Genomics-enabled analysis of the emergent disease cotton bacterial blight

Anne Z. Phillips¹,², Jeffrey C. Berry¹✉, Mark C. Wilson¹✉, Anupama Vijayaraghavan¹, Jillian Burke¹,
J. Imani Bunn¹, Tom W. Allen³, Terry Wheeler⁴, Rebecca Bart¹✉

¹Donald Danforth Plant Science Center, St. Louis MO, United States of America
²Department of Biology, Washington University in Saint Louis, St. Louis, MO, United States of America
³Delta Research and Extension Center, Mississippi State University, Stoneville, MS, United States of America
⁴Texas AgriLife Research, Texas AgriLife Extension Service, Lubbock, TX, United States of America

*Corresponding Author
Email: rbart@danforthcenter.org

*these authors contributed equally to this work.
Abstract

Cotton bacterial blight (CBB), an important disease of *Gossypium hirsutum* in the early 20\textsuperscript{th} century, had been controlled by resistance genes for over half a century. Recently, CBB re-emerged as an agronomic problem in the United States. Here, a comparative genomics analysis of host and pathogen was conducted. Phylogenetic analysis revealed that strains from the current outbreak cluster with race 18 *Xanthomonas citri* pv. *malvacearum* (*Xcm*) strains. Type three effector repertoires of 16 *Xcm* isolates reveal 24 conserved effectors as well as nine variable effectors. In addition, virulent race 18 strains contain 3 to 5 more effectors than non-race 18 strains. Genome assemblies for two geographically and temporally divergent strains of *Xcm*, yielded circular chromosomes and accompanying plasmids. These genomes encode eight and thirteen distinct transcription activator-like effector genes. RNA-sequencing revealed that both strains induced 52 conserved gene targets in diverse cotton cultivars, including a homeologous pair of genes, with homology to the known susceptibility gene, MLO. In contrast, the two strains of *Xcm* induced different SWEET sugar transporters and in one case, only one homeolog was significantly induced. Subsequent genome wide analysis revealed the overall expression patterns of the homeologous gene pairs in cotton after inoculation by *Xcm*. These data reveal host-pathogen specificity in the *Xcm-G. hirsutum* pathosystem, give explanations for the CBB reemergence, and strategies for future development of resistant cultivars.

Author Summary
Cotton bacterial blight (CBB), caused by *Xanthomonas citri* pv. *malvacearum* (*Xcm*), significantly limited cotton yields in the early 20th century but has been controlled by classical resistance genes for more than 50 years. In 2011, the pathogen re-emerged with vengeance. In this study, we compare diverse pathogen isolates and cotton varieties to further understand the virulence mechanisms employed by *Xcm* and to identify promising resistance strategies. We generate fully contiguous genome assemblies for two diverse *Xcm* strains and identify pathogen proteins used to modulate host transcription and promote susceptibility. RNA-sequencing of infected cotton reveals novel putative gene targets for the development of durable *Xcm* resistance. Together, the data presented reveal the underlying cause of CBB re-emergence in the U.S. and highlight several promising routes towards the development of durable resistance including classical resistance genes and potential manipulation of susceptibility targets.

**Introduction**

Upland cotton (*Gossypium hirsutum* L.) is the world’s leading natural fiber crop. Cotton is commercially grown in over 84 countries and in the United States is responsible for $74 billion annually [1, 2]. Numerous foliar diseases affect cotton throughout the world’s cotton growing regions. Historically, one of the most significant foliar diseases has been bacterial blight, caused by *Xanthomonas citri* pv. *malvacearum*. Cotton bacterial blight significantly limited cotton yield in the late 20th century. In the 1940’s and 1950’s, breeders identified and introgressed multiple resistance loci into elite germplasm [3-5]. This strategy proved durable for over half a century. In 2011, cotton bacterial blight (CBB) returned and caused significant losses to farmers in the southern United States, more specifically in Arkansas and Mississippi.
Nonetheless, CBB has received little research focus during the last several decades because this disease had been considered “tamed”. Modern molecular and genomic technologies can now be employed expeditiously to deduce the underlying cause of the disease re-emergence and pinpoint optimized routes towards the development of durable resistance.

CBB is caused by *X. citri pv. malvacearum (Xcm)*; however, the pathogen has previously been placed within other species groupings [6-9]. The *Xcm* pathovar can be further divided into at least 19 races according to virulence phenotypes on a panel of historical cotton cultivars: Acala-44, Stoneville 2B-S9, Stoneville 20, Mebane B-1, 1-10B, 20-3, and 101-102.B [10, 11].

Historically, the most common race observed in the U.S. has been race 18, which was first isolated in 1973 [12]. This race is highly virulent, causing disease on all cultivars in the panel except for 101-102.B. CBB can occur at any stage in the plant’s life cycle and on any aerial organ. Typical symptoms include seedling blight as either pre or post-emergent damping-off, black arm on petioles and stems, water-soaked spots on leaves and bracts, and most importantly boll rot. The most commonly observed symptoms are the angular-shaped lesions on leaves that, in some cases, can coalesce and result in a systemic vein infection where leaf lesions coalesce on major leaf veins. Disease at each of these stages can cause yield losses either by injury to the plant or direct damage to the boll. No effective chemical treatments for the disease have been released to date. Therefore, the most important methods to reduce loss as a result of CBB include field methods that rely on cultivation to reduce potential sources of overwintering inoculum and planting cultivars with known sources of resistance.

