vol. 182, no. 1 the american naturalist july 2013

E-Article

On the Genetic Architecture of Cytoplasmic Incompatibility: Inference from Phenotypic Data

Igor Nor,[1,2] Jan Engelstädter,[3] Olivier Duron,[4] Max Reuter,[5] Marie-France Sagot,[1,2] and Sylvain Charlat[1,*]

1. Laboratoire Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique (CNRS), Université Lyon 1, Bâtiment Mendel, 43 boulevard du 11 novembre, 69622 Villeurbanne, France; 2. Institut national de recherche en informatique et en automatique, Grenoble Rhône-Alpes, Saint Ismier, France; 3. Institute of Integrative Biology and Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Universitätsstrasse 16, ETH Zentrum, CHN K12.1, 8092 Zurich, Switzerland; and School of Biological Sciences, University of Queensland, Brisbane, Queensland 4072, Australia; 4. Institut des Sciences de l'Evolution, CNRS, Université Montpellier II, Place Eugène Bataillon, CC065, 34095 Montpellier Cedex 05, France; 5. Research Department of Genetics, Evolution and Environment, Faculty of Life Sciences, University College London, Gower Street, London WC1E 6BT, United Kingdom

Submitted May 23, 2012; Accepted January 7, 2013; Electronically published May 7, 2013

abstract: Numerous insects carry intracellular bacteria that manipulate the insects' reproduction and thus facilitate their own spread. Cytoplasmic incompatibility (CI) is a common form of such manipulation, where a (currently uncharacterized) bacterial modification of male sperm induces the early death of embryos unless the fertilized eggs carry the same bacteria, inherited from the mother. The death of uninfected embryos provides an indirect selective advantage to infected ones, thus enabling the spread of the bacteria. Here we use and expand recently developed algorithms to infer the genetic architecture underlying the complex incompatibility data from the mosquito *Culex pipiens*. We show that CI requires more genetic determinants than previously believed and that quantitative variation in gene products potentially contributes to the observed CI patterns. In line with population genetic theory of CI, our analysis suggests that toxin factors (those inducing embryo death) are present in fewer copies in the bacterial genomes than antitoxin factors (those ensuring that infected embryos survive). In combination with comparative genomics, our approach will provide helpful guidance to identify the genetic basis of CI and more generally of other toxin/antitoxin systems that can be conceptualized under the same framework.

Keywords: Wolbachia, cytoplasmic incompatibility, model, toxin-antitoxin systems, gene-for-gene systems.

Introduction

Cytoplasmic incompatibility (CI) is a form of conditional sterility induced by maternally inherited intracellular bacteria in numerous arthropod species (Bourtzis et al. 2003; Engelstädter and Telschow 2009). The bacteria, when present in a male, induce developmental arrest of its offspring unless the fertilized embryo carries the same symbiont, inherited from its mother. This protection confers a fertility benefit to infected embryos and therefore promotes the spread of the infection. CI has been reported in two lineages of bacteria (*Wolbachia* and *Cardinium*; Hunter et al. 2003), where it appears to have evolved independently (Penz et al. 2012).

Despite CI's widespread occurrence, its molecular basis remains elusive (but see Serbus et al. 2008 for reviews on this area). A useful conceptualization of the phenomenon is provided by the mod/resc (modification/rescue) model (Werren 1997). This model proposes that CI involves a toxin (the mod factor), deposited by symbionts in the male's sperm, that induces the death of the zygote unless neutralized by an antidote produced by symbionts present in the egg (the resc factor). The mod/resc model makes no assumption about the actual nature of the mod and resc factors. More-concrete models have been proposed. One of them, the Lock-Key model, assumes that the mod and resc factors are distinct molecules and that the rescue of modified sperm implies a direct interaction between the Lock (produced in sperm) and the Key (produced in the egg). The Lock-Key model currently represents a satisfactory working hypothesis, for it is at the same time specific with regard to interactions between biological molecules (e.g., proteins) and consistent with a number of empirical observations (Poinsot et al. 2003; see "Discussion" for a more detailed presentation of an alternative model proposed by Bossan et al. 2011). Thus, the Lock-Key model can explain not only incompatibility between infected males and uninfected females but also "bidirectional incompatibility," where crosses between males and females carrying different symbionts are incompatible. Bidirectional incompatibility has been observed in a number of

* Corresponding author; e-mail: sylvain.charlat@univ-lyon1.fr.

Am. Nat. 2013. Vol. 182, pp. E15–E24.
DOI: 10.1086/670612

species, a classical example coming from the fruit fly *Drosophila simulans* (fig. 1), where the CI pattern among native *Wolbachia* infections can be explained by assuming that each strain carries its own Lock-Key pair (Merçot and Charlat 2004).

