using the % of fruit setting 14 days after self-pollination, taken as a quantitative variable to assess SC/SI status. These individuals were genotyped using 16480 SNPs (GBS) stored in the TropGENE-DB (http://tropgenedb.cirad.fr/tropgene/) and with a minor allele frequency > 0.05. The structure of the population was determined with a subset of 150 SNPs distributed over all chromosomes, and a bayesian clustering method implemented in the STRUCTURE software (Pritchard et al., 2000), with a burning period of 100,000 iterations, 500,000 Markov Chain Monte Carlo repetitions and ten independent runs.

**Search for candidate genes and gene expression analyses**

The expression of candidate genes, identified in the CH1 and CH4 regions using the Criollo cocoa genome sequence, version V1, available at genome (http://cocoagendb.cirad.fr/tools.html) (Argout et al., 2011), was analysed during SC and SI reactions carried out on the SCA 6 accession. Q-PCR were conducted using a Roche LightCycler 480 Real-time PCR System and a SYBR Green dye included in the supermix to detect dsDNA amplification products, and with two reference genes used for normalization: Tc04_g000050 (Isocitrate dehydrogenase) and Tc08_g003640 (Tubulin beta-6 chain).

**Immunolocalization of Tc01_g007270 and Tc01_g007290 proteins**

Rabbit polyclonal antibodies, anti-Tc01_g007270 and anti-Tc01_g007290, were produced by Eurogentec (Anti-peptide Speedy 28-Day https://secure.eurogentec.com/speedy.html) using synthetized peptides as antigen. Peptides sequences LGNDKTVRIWTQENE, corresponded to residues 310-324 of Tc01_g007270 protein and RSVDKSNDSESVS corresponded to residues 478-492 of Tc01_g007290 protein. Immunolocalization was revealed with a Alexa Fluor 488 dye conjugated goat anti-rabbit antibodies (Interchim, France, Montluçon). The microscope imaging was performed in Montpellier RIO Imaging Platform (http://www/mri/cnrs.fr) with a confocal microscope (LSM510, Meta; Carl Zeiss Micro Imaging).

**Identification of diagnostic markers**

Eleven SSR markers (Table 1), identified in the CH1 and CH4 regions, were used to genotype a population of 710 individuals, and to establish predictions for SC/SI according to the genotype or allele, considered alone or in combination with one or two other markers.

**Table 1:** SSR or INDEL markers defined in the CH1 and CH4 regions and used to establish predictions of SC or SI cocoa plants

<table>
<thead>
<tr>
<th>Markers</th>
<th>Chromosome</th>
<th>Position</th>
<th>type</th>
<th>5'-3' forward primer</th>
<th>Tm</th>
<th>5'-3' reverse primer</th>
<th>Tm</th>
<th>PCR product size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mS1 103</td>
<td>CH1</td>
<td>4021626</td>
<td>[I][I]</td>
<td>ATCTGTCGACATCCTTC</td>
<td>55,1</td>
<td>TCAAGGGATCACTCCAAAA</td>
<td>55,2</td>
<td>299</td>
</tr>
<tr>
<td>mS4 107</td>
<td>CH1</td>
<td>4130575</td>
<td>[A][A]</td>
<td>GAGAGCTACCGGACAGAACA</td>
<td>54,4</td>
<td>ACCTTGGCAGACACACACA</td>
<td>54,6</td>
<td>221</td>
</tr>
<tr>
<td>mS1 7</td>
<td>CH4</td>
<td>20673</td>
<td>[AG][A]</td>
<td>TCTGAGGAGGCTGAGGAG</td>
<td>55,5</td>
<td>TGGGCAAGAGATTGGG</td>
<td>55,7</td>
<td>183</td>
</tr>
<tr>
<td>mS1 2</td>
<td>CH4</td>
<td>43494</td>
<td>[I][A]</td>
<td>ATCTGAGGATACAGAAGAAG</td>
<td>55,1</td>
<td>ATTGAGATGTTGGTGGT</td>
<td>55,1</td>
<td>268</td>
</tr>
<tr>
<td>mS1 3</td>
<td>CH1</td>
<td>119995</td>
<td>[I][I]</td>
<td>CAGATGCTACTTCGAGGG</td>
<td>55,7</td>
<td>AAAGTTCAATCCATTTC</td>
<td>55,6</td>
<td>265</td>
</tr>
<tr>
<td>mS1 458</td>
<td>CH4</td>
<td>136890</td>
<td>[A][I]</td>
<td>GCAGCAAGAAGATCTCAGCA</td>
<td>59,3</td>
<td>TGCAAGCGGATGATCTT</td>
<td>59,3</td>
<td>284</td>
</tr>
<tr>
<td>mS1 460</td>
<td>CH4</td>
<td>139970</td>
<td>[C][B]</td>
<td>TGAGGCAAGCAAGCAAGAAGA</td>
<td>58,7</td>
<td>CCGAGCAAGCAAGCAAGAG</td>
<td>58,7</td>
<td>117</td>
</tr>
<tr>
<td>mS1 315</td>
<td>CH4</td>
<td>233706</td>
<td>[C][C]</td>
<td>CAGAGGGAGCTGAGTGGTAG</td>
<td>55,2</td>
<td>ATGGTGCCGAGTTGGG</td>
<td>55,2</td>
<td>206</td>
</tr>
</tbody>
</table>