Most pathogenic bacteria assemble the type three secretion system (T3SS), a needle-like structure, to inject diverse type three effectors (T3Es) into the plant cell to suppress
immunity and promote disease [13-17]. For example, transcription activator-like (TAL) effectors influence the expression levels of host genes by binding directly to host gene promoters in a sequence-specific way [18]. Up-regulated host genes that contribute to pathogen virulence are termed susceptibility genes and may be modified through genome editing for the development of resistant crop varieties [19].

Plants have specialized immune receptors, collectively known as nucleotide-binding leucine rich repeat receptors (NLRs), that recognize, either directly or indirectly, the pathogen effector molecules [20, 21]. Historically, this host-pathogen interaction has been termed the ‘gene-for-gene’ model of immunity, wherein a single gene from the host and a single gene from the pathogen are responsible for recognition [22]. Recognition triggers a strong immune response that often includes a localized hypersensitive response (HR) in which programmed cell death occurs around the infection site [23]. Nineteen CBB resistance loci have been reported in Gossypium hirsutum breeding programs; however, none have been molecularly identified [8, 24].

Here we combine comparative genomics of the pathogen Xcm with transcriptomics of the host to identify the molecular interactions underlying this re-emergent disease. This will inform the development of durable resistance strategies.

Results

CBB Reemergence in the US

In 2011, farmers, Extension specialists, and Certified Crop Advisers in Missouri, Mississippi, and Arkansas observed cotton plants exhibiting symptoms of CBB. While a major limiting factor for cotton production through the 1950s, this disease had been controlled by
agricultural practices such as acid-delinting seed as well as planting resistant cultivars. Prior to the widespread observation of CBB in the mid-southern U.S., isolated, sporadic instances of the disease were generally detected on an annual basis. Reemergence of the disease occurred rapidly during 2011. Widespread infected plant material was observed throughout much of the production area, but appeared to be centered around Clarksdale, Mississippi. Much of the infestation in the Arkansas production system was reported to have originated from several infested seed lots [25]. The disease has since spread through much of the cotton belt in the southern U.S. (Figs 1 and S1).

In 2014, diseased cotton leaves were collected from three sites across Mississippi and Koch’s postulates were conducted to prove causality [26]. PCR amplification of the 16S rRNA gene confirmed that the causal agent was a member of the Xanthomonas genus. Multi locus sequence type (MLST) analysis and maximum-likelihood analysis were performed using concatenated sections of the gltA, lepA, lacF, gyrB, fusA and gap-1 loci (Fig 2a) for increased phylogenetic resolution. The newly sequenced strains were named MS14001, MS14002 and MS14003 and were compared to four previously published Xcm genomes and thirty-six additional Xanthomonas genomes representing thirteen species (Tables S1, S2). MS14001, MS14002 and MS14003 grouped with the previously published Xcm strains as a single polytomy, further confirming that the current disease outbreak is CBB and is caused by Xcm. The species designation reported here is consistent with previous reports [6, 7]. To date, CBB has been reported from at least eight out of the sixteen states that grow cotton (Fig 1).

Contemporary U.S. Xcm strains cluster phylogenetically with historical race 18 strains.
Race groups have been described for *Xcm* strains by analyzing compatible (susceptible) and incompatible (resistant) interactions on a panel of seven cotton cultivars. In general, race groups tend to be geographically distinct. For example, as mentioned previously, race 18 is prevalent in the U.S. while race 20 is a highly virulent strain reported from several African countries [7]. Consequently, one possible explanation for the recent outbreak of CBB would be the introduction of a new race of *Xcm* capable of overcoming existing genetic resistance. Unfortunately, only 2 cultivars of the original cotton panel plus three related cultivars, were available and these cultivars were not sufficient to determine whether a new race had established within the U.S. Consequently, twelve *Xcm* strains were sequenced using Illumina technology to determine the phylogenetic relationship between recent isolates of *Xcm* and historical isolates. Isolates designated as race 1, race 2, race 3, race 12 and race 18 have been maintained at Mississippi State University. Additional isolates were obtained from the Collection Française de Bactéries associées aux Plantes (CFBP) culture collection. Together, these isolates include eight strains from the US, three from Africa, and one from South America and span collection dates ranging from 1958 through 2014 (Fig 1). Illumina reads were mapped to the *Xanthomonas citri* subsp. *citri* strain Aw12879 (565918 [RefSeq]) using Bowtie and single nucleotide polymorphisms (SNPs) were identified using Samtools [27, 28]. Only regions of the genome with at least 10x coverage for all genomes were considered. This approach identified 17,853 sites that were polymorphic in at least one genome. Nucleotides were concatenated and used to build a neighbor-joining tree (Fig 2b). This analysis revealed that recent *Xcm* isolates grouped with the race 18 clade. Notably, the race 18 clade is phylogenetically distant from the other *Xcm* isolates.
Contemporary US Xcm strains have conserved virulence protein arsenals and disease phenotypes with historical race 18 strains.

Type three effector (T3E) profiles from sixteen Xcm isolates were compared to determine whether a change in the virulence protein arsenal of the newly isolated strains could explain the re-emergence of CBB. Genomes from 12 Xcm isolates were de novo assembled with SPAdes and annotated with Prokka based on annotations from the X. euvesicatoria (aka. X. campestris pv. vesicatoria) 85-10 genome (NCBI accession: NC_007508.1). T3Es pose a particular challenge for reference based annotation as no bacterial genome contains all effectors. Consequently, an additional protein file containing known T3Es from our previous work was included within the Prokka annotation pipeline [13, 29]. This analysis revealed 24 conserved and 9 variable Xcm T3Es (Fig 3a). Most race 18 isolates contain more effectors than other isolates that were sequenced. The recent Xcm isolates (MS14002 and MS14003) were not distinguishable from historical race 18 isolates, with the exception of XcmNI86 isolated from Nigeria in 1986, which contains mutations in XopE2 and XopP.