In contrast to this simple case, other incompatibility relationships do not always fit such a straightforward interpretation. This is the case in the mosquito *Culex pipiens*, the species where *Wolbachia* was first described (Hertig and Wolbach 1924) and its causal link with incompatibility first established (Yen and Barr 1971). In this classical study system, the complexity and variability of incompatibility patterns have long been recognized (Laven 1967). The hypothesis that host nuclear variation could be responsible for such complexity, although appealing and theoretically sound (Rousset et al. 1991), has been repeatedly ruled out by empirical data (Ghelelovitch 1952; Laven 1953, 1957, 1967; Barr 1966; Irving-Bell 1983; Duron et al. 2006, 2012; Walker et al. 2009; Atyame et al. 2011; but see Sinkins et al. 2005). Figure 2 shows the compatibility matrix between 19 *C. pipiens* lines compiled from earlier studies (Duron et al. 2006, 2007). Close inspection of the *C. pipiens* data reveals incompatibility relationships that cannot be accounted for by a two-locus Lock-Key mechanism. The colored cells in figure 2 provide an illustration of this. Lines A and B are mutually compatible in both reciprocal crosses (yellow cells). This observation is compatible with the two-locus Lock-Key model, under which we would infer that bacterial strains present in A and B carry the same pair of genes, say $Lock_1$ and Key_1. However, contradictions arise when we include a third line, C, in our analysis. Focusing on the blue cells, we see that C females are compatible with A males, implying that the bacterial strain in line C carries the Key_1 gene; but at the same time C females are incompatible with B males, implying that this bacterial strain would not carry Key_1. In other words, the contradictions between the compatibility patterns observed in A-C and B-C matings show that a simple two-locus Lock-Key model cannot account for the incompatibility patterns observed in *C. pipiens*. The data thus call for an extended model that allows us to explain intransitive relationships such as the ones just described.

A straightforward way of accounting for complex patterns is to extend the Lock-Key model to more than two loci. Thus, the incompatibility relationships between lines A, B, and C above could be explained by assuming that the bacterial strain in B carries an additional Lock ($Lock_2$) at another locus, which is rescued by a Key_2 gene present in A and B but not in C. We would thus infer genotypes {$Lock_1$, Key_1, Key_2}, {$Lock_1$, $Lock_2$, Key_1, Key_2}, and {$Lock_2$, Key_2} for the bacterial strains in A, B, and C, respectively. Another, not mutually exclusive way of extending the Lock-Key model is to allow for quantitative variation of allelic products. Different symbiont strains could thus differ in the quantity of Lock and/or Key gene products, for example, through effects of varying levels of gene expression. In this case, rescue would require the production not only of the right type of Key molecule but also of a sufficient quantity to neutralize all Lock molecules transmitted in the sperm.

	Fem *w*Ha	Fem *w*No	Fem *w*Ri	Fem -
Male *w*Ha	1	0	0	0
Male *w*No	0	1	0	0
Male *w*Ri	0	0	1	0
Male -	1	1	1	1

Figure 1: Incompatibility pattern seen in *Drosophila simulans* between males and females carrying native cytoplasmic incompatibility (CI)–inducing strains (*w*Ha, *w*No, and *w*Ri) or no infection (minus sign). "Fem" stands for "female"; 1 indicates a compatible cross (offspring survival), whereas 0 indicates an incompatible cross. Under the Lock-Key model, this pattern can be explained by assuming that each *Wolbachia* strain carries its own Lock-Key pair, so that only crosses between males and females carrying the same strain are compatible. In addition to these three CI-inducing strains, additional *Wolbachia* strains are found in natural *D. simulans* populations that vary in their ability to induce or rescue CI (e.g., mod−/resc+, or Lock−/Key+ strains; Merçot and Poinsot 1998). This particular phenotype suggests that Locks and Keys are encoded by different genes, but it is readily explained within the frame of the two-locus Lock-Key model, since it does not imply the existence of multiple Lock or multiple Key genes.

While this simple example, involving three *Wolbachia* strains only, can be treated through verbal reasoning, finding the smallest number of Lock and Key factors accounting for a large CI data set becomes a computationally difficult task. Specifically, its algorithmic equivalence to a known NP-complete graph-theoretical problem (which implies that it is computationally intractable for large data sets) was recently demonstrated, but an effective method was developed (Nor et al. 2012). Here we apply this method to the *Culex* data set and explore the biological implications of its outcome as it concerns the genetic architecture and population genetics of CI. We further expand the original binary model (assuming that Lock and Key gene products are either present or absent) to include potential quantitative variation in the gene products. Because of its simplicity, the binary model is more tractable, allowing us to precisely examine the nature and diversity of all the solutions. The quantitative model cannot be analyzed to the same level of detail, but it is potentially important from an empirical point of view. Taken together, our results shed new light on the evolution of incompatibility types and will provide guidance for genomic studies aiming to identify the genetic basis of CI.