Allelic frequencies and identification of alleles specific to Amelonado and Criollo SC varieties were search in a subset of 108 T. cacao genotypes, capturing the diversity of the T. cacao genetic groups. Reference SSR profiles allowing cocoa breeders to characterise clones for potential SC/SI status were established for these SSR (Table 2).

Prediction analyses were carried out according to genotype or allele presence, using SAS software modules (SAS Institute Inc., 2004) using Fisher’s Exact Test and PROC FREQ. Probabilities.

**Table 2:** Newly references SSR profiles were established for a collection of diverse T. cacao clones available from international germplasm collections, and which could be used as standards for prediction of self-compatibility status of cocoa trees.
Study of skewed segregations linked to the incompatibility system

Fine mapping of restricted genome regions including the skewed segregations in the F2 progeny:

• In the skewed CH1 and CH4 regions, respectively 22 and 26 new markers segregating in the F2 progeny could be defined. The expected a (homozygous)/h (heterozygous)/b (homozygous) genotype segregations for this F2 was 1/2/1, whereas only two genotypes were predominant in these regions, with the “a” Amelonado genotype totally absent in a small region of 70.3 Kb of CH1, and the scavina 6 allele (b) totally absent in a region of 257.2 Kb of CH4 (Fig 1).
Fig 1: Segregations observed in the F2 progeny at the level of the CH1 and CH4 genome regions.

Skewed segregations observed in other progenies (Table 3).

Self-progeny from IMC 60: among the 39 self-fertilized seeds identified from the 550 individuals, skewed segregations, with a complete absence of one homozygous genotype, were only observed in the CH4 region and not in the CH1 region.

Segregations observed in UF 676 x ICS 95 involving the Amelonado CH1 allele: Each Trinitario parent of this progeny, hybrid between Amelonado and Criollo genotypes, is heterozygous for the mSI_103 (223) Amelonado allele and the mSI_103 (228) allele originated from Criollo. No skewed segregation was observed in this progeny.

Segregations observed in the Self-compatible F2 plants, homozygous for the CH4 loci (Sf/Sf) and heterozygous for the CH1 loci, after self pollinations: Two F2 plants, BR36 and BR59, homozygous Sf/Sf for the CH4 locus and heterozygous for the CH1 locus, were self-pollinated. It was observed in both cases, as in the F2 progeny, a total absence of the “a” genotype corresponding to plants homozygous for the Amelonado allele (mSI_103-223) for the CH1 locus. These results show clearly that even if the plants are SC due to the Sf/Sf/CH4 genotype, the CH1 locus is still functional, and a genotypic selection could still be observed.

Table 3: segregations observed at the level of CH1 and CH4 loci for other progenies.
**Origin of genotypic selection in self pollinated young fruits of BR59**

In order to check if the genotypic selection provided by the CH1 locus could result from gamete non fusion, the proportion of fertilized ovules were assessed in young fruits of BR59 seven days after self-pollination. If the skewed segregations observed in the CH1 region reflect gamete non fusion events, a proportion of 25% aborted ovules is expected. The proportion of aborted ovules was observed similar between self pollinated ovaries and ovaries cross-pollinated with a compatible pollen from a different origin (CCN51), taken as control (respectively 11 % and 9.2 %). This shows that in that case, the SI reaction did not result in gamete non fusions and consequently aborted ovules.

**Association studies for the fruit setting linked to self-incompatibility**

Marker/incompatibility trait association studies, conducted on the 570 individuals evaluated by 0/1 for SC/SI trait using 16480 SNP, revealed positive associations for 11 SNP markers, all located at the top of CH4 from the positions 189447 bp to 1328172 bp (Fig. 2). Only one marker, the 3673000|F|0-37:G>A located at position 189447 bp is included in the restricted region identified by fine mapping which contained 36 other SNP markers subjected to GWAS. This marker is located inside the Tc04_g000230 gene, an ortholog to the GEX1 gene of *Arabidopsis thaliana*. No association was detected in CH1. These observations revealed that only the CH4 locus is associated to fruit drop, contrary to the CH1 locus.