Analysis of the genomic sequence of T3E revealed presence/absence differences, frameshifts and premature stop codons. However, this analysis does not preclude potential allelic or expression differences among the virulence proteins that could be contributing factors to the re-emergence of CBB. Therefore, newly isolated strains may harbor subtle genomic changes that have allowed them to overcome existing resistance phenotypes. Many commercial cultivars of cotton are reported to be resistant to CBB [30-32]. Based on these previous reports, we selected commercial cultivars resistant and susceptible (6 of each) to CBB. In addition, we included 5 available varieties that are related to the historical panel as well as 2
parents from a nested association mapping (NAM) population currently under development [33]. All varieties inoculated with the newly isolated Xcm strains exhibited inoculation phenotypes consistent with previous reports for these varieties (Figs 3b,c). In these assays, brightfield and near infrared (NIR) imaging were used to distinguish water-soaked disease symptoms from rapid cell death (hypersensitive response) that is indicative of an immune response. These data confirm that existing resistance genes present within cotton germplasm are able to recognize the newly isolated Xcm strains and trigger a hypersensitive response.

Together, the phylogenetic analysis, effector profile conservation and cotton inoculation phenotypes, confirm that the recent outbreak of Xcm in the US represents a re-emergence of race 18 Xcm and is not the result of a dramatic shift in the pathogen.

The USDA Agricultural Marketing Service (AMS) publishes the percentage of upland cotton cultivars planted in the U.S. each year (www.ams.usda.gov/mnreports/cnnavar.pdf). In 2016, only 25% of the total cotton acreage was planted with resistant cultivars (Fig 3d), based on previously published CBB phenotypes for these cultivars. This is part of a larger downward trend in which the acreage of resistant cultivars has fallen each year since at least 2009.

**Comparative genome analysis for two Xcm strains**

Differences in virulence were observed among Xcm strains at the molecular and phenotypic level. In order to gain insight into these differences, we selected two strains from our collection that differed in T3E content, virulence level, geography of origin and isolation date. AR81009 was isolated in Argentina in 1981 and is one of the most virulent strains investigated in this study; MS14003 was isolated in Mississippi in 2014 and causes comparatively slower and diminished leaf symptoms. However, both strains are able to multiply
and cause disease on susceptible varieties of cotton (S2 Fig). Full genome sequences were
generated with Single Molecule Real-Time (SMRT) sequencing. Genomes were assembled using
the PacBio Falcon assembler which yielded circular 5Mb genomes and associated plasmids.
Genic synteny between the two strains was observed with the exception of two 1.05 Mb
inversions (Fig 4). Regions of high and low GC content, indicative of horizontal gene transfer,
were identified in both genomes. In particular, a 120kb insertion with low GC content was
observed in AR81009. This region contains one T3E as well as two annotated type four
secretion system related genes, two conjugal transfer proteins, and two multi drug resistant
genres. MS14003 contained three plasmids of the sizes 52.4, 47.4, and 15.3kb while AR81009
contained two plasmids of the sizes 92.6 and 22.9kb. Analysis of homologous regions among
the plasmids was performed using progressiveMauve [34]. This identified four homologous
regions greater than 1kb that were shared among multiple plasmids (Fig 4).

The AR81009 genome encodes twelve TAL effectors that range in size from twelve to
twenty three repeat lengths, six of which reside on plasmids. The MS14003 genome encodes
eight TAL effectors that range in size from fourteen to twenty eight repeat lengths, seven of
which reside on plasmids (Fig 5a). Three incomplete TAL effectors were also identified within
these genomes. A 1-repeat gene with reduced 5’ and 3’ regions was identified in both strains
directly upstream of a complete TAL effector. In addition, a large 4kb TAL effector was
identified in AR81009 with a 1.5 kb insertion and 10 complete repeat sequences. The tool
AnnoTALE was used to annotate and group TAL effectors based on the identities of the repeat
variable direseidues (RVDs) in each gene [35]. Little homology was identified among TAL
effectors within and between strains; only two TAL effectors were determined to be within the
same TAL class between strains (TAL19b of AR81009 and TAL19 of MS14003) and two within strain MS14003 (TAL14b and TAL16). Both strains express TAL effector proteins as demonstrated through western blot analysis using a TAL effector specific antibody (Fig 5b).

However, the complexity of TAL effector repertoires within these strains prevented complete resolution of each individual TAL effector.

Transcriptome changes induced by Xcm in G. hirsutum.

An RNA-sequencing experiment was designed to determine whether AR81009 and MS14003 incite different host responses during infection (Fig 6a). Isolates were inoculated into the phylogenetically diverse G. hirsutum cultivars Acala Maxxa and DES 56 (Fig 6b) [33]. Infected and mock-treated tissue were collected at 24 and 48 hours post inoculation. First, we considered global transcriptome patterns of gene expression. Fifty-two genes were determined to be induced in all Xcm-G. hirsutum interactions at 48 hours (S3 Table). Of note among this list of genes is a homeologous pair of genes with homology to the known susceptibility target, MLO [36-39]. Gene induction by a single strain was also observed; AR81009 and MS14003 uniquely induced 127 and 16, G. hirsutum genes, respectively (Fig 6c). The increased number of genes induced by AR81009 correlates with the observed severe leaf symptoms caused by this strain. In contrast, the average magnitude of gene induction between the two strains was not significantly different (S3 Fig). Both Xcm strains caused more genes to be differentially expressed in DES 56 than in Acala Maxxa. Among the 52 genes significantly induced by both strains, sixteen conserved targets are homeologous pairs, whereas seventeen and fifteen genes are encoded by the A and D sub-genomes, respectively (Tables 1 and S3). It has been previously reported that homeologous genes encoded on the G. hirsutum A and D sub-genomes are
differentially regulated during abiotic stress [40]. A set of approximately 10,000 homeologous gene pairs were selected and differential gene expression was assessed (Fig 7). For each pairwise comparison of Xcm strain and G. hirsutum cultivar, a similar number of genes were differentially expressed in each of the A and D subgenomes. However, some homeologous pairs were up or down regulated differentially in response to disease, indicating a level of sub-genome specific responses to disease. For example, SWEET sugar transporter gene Gh_D12G1898 in the D genome is induced over four fold during infection with Xcm strain AR81009, but the homeolog Gh_A12G1747 in the A genome is not. This relationship is further explored below.