Line ID (males \ females)	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
A	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1
B	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1
C	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
D	1	1	1-0	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
E	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
F	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1
G	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
H	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
I	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
J	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
K	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
L	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
M	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
N	0	1	0	1	1	1-0	1	0	0	0	0	1	1	1	1	1	1	1	1
O	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
P	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
Q	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
R	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1
S	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1

Figure 2: *Culex pipiens* compatibility matrix. Rows represent males and columns represent females. Compatibility is indicated by entries of 1 and incompatibility by 0. Cells containing an entry "1-0" indicate crosses producing intermediate and variable hatch rates, possibly corresponding to mosquito lines segregating more than a single *Wolbachia* clone; for this analysis, these crosses are considered compatible (value of 1). Gray cells represent missing data; for this analysis, these cells were assigned a 0 or a 1 on basis of the frequency of 0s and 1s is the remaining cells of the corresponding column and row. Colored cells illustrate one among many cases of nontransitive relationships in this matrix, detailed in the text. The mosquito line IDs correspond to the following original names (Duron et al. 2006, 2007): A, LaVar; B, Bifa-A; C, Bifa-B; D, Kol; E, Keo-A; F, Keo-B; G, Tunis; H, Istanbul; I, Aus; J, Slab; K, MaClo; L, Kara-C; M, Manille-A; N, Manille-B; O, Ep-A; P, Ep-B; Q, Cot-A; R, Cot-B; S, Bismuth.

Results

Binary Model

We first focus on determining the minimum number of Lock and Key loci required to explain the data set given in figure 2 under a "binary" model, that is, a model assuming no quantitative variation of gene products. A detailed description of the algorithm used to solve this problem can be found in Nor et al. (2012). Here, we provide a brief and simplified description of the main steps of the procedure.

The inference method uses input in the form of a compatibility matrix (**C**), an $n \times n$ matrix describing the observed compatibility relationships among n host lines, with males in rows and females in columns. For the *Culex pipiens* data set, the content of the **C** matrix is directly given by figure 2. For each entry C_{ij} of this matrix, a value of 0 indicates that the cross between males of line i and females of line j is incompatible, while a value of 1 indicates that it is compatible. Save for two cases, neglected in this analysis, no intermediate levels of incompatibility are observed in *C. pipiens*, so that a discrete code (0 or 1) is sufficient to describe the data.

On the basis of the compatibility matrix, we aim at determining the most parsimonious pairs of matrices **L** and **K** that describe the Lock and Key factors carried by the symbiont strains present in the n host lines. Both matrices are of dimensions $n \times f$, with n strains and f Lock and f Key factors. The matrices are binary, and for each entry L_{ij} or K_{ij}, 0 and 1 indicate the absence and presence, respectively, of factor j in strain i. We assume that a Key can match only a single Lock and that different Locks and different Keys act independently. A cross between a female of host line i and a male of host line j is compatible only if strain i carries at least all the Key factors matching the Lock factors present in strain j (i.e., $K_{ik} \geq L_{jk} \; \forall \; k \in \{1, 2, \ldots, f\}$). The problem of finding a parsimonious CI genetic architecture for a given incompatibility matrix **C** can therefore be stated as the problem of finding pairs of matrices **L** and **K** that satisfy this condition for all crosses and have a minimum number f of columns.

As recently demonstrated, finding the minimum number of Lock and Key factors is algorithmically equivalent to finding the smallest number of rectangles containing only entries of 0 that are required to cover all entries of 0 in the **C** matrix, that is, to include every entry of 0 in

at least one rectangle (Nor et al. 2012; see fig. 3 for a simple example). Notably, a rectangle of 0s is defined as a set of rows and columns whose intersections are only 0s. In other words, rectangles are not necessarily continuous. As an example, the intersections of rows C and G with columns A and J in the *Culex* **C** matrix (fig. 2) make a rectangle of 0s. This problem is known to belong to the class of NP-complete problems. The following approach, based on parameterized complexity theory, was therefore developed. The analysis begins by reducing, when possible, the size of the input **C** matrix by recursively applying the kernelization rules (Nor et al. 2012). In the particular case of the *C. pipiens* matrix, these rules reduce the matrix enough to be able to find the exact solution to the problem. In other cases, however, a sufficient reduction might not be achievable, and it may be necessary to use the heuristics also described in Nor et al. (2012).

The analysis of the *C. pipiens* data set shows that a minimum of eight pairs of Lock and Key factors are required to explain the observed CI pattern under the binary model. Thanks to the tractability of the binary model, we were able to explore all the solutions to the *C. pipiens* matrix in more detail. This analysis showed that there are exactly 3,976 different Lock and Key matrix pairs that are solutions of minimum size to the incompatibility matrix