**Search for candidate genes potentially involved in the *T. cacao* self-incompatibility system in the CH1 and CH4 genome regions and analysis of their expression.**

Candidate genes potentially involved in cocoa self-incompatibility were first searched in the CH1 and CH4 regions without recombinant plants identified by fine mapping and the expression of twelve candidate genes was analysed during SC and SI reactions.

**CH1 Region**: In this region, only nine genes have been annotated in the whole genome sequence V1 of Criollo (Argout et al., 2011), from Tc01_g007220 to Tc01_g007300. Among them three genes are considered as possible candidate genes:
- **Tc01_g007220** is an ortholog of *Arabidopsis thaliana* BAM1 which encodes CLAVATA1-related Leucine rich repeat receptor like kinases having an important role in early anther development and development of both male and female gametophyte (Hord et al., 2006; DeYoung et al., 2006). Tc01_g007220 is differentially expressed between SC and SI reactions at later stages with a higher expression during SC reaction.
- **Tc01_g007270** is homologous to the COMPASS-like H3K4 histone methylase component WDR5a from *Arabidopsis thaliana* (At3g49660). It is a transducin WD-40 repeat-containing protein acting as a site of protein-protein interactions playing central roles in biological processes (Stirniman et al., 2010)
- **Tc01_g007290**, is a putative transmembrane transporter, part of the Major facilitator superfamily (MFS-1), and homologous to *A. thaliana* At5g65687 gene, a probable sphingolipid transporter spinster homolog 1. This gene is also predicted to interact with several genes involved in ubiquination, an important process during self-incompatibility reactions. Tc01_g007270 and Tc01_g007290 are differentially expressed between SC/SI reactions at early stages (2h to 8h) with a much higher expression observed for Tc01_g007290 during SI reactions.

![Fig 2: Genome-wide association analysis of self-incompatibility traits and localisation of the positive markers in the CH4 region. Only one marker is inside the genomic region defined by fine mapping and is located in a GEX1 orthologous gene.](image)
Immunolocalization experiments conducted with Tc01_g007270 and Tc01_g007290 showed that these genes are expressed in the cell layers of the style where are progressing the pollen tubes and in those surrounding the embryo sac (Fig. 3)

**Fig 3:** Immunolocalization of Tc01_g007270 and Tc01_g007290 proteins. Observations (confocal microscope) of sections of ovules, ovary and style of Scavina 6 accession 24H after pollinations by a compatible pollen (SC) and probed with antibodies of Tc01_g007270 and Tc01_g007290 proteins (green points).

**CH4 region:** The region of 257.270 Kb identified by fine mapping includes 30 genes annotated in the Criollo genome V1. Among them, nine genes were identified as candidates potentially involved in the SI reactions.

- **Tc04_g000160** is homologous to the Voltage-dependent L-type calcium channel and could have a role in Ca<sup>2+</sup> influx into stigma papilla cells mediating the SI signalling as was shown in the *Brassicaceae*

- **Tc04_g000170** is ortholog to a Kanadi transcription factor, potentially interacting with an Auxin factor response 3

- **Tc04_g000190, Tc04_g000230, Tc04_g000240** and **Tc04_g000260** are homologous to GEX1 (*Gamete Expressed Protein - At5g55490*) *A. thaliana* genes potentially involved in early embryo development (Alandete-Saez *et al.*, 2011). They have a dual function during male and female gametophyte development and early embryogenesis and are required for correct pollen maturation. Using STRING network view, it was shown that in *A. thaliana*, *GEX1* is interacting with the *HAPLESS2* gene involved in male fertility, essential for pollen tube guidance, successful gamete attachment and fertilization. Among them only Tc04_g000190 and Tc04_g000240 displayed a differential expression between SC/SI reactions compared to un-pollinated ovules.

- **Tc04_g000300** is ortholog to a ferredoxin thioredoxin reductase, but not differentially expressed between SC/SI reactions.

- **Tc04_g000320** is an ortholog of a Zinc finger AN1 domain-containing stress-associated protein 12 (PMZ) gene from *A. thaliana*. PMZ could interact with several genes associated to protein degradation through the ubiquination complex. Tc04_g000320 is significantly more expressed at the stage 12-24h during the SC reaction.