**Table 1:** Eight homeologous pairs of *Gossypium hirsutum* genes are upregulated in both Acala Maxxa and DES 56 varieties 48 hours post inoculation with *Xanthomonas citri pv. malvacearum* strains MS14003 and AR81009.

<table>
<thead>
<tr>
<th>A Genome</th>
<th>D Genome</th>
<th>Gene Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gh_A02G0615</td>
<td>Gh_D02G0670</td>
<td>Seven transmembrane MLO family protein</td>
</tr>
<tr>
<td>Gh_A03G0560</td>
<td>Gh_D03G0971</td>
<td>Pectate lyase family protein</td>
</tr>
<tr>
<td>Gh_A05G2012</td>
<td>Gh_D05G2256</td>
<td>Protein of unknown function DUF688</td>
</tr>
<tr>
<td>Gh_A06G0439</td>
<td>Gh_D06G0479</td>
<td>basic chitinase</td>
</tr>
<tr>
<td>Gh_A07G1129</td>
<td>Gh_D07G1229</td>
<td>Protein of unknown function (DUF1278)</td>
</tr>
<tr>
<td>Gh_A10G0257</td>
<td>Gh_D10G0257</td>
<td>Protein E6</td>
</tr>
<tr>
<td>Gh_A10G1075</td>
<td>Gh_D10G1437</td>
<td>Pectin lyase-like superfamily protein</td>
</tr>
<tr>
<td>Gh_A13G1467</td>
<td>Gh_D13G1816</td>
<td>pathogenesis-related 4</td>
</tr>
</tbody>
</table>

**Different strains of Xcm target distinct SWEET transporters in G. hirsutum.**

SWEET sugar transporter genes are commonly targeted and upregulated by TAL effectors in *Xanthomonas* plant interactions [19, 41-43]. Surprisingly, no SWEET genes were detected in the above list of conserved targets. However, of the 54 SWEET sugar transporter genes encoded by the *G. hirsutum* genome, three were upregulated greater than 4 fold in response to inoculation by one of the two Xcm strains (Fig 8). Potential TAL effector binding
sites were identified using the program TALEnt [44]. MS14003 significantly induces the homeologs Gh_A04G0861 and Gh_D04G1360 and contains three TAL effectors predicted to bind within the 300bp promoter sequences of at least one of these genes (Fig 8a). In contrast, AR81009 significantly induces Gh_D12G1898 but not its homeolog Gh_A12G1747 (Fig 8b). TAL14a, TAL14c, and TAL16b from AR81009 are all predicted to bind to the Gh_D12G1898 promoter however the latter two are also predicted to bind to the homeolog Gh_A12G1747. We note that while Gh_A12G1747 did not pass the four fold cut off for gene induction, this gene is slightly induced in DES 56 compared to mock inoculation.

Discussion

Cotton Bacterial Blight was considered controlled in the U.S. until an outbreak was observed during the 2011 growing season in Missouri, Mississippi and Arkansas [45]. Until 2011, seed sterilization, breeding for resistant varieties, and farming techniques such as crop rotation and sterilizing equipment prevented the disease from becoming an economic concern [46]. The number of counties reporting incidence of CBB has increased from 17 counties in 2011 to 77 counties in 2015 [47-49]. This paper investigates the root of the re-emergence and identifies several routes towards control of the disease.

When the disease was first recognized as re-emerging, several possible explanations were proposed including: (1) A highly virulent race of the pathogen that had been introduced to the U.S.; (2) Historical strains of Xcm that had evolved to overcome existing resistance; and (3) Environmental conditions over the last several years that had been particularly conducive to the disease. Here, we present evidence that the re-emergence of CBB is not due to a large genetic change or race shift in the pathogen as has been previously suggested. Rather, the re-
The emergence of the disease is likely due to large areas of susceptible cultivars being planted. The presented data do not rule out potential environmental conditions that may also have contributed to the re-emergence. In this context, environmental conditions includes disease conducive temperature and humidity as well as potentially contaminated seed or other agronomic practices that may have perpetuated spread of the disease outbreaks. Importantly, the presented data confirm that the presence of resistance loci could be deployed to prevent further spread of this disease. However, since many of the most popular farmer preferred varieties lack these resistance traits, additional breeding or biotechnology strategies will be needed to maximize utility. Notably, the current Xcm isolates characterized in this study all originate from Mississippi cotton fields in 2014. During the 2015 and 2016 growing seasons, resistant cotton cultivars were observed in Texas with symptoms indicative of bacterial infection yet distinct from CBB. Additional work is underway to identify and characterize the causal agent(s) of these disease symptoms.

While resistant cotton cultivars were identified for all strains in this study, variability in symptom severity was observed for different strains when inoculated into susceptible cultivars. Two strains in particular, MS14003 and AR81009, have different effector profiles as well as different disease phenotypes. Comparative genomic analysis of the two pathogens revealed many differences that may contribute to the relative disease severity phenotypes. Similarly, transcriptomic analysis of two cultivars of G. hirsutum inoculated with these strains confirm that the genomic differences between the two strains result in a divergence in their molecular targets in the host.
Over the past decade, susceptibility genes have become targets for developing disease tolerant plants [50, 51]. These genes are typically highly induced during infection [52]. Therefore, RNA-Seq of infected plants has become a preferred way to identify candidate susceptibility genes. Once identified, genome editing can be used to block induction of these genes [53]. We report a homeologous pair of genes that are homologs of the MLO gene as targeted by both Xcm strains in both cotton cultivars. This conservation makes it an excellent candidate for future biotechnology efforts. Because the potential importance of this gene in cotton biology is unknown, the effect of disrupting this gene in cotton physiology must first be explored. However, knock-outs of MLO genes in other systems has led to durable resistance against powdery mildew but also oomycetes and bacteria such as Xanthomonas [36, 39]. The dual purpose of host susceptibility genes has been observed previously. For example, the rice Xa13 (aka. Os8N3 and OsSWEET11) gene is required for pollen development but also targeted by a rice pathogen during infection [54]. Xa13 is a member of the SWEET sugar transporters implicated in many pathosystems. In this case, the induction of Xa13 for pathogen susceptibility is mediated by a TAL effector. Of the 54 SWEET genes in the G. hirsutum genome, three are significantly upregulated during Xcm infection. In contrast to MLO, no single SWEET gene was induced by both pathogen strains in both hosts. Analysis of SWEET gene expression after inoculation revealed a context for polyploidy in the G. hirsutum-Xcm pathosystem. We observed a difference in induction between the Gh_A12G1747 and Gh_D12G1898 SWEET genes. Future research may investigate the diploid ancestors of tetraploid cotton to further explore the evolution of host and pathogen in the context of ploidy events [55].
Multiple putative TAL effector binding sites were identified within each up-regulated SWEET promoter. These observations suggest that TAL14b, TAL28a and TAL28b from MS14003 may work independently or in concert to induce the homeologs Gh_A04G0861 and Gh_D04G1360. Further, TAL14a from AR81009 is likely responsible for the upregulation of Gh_D12G1898. Whether additional TAL effectors are involved in these responses is not clear. It is possible that not all the TAL effectors are expressed. Similarly, genome organization in the host, such as histone modifications or other epigenetic regulation may be affecting these interactions. Future research will investigate these mechanisms further. However, these experiments will be difficult as most Xcm strains are not amenable to conjugation nor electroporation.

Collectively, the data presented here suggest that the wide-spread planting of CBB-susceptible cultivars has contributed to the re-emergence of CBB in the southern U.S. It is possible that a reservoir of race 18 Xcm was maintained in cotton fields below the level of detection due to resistant cultivars planted in the 1990s and early 2000s. Alternatively, the pathogen may have persisted on an alternate host or was brought in on contaminated seed as has previously been suggested [9, 10]. Regardless of the cause of the re-emergence, the genomic comparisons among pathogen races and host cultivars has identified several possible routes towards resistance. These include the use of existing effective resistance loci as well as the potential disruption of the induction of susceptibility genes through genome editing. The latter is an attractive strategy in part because of recent progress in genome editing [56, 57]. In summary, within a relatively short time frame, through the deployment of modern molecular and genomic techniques, we were able to identify the cause of the re-emergence of cotton
bacterial blight and generate data that can now be rapidly translated to effective disease control strategies.

**Methods**

*Xcm* strain isolation and manipulation:

New strains were isolated from infected cotton leaves by grinding tissue in 10mM MgCl₂ and culturing bacteria on NYGA media. The most abundant colony type was selected, single colony purified and then 16S sequencing was used to confirm the bacterial genus as previously described [58]. In addition, single colony purified strains were re-inoculated into cotton leaves and the appearance of water soaked symptoms indicative of CBB infection was confirmed. Both newly isolated strains as well as strains received from collaborators were used to generate a rifampicin resistance version of each strain. Wildtype strains were grown on NYGA, then transferred to NYGA containing 100µg/ml rifampicin. After approximately 4-5 days, single colonies emerged. These were single colony purified and stored at -80°C. The rifampicin resistant version of each *Xcm* strain was used in all subsequent experiments reported in this manuscript unless otherwise noted.

Plant inoculations

*Xcm* strains were grown on NYGA plates containing 100µg/ml rifampicin at 30°C for two days before inoculations were performed. Disease assays were conducted in a growth chamber set at 30°C and 80% humidity. Inoculations were conducted by infiltrating a fully expanded leaf with a bacterial solution in 10mM MgCl₂ (OD₆₀₀ specified within each assay).

Cotton Cultivar Statistics
Area of cotton planted per county in the United States in 2015 was obtained from the USDA National Agricultural Statistics Service:

www.nass.usda.gov/Statistics_by_Subject/result.php?7061F36A-A4C6-3C65-BD7F-129B702CFBA2&sector=CROPS&group=FIELD%20CROPS&comm=COTTONUSDA. Estimated percentage of upland cotton planted for each variety was obtained from the Agricultural Marketing Service (AMS): www.ams.usda.gov/mnreports/canvar.pdf. CBB disease phenotyping data from 2009-2016 was determined via cotyledon scratch assays and/or field trials sprayed with virulent Xcm isolates and has previously been described [59-61].