- **Tc04_g000330** is an ortholog of the Arm repeat-Containing protein ARC1 gene from *Brassica napa* (AGP76183.1). The ARC1 protein is involved in SI signalling in *Brassica* and targets proteins for degradation during SI response. Tc04_g000330 is differentially expressed, at the stage 12h-24h, where a higher expression is observed for SC reaction, which is not in agreement with what is expected during the SI reaction.

**Self-compatibility/incompatibility predictions**

**Allele specificity:** Specific Amelonado and Criollo alleles were identified among the CH4 markers, predicting SC trees homozygous for these alleles:
The 161 bp allele of mSI_460 is specific to Amelonado. Its low frequency in the Nanay and Nacional groups probably reflects introgression of Amelonado in some individuals of these 2 groups. The allele 217 of mSI_303 is specific to Criollo. These 2 alleles could be used to screen for SC plants, homozygous for these alleles.

**Genotype analysis:** Other *T. cacao* allele combinations than those identified in Amelonado and Criollo could result in SC plants. Dominance relationships exist between cacao S-alleles and the multi-allelic status of SSR markers allowed a better discrimination of S-haplotype interactions. When only one genotype was considered, 24 genotypes from 7 markers were significantly associated to cocoa self-incompatibility, by Fisher’s Exact test (q-value < 0.05). Among those genotypes, 16, when present, resulted in higher frequency of SC plants and 10 resulted in higher frequency of SI plants. When two genotypes were considered, 178 combinations of genotypes were significantly associated to SI in cacao, with 134 combinations resulting in more SC clones and 44 in more SI clones.

**Allele effect:** Among the 128 alleles tested, 43 were significantly associated with SI status by the Fisher’s Exact test.

Examples of predictions are reported in Table 4

### Table 4: examples of predictions. Probabilities of an individual to be self-incompatible (prob-SI) or self-compatible (Prob-SC) according to its genotype

<table>
<thead>
<tr>
<th>marker1</th>
<th>marker2</th>
<th>genotype</th>
<th>S_0</th>
<th>S_1</th>
<th>Prob-SI</th>
<th>Prob-SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSI_303</td>
<td>mSI_7</td>
<td>227/227</td>
<td>17</td>
<td>57</td>
<td>0.23</td>
<td>0.77</td>
</tr>
<tr>
<td>mSI_303</td>
<td>mSI_411</td>
<td>227/227</td>
<td>200/200</td>
<td>0</td>
<td>50</td>
<td>0.00</td>
</tr>
<tr>
<td>mSI_303</td>
<td>mSI_35</td>
<td>227/227</td>
<td>286/286</td>
<td>1</td>
<td>53</td>
<td>0.02</td>
</tr>
<tr>
<td>mSI_35</td>
<td>mSI_303</td>
<td>225/227</td>
<td>1</td>
<td>14</td>
<td>0.07</td>
<td>0.93</td>
</tr>
<tr>
<td>mSI_460</td>
<td>mSI_35</td>
<td>119/119</td>
<td>227/227</td>
<td>28</td>
<td>86</td>
<td>0.25</td>
</tr>
<tr>
<td>mSI_107</td>
<td>mSI_35</td>
<td>161/161</td>
<td>1</td>
<td>53</td>
<td>0.02</td>
<td>0.98</td>
</tr>
<tr>
<td>mSI_107</td>
<td>mSI_458</td>
<td>239/245</td>
<td>292/292</td>
<td>0</td>
<td>27</td>
<td>0.00</td>
</tr>
<tr>
<td>mSI_2</td>
<td>mSI_303</td>
<td>280/283</td>
<td>35</td>
<td>6</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td>mSI_303</td>
<td>mSI_303</td>
<td>231/233</td>
<td>23</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>mSI_458</td>
<td>mSI_460</td>
<td>231/231</td>
<td>24</td>
<td>0</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>mSI_460</td>
<td>mSI_460</td>
<td>306/306</td>
<td>26</td>
<td>2</td>
<td>0.93</td>
<td>0.07</td>
</tr>
<tr>
<td>mSI_7</td>
<td>mSI_7</td>
<td>155/161</td>
<td>22</td>
<td>1</td>
<td>0.96</td>
<td>0.04</td>
</tr>
<tr>
<td>mSI_7</td>
<td>mSI_7</td>
<td>145/146</td>
<td>24</td>
<td>2</td>
<td>0.92</td>
<td>0.08</td>
</tr>
<tr>
<td>mSI_7</td>
<td>mSI_7</td>
<td>194/197</td>
<td>22</td>
<td>1</td>
<td>0.96</td>
<td>0.04</td>
</tr>
</tbody>
</table>

### Discussion- conclusion

The *T. cacao* self–incompatibility system has already been described as a late incompatibility system having both gametophytic and sporophytic aspects and dominance relationships between alleles (Cope, 1958, 1962; Glendinning, 1960). Our results confirmed its gametophytic and sporophytic features, and led to the identification of two independent loci involved in the SI system through two different late acting mechanisms. The fine mapping and GWAS analyses lead to restricted chromosome regions where several candidate genes could be identified.