**Bacterial Sequencing and Phylogenetics**

Illumina based genomic datasets were generated as previously described [29]. Paired-end Illumina reads were trimmed using Trimmomatic v0.32 (ILLUMINACLIP:TruSeq-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) [62]. Genome assemblies were generated using the SPAdes *de novo* genome assembler [63]. Strain information is reported in Supplemental Table 1. Similar to our previously published methods [29], the program Prokka was used in conjunction with a T3E database to identify type three effector repertoires for each of the 12 Xcm isolates as well as four Xcm genomes previously deposited on NCBI (S2Table) [64].

Multi-locus sequence analysis was conducted by concatenating sequences of the gltA, lepA, lacF, gyrB, fusA and gap-1 loci obtained from the Plant-Associated Microbes Database (PAMDB) for each strain as previously described [65]. A maximum-likelihood tree using these concatenated sequences was generated using CLC Genomics 7.5.

**Variant Based Phylogeny**
A variant based dendrogram was created by comparing 12 Illumina sequenced Xcm genomes to the complete Xanthomonas citri subsp. citri strain Aw12879 reference genome (565918 [RefSeq]) on NCBI. Read pairs were aligned to the reference genome using Bowtie2 v2.2.9 with default alignment parameters [27]. From these alignments, single nucleotide polymorphisms (SNPs) were identified using samtools mpileup v1.3 and the bcftools call v1.3.1 multi-allelic caller [28]. Using Python v2.7, the output from samtools mpileup was used to identify loci in the X. citri subsp. citri reference genome with a minimum coverage of 10 reads in each Xcm genome used Python version 2.7 available at http://www.python.org. Vcftools v0.1.14 and bedtools v2.25.0 were used in combination to remove sites marked as indel, low quality, or heterozygous in any of the genomes [66, 67]. Remaining loci were concatenated to create a FASTA alignment of confident loci. Reference loci were used where SNP's were not detected in a genome. The resulting FASTA alignment contained 17853 loci per strain. This alignment was loaded into the online Simple Phylogeny Tool from the ClustalW2 package to create a neighbor joining tree of the assessed strains [68, 69]. Trees were visualized using FigTree v1.4.2.

**Genome Assembly**

Single Molecule, Real Time (SMRT) sequencing of Xcm strains MS14003 and AR81009 was obtained from DNA prepped using a standard CTAB DNA preparation. Blue Pippin size selection and library preparation was done at the University of Delaware Sequencing Facility. The genomes were assembled using FALCON-Integrate (https://github.com/PacificBiosciences/FALCON-integrate/commit/cd9e93) [70]. The following parameters were used: Assembly parameters for MS14003: length_cutoff = 7000;
length_cutoff_pr = 7000; pa_HPCdaligner_option = -v -dal8 -t16 -e.70 -l2000 -s240 -M10;

ovlp_HPCdaligner_option = -v -dal8 -t32 -h60 -e.96 -l2000 -s240 -M10; falcon_sense_option = --

output_multi --min_idt 0.70 --min_cov 5 --local_match_count_threshold 2 --max_n_read 300 --
n_core 6; overlap_filtering_setting = --max_diff 80 --max_cov 160 --min_cov 5 --bestn 10;

Assembly parameters for AR8109: length_cutoff = 8000; length_cutoff_pr = 8000;

pa_HPCdaligner_option = -v -dal8 -t16 -e.72 -l2000 -s240 -M10; ovlp_HPCdaligner_option = -v -
dal8 -t32 -h60 -e.96 -l2000 -s240 -M10; falcon_sense_option = --output_multi --min_idt 0.72 --

min_cov 4 --local_match_count_threshold 2 --max_n_read 320 --n_core 6;

overlap_filtering_setting = --max_diff 90 --max_cov 300 --min_cov 10 --bestn 10. Assemblies

were polished using iterations of pbalign and quiver, which can be found at

https://github.com/PacificBiosciences/pbalign/commit/cda7abb and

https://github.com/PacificBiosciences/GenomicConsensus/commit/43775fa. Two iterations

were run for Xcm strain MS14003 and 3 iterations for AR81009. Chromosomes were then

reoriented to the DnaA gene and plasmids were reoriented to ParA. The assemblies were

checked for overlap using BLAST, and trimmed to circularize the sequences [71]. TAL effectors

were annotated and grouped by RVD sequences using AnnoTALE [35]. Homologous regions

among plasmids that are greater than 1 kb were determined using progressiveMauve [34].

Genomic comparisons between the MS14003 and AR81009 chromosomes were visualized using

Circos [72]. Single-copy genes on each of the chromosomes were identified and joined using

their annotated id’s. Lines connecting the two chromosomes represent these common genes

and their respective positions in each genome. A sliding window of 1KB was used to determine

the average GC content. Methylation was determined using the Base Modification and Motif
Western Blot Analysis

Western Blot analysis of Transcription Activator-Like (TAL) effectors was performed using a TAL specific antibody [43]. Briefly, bacteria were suspended in 5.4 pH minimal media for 4.5 hours to induce effector production and secretion. Pellet was then suspended in laemmli buffer at 95 degrees Celsius for three minutes to lyse the cells. Freshly boiled samples were then loaded onto a 4-6% gradient gel and run for several hours to ensure sufficient separation of the different sized TAL effectors. Polyclonal rbTal10 antibody was used to visualize all TALs.

Gene Expression Analysis

Susceptible cotton were inoculated with Xcm using a needleless syringe at an OD_{600} of 0.5. Infected and mock-treated tissue were collected and flash frozen at 24 and 48 hours post inoculation. RNA was extracted using the Sigma tRNA kit. RNA-sequencing libraries were generated as previously described [73].

Raw reads were trimmed using Trimmmomatic [62]. The Tuxedo Suite was used for mapping reads to the TM-1 NBI Gossypium hirsutum genome [74], assembling transcripts, and quantifying differential expression [27, 75].

Homeologous pairs were identified based on syntenic regions with MCScan [76]. A syntenic region is defined as a region with a minimum of five genes with an average intergenic distance of 2 and within extended distance of 40. All other values are set to the default.

Bioinformatic prediction of TAL effector binding sites on the G. hirsutum promoterome was performed using the TAL Effector-Nucleotide Targeter (TALEnt) [44]. In short, the regions of
the genome that were within 300 basepairs of annotated genes were queried with the RVD’s of MS14003 and AR81009 using a cutoff score of 4. Promiscuously binding TALs 16 from MS14003 and 16a from AR81009 were removed from analysis.

**Fig Legends**

**Fig 1: Cotton Bacterial Blight (CBB) symptoms and reemergence across the southern United States.** (Left) Typical CBB symptoms present in cotton fields near Lubbock, TX during the 2015 growing season include angular leaf spots, boll rot, and black arm rot. Acres of cotton planted per county in the United States in 2015 (blue) and counties with confirmed CBB in 2015 (red outline). Statistics on cotton planted in the U.S. were acquired from the USDA. CBB was reported by Extension agents, Extension specialists, and Certified Crop Advisers in their respective states, and compiled by Tom Allen.

**Fig 2: Phylogenetic analysis of Xcm isolates and 13 species of Xanthomonas** A) MLST (Multi Locus Sequence Typing) analysis of 12 Illumina sequenced Xcm isolates (this paper) and 40 other Xanthomonads using concatenated sections of the gltA, lepA, lacF, gyrB, fusA and gap-1 loci. B) SNP based Neighbor-Joining Tree generated from 17853 variable loci between 14 Xcm isolates and the reference genome *Xanthomonas citri* subsp. *citri* strain Aw12879. The tree was made using the Simple Phylogeny tool from ClustalW2.

**Fig 3: Molecular and phenotypic analysis of Xcm and G. hirsutum interactions.** A) Type three effector profiles of Xcm isolates were deduced from de novo, Illumina based genome assemblies. Effector presence absence was determined based on homology to known type three effectors using the program Prokka. B) Commercial and public *G. hirsutum* cultivars were inoculated with 14 Xcm isolates. Susceptible (S) indicates water soaking symptoms. Resistant (R) indicates a visible hypersensitive response. Plants were screened with a range of inoculum concentration from OD_{600} = 0.001-0.5. C) Disease symptoms on *G. hirsutum* cultivars Stoneville 5288 B2F and DES 56 after inoculation with Xcm strain AR81009 (OD_{600} = 0.05). Symptoms are visualized under visible (VIS) and near infrared (NIR) light. D) The proportion of US fields planted with susceptible and resistant cultivars of *G. hirsutum* was determined based on
planting acreage statistics from the USDA-AMA and disease phenotypes based on previous reports for common cultivars [59-61].

**Fig 4: SMRT sequencing of two phenotypically and geographically diverse Xcm isolates:**

**MS14003 and AR81009.** Circos plot comparing the circular genomes. Tracks are as follows from inside to outside: synteny of gene models; GC Content; Methylation on + and – strands; location of type three effectors (teal) and TAL effectors (red). On each side, accompanying plasmids are cartooned. Type three effector repertoires and the type IV secretion systems were annotated using Prokka, homologous regions greater than 1kb were identified using MAUVE, and TAL effectors were annotated using AnnoTALE.

**Fig 5: SMRT sequencing and western blot reveal diverse TAL effector repertoires between Xcm strains MS14003 and AR81009.** A) Gene models of TAL-effectors identified by AnnoTALE. Blue and Green highlighted gene models represent TALs grouped in the same clade by RVD sequence using AnnoTALE. B) Western Blot of TAL effectors using polyclonal TAL-specific antibody.

**Fig 6: RNA-Sequencing analysis of infected G. hirsutum tissue demonstrates transcriptional changes during CBB.** A) Disease phenotypes of Xcm strains MS14003 and AR81009 on G. hirsutum cultivars Acala Maxxa and DES 56, 7dpi. B) RNA-Seq Experimental Design: Acala Maxxa and DES 56 were inoculated with Xcm strains MS14003 and AR81009 at an OD of 0.5 and a mock treatment of 10mM MgCl₂. Inoculated leaf tissue was collected at 24 and 48 hpi (before disease symptoms emerged). C) Venn diagram of upregulated G. hirsutum genes (Log2(fold change in FPKM) ≥ 2 and p value ≤ 0.05) in response to Xcm inoculation. Cuffdiff output was parsed using a custom script and visualized with the VennDiagram package in R.

**Table 1: RNA-Seq analysis reveals that 8 homeologous pairs of G. hirsutum genes are upregulated in both Acala Maxxa and DES 56 cultivars 48 hours post inoculation with Xcm strains MS14003 and AR81009 at Log2(fold change in FPKM) ≥ 2 and p value ≤ 0.05).**

Homeologous pairs were identified using genic synteny.

**Fig 7: Expression of homeologous pairs across the A and D G. hirsutum genomes in response to Xcm inoculation.** Genes considered up or down regulated meet both differential expression from mock significance of q-value < 0.05 and the absolute value of the log2 fold change is
greater than 2. A) Acala Maxxa inoculated with MS14003 B) DES 56 inoculated with MS14003 C)
Acala Maxxa inoculated with AR 81009 D) DES 56 inoculated with AR81009.