The CH1 locus, identified for the first time in this study, acts prior gamete fusion, but after pollen tube germination. Indeed, all ovules were fertilized which reflects a gametic selection that happened prior to the gamete fusion step. This selection could involve a lack of penetration or migration of the sperm nuclei in the embryo sac as was already observed in SCA 24 (closely related to SCA 6 used as pro-genitor in this study) (Ford and Wilkinson, 2012). Two main candidate genes were identified in the 70.3 kb CH1 region without recombinants plants, identified by fine mapping (*Tc01_g0007270* and *Tc01_g0007290*). Their proteins did not seem to be present in the pollen tubes themselves, observed at 8 or 24 h after pollination, but were apparent in the style cells where there were growing pollen tubes, and in several embryo sac cell layers.

The CH4 locus is involved in the SI reaction through a different mechanism, strongly associated with fruit drop. All positive associations identified by the GWAS were determined in the CH4 region where a gametic selection (probably linked to the observations on gamete non-fusions made by several authors: Cheesman 1927; Cope 1939b, 1940, 1958, 1962; Posnette, 1940; Knight and Rogers, 1955; Bouharmont, 1960; Ford and Wilkinson,
2012) was also observed in the F2 progeny under study, as in the self-progeny of IMC 60, another SI clone. In these two progenies, only the use of mentor pollen, which prevented the fruit from dropping, had allowed the seed development in ovules where gametic fusions were effective.

In the region of 257.270 kb, without recombination identified by fine mapping in the CH4, several candidate genes were identified, potentially involved mainly in ubiquination steps or in early embryo or gamete development and guidance, as the GEX1 (gamete expressed protein) orthologous genes. Their potential role in the T. cacao self-incompatibility system was enhanced by the results of GWAS for which the only association identified in the region without recombinant identified by fine mapping was located in a GEX1 orthologous gene. Recently, it has been observed that the non-fusion of gametes, as observed after pollination of T. cacao using incompatible pollen, could result from two different and independent mechanisms: an incomplete migration of the sperm nucleus, which fails to reach the female nuclei, or a successful sperm nucleus migration and reaching of the female nuclei, but followed by a non-fusion of gametes (Ford and Wilkinson, 2012). The role of GEX1 genes, interacting with the HAPLESS2 gene that is known to be essential for pollen tube guidance, successful gamete attachment and fertilization (von Besser et al., 2006; Mori et al., 2006), could be determinant in this late manifestation of SI in T. cacao. However, Tc04_g000160, Tc04_g000320, and Tc04_g000330, orthologs to genes involved in SI signalling in Brassica and protein degradation during the SI response, cannot be ruled out and could also participate in the pathway leading to self-incompatibility.

The practical output from this study is the identification of markers that could help breeders to select for SC plants in cocoa populations at an early stage in the breeding process. Indeed, the identification of highly polymorphic SSR markers in the CH4 region gives the potential to have a strong linkage disequilibrium between SSR alleles and incompatibility alleles, leading to a good ability of SSR to predict SI or SC genotypes. Specific Amelonado and Criollo alleles, identified in this study, will be particularly efficient in predicting SC varieties.

**Acknowledgements**

This project was supported by Agropolis Fondation under the reference ID 1403-046 through the “Investissements d’avenir” programme (Labex Agro: ANR-10-LABX-0001-01). We also thank Mondelēz International for their financial contribution to this project. We thank Marc Lartaud and Geneviève Conéjéro for their helpful advice in immunolocalization experiments and imaging. We thank Neba Ngwa AKONGNWI and Semi Melliti for their help in managing field pollinations in Cameroun and greenhouse in Montpellier.

All details about markers, primers and gene expression values are reported in *Journal of Experimental Botany*, Volume 68, Issue 17, 13 October 2017, Pages 4775-4790, [https://doi.org/10.1093/jxb/erx293](https://doi.org/10.1093/jxb/erx293)
Hord CLH, Chen C, DeYoung BJ, Clark SE, Ma H. 2006. The BAM1/BAM2 receptor-like kinases are important regulators of Arabidopsis early anther development. Plant Cell 18,1667-1680.
Posnette AF. 1940. Self-incompatibility in cocoa (Theobroma spp.). Tropical Agriculture, Trinidad and