Fig 8: Three candidate G. hirsutum susceptibility genes are targeted by two different Xcm strains. (left) Bioinformatically predicted Xcm TAL Effector binding sites on the 300bp promoter region of four SWEET genes. These were predicted with TALEsf using a quality score cutoff of 4.
(right) Heat-map of Cuffdiff results of significantly upregulated G. hirsutum SWEET genes (p ≤ 0.05) with a Log2 (fold change in FPKM) ≥ 2, 48 hours after inoculation with Xcm.

References


Milestones in Microbiology. 1876;1556.


Supporting Information

S1 Table: Illumina and SMRT sequenced Xcm genomes described in this paper.

S2 Table: Xanthomonas genomes previously deposited on NCBI that are referenced in this paper.

S3 Table: RNA-Seq analysis reveals that 52 genes are induced in all Xcm-G. hirsutum interactions at 48 hours ((p ≤ 0.05) with a Log2 (fold change in FPKM) ≥ 2).

S1 Fig: Maps of CBB incidence in the US from 2011-2012 and 2014-2016. CBB incidence was reported by farmers, Extension specialists and Certified Crop Advisers in their respective states for the years 2011-2012 and 2014-2016, and compiled by Tom Allen. CBB reports for 2013 were infrequent.
S2 Fig: Growth assay of MS14003 and AR81009 on cotton cultivars Acala Maxxa and DES 56.

S3 Fig: Expression levels of significantly upregulated genes with a Log2 fold change of 2 in G. hirsutum

A) All significantly upregulated genes with a Log2 fold change of 2

B) All significantly upregulated genes (p ≤ 0.05) with a Log2 (fold change in FPKM) ≥ 2 that are unique to each cultivar/pathovar disease interaction in G. hirsutum.
### Table A: Type Three Effectors

<table>
<thead>
<tr>
<th>Effector Present</th>
<th>24 Conserved</th>
<th>XopAK</th>
<th>XopC</th>
<th>HopAl2</th>
<th>XopE2</th>
<th>XopJ</th>
<th>XopAl</th>
<th>HopAl1</th>
<th>XopP</th>
<th>XopAO</th>
<th>TAL Effectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>SU58011</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>SU92012</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Race 3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Race 12</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MA81010</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Xcm004</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>AR81009</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>SU44</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Race 2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>BF1</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>NI86</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MS14002</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Race 18</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Xcm014</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>MS14003</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>BF2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>

### Table B: Commercial Varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>R/S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fibermax 989</td>
<td>R</td>
</tr>
<tr>
<td>Fibermax 1830 GLT</td>
<td>R</td>
</tr>
<tr>
<td>Fibermax 2334 GLT</td>
<td>R</td>
</tr>
<tr>
<td>Fibermax 2484 B2F</td>
<td>R</td>
</tr>
<tr>
<td>Deltapine 1133 B2RF</td>
<td>R</td>
</tr>
<tr>
<td>Stoneville 5288 B2F</td>
<td>R</td>
</tr>
<tr>
<td>Deltapine 0912 B2RF</td>
<td>S</td>
</tr>
<tr>
<td>Deltapine 1028 B2RF</td>
<td>S</td>
</tr>
<tr>
<td>Deltapine 1034 B2RF</td>
<td>S</td>
</tr>
<tr>
<td>Deltapine 1048 B2RF</td>
<td>S</td>
</tr>
<tr>
<td>PhytoGen 499 WRF</td>
<td>S</td>
</tr>
<tr>
<td>Stoneville 4946 GLB2</td>
<td>S</td>
</tr>
</tbody>
</table>

### Table C: CBB Panel

<table>
<thead>
<tr>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acala-44</td>
</tr>
<tr>
<td>Gregg</td>
</tr>
<tr>
<td>Mebane</td>
</tr>
<tr>
<td>Stoneville 2B</td>
</tr>
<tr>
<td>Stoneville 20</td>
</tr>
<tr>
<td>NAM</td>
</tr>
<tr>
<td>Acala-Maxxa</td>
</tr>
<tr>
<td>DES 56</td>
</tr>
</tbody>
</table>

### Table D: Percentage of Cotton Acreage Planted

<table>
<thead>
<tr>
<th>Year</th>
<th>Susceptible</th>
<th>Resistant</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>70</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>2010</td>
<td>60</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>2011</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>2012</td>
<td>40</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>2013</td>
<td>30</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>2014</td>
<td>20</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>2015</td>
<td>10</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>2016</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

*Notes: VIS = Visible Spectrum, NIR = Near Infrared Range*
### Expression of Selected SWEET Genes in Mock and Inoculated Cotton, 48 Hpi

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mock (FPKM)</th>
<th>MS14003 (FPKM)</th>
<th>AR81009 (FPKM)</th>
<th>Expression (FPKM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A04G0861</td>
<td>11.83</td>
<td>403.75</td>
<td>2.61</td>
<td>10.73 468.54 10.47</td>
</tr>
<tr>
<td>D04G1360</td>
<td>1.88</td>
<td>190.09</td>
<td>1.60</td>
<td>0.73   279.72 5.52</td>
</tr>
<tr>
<td>A12G1747</td>
<td>117.23</td>
<td>102.12</td>
<td>144.52</td>
<td>92.05  74.32  277.98</td>
</tr>
<tr>
<td>D12G1898</td>
<td>93.21</td>
<td>77.79</td>
<td>888.87</td>
<td>69.40  63.66  850.18</td>
</tr>
<tr>
<td>Acala Maxxa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mock</td>
<td></td>
<td>MS14003</td>
<td>AR81009</td>
<td></td>
</tr>
<tr>
<td>DES 56</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Expression color scale:
- **800**
- **600**
- **400**
- **200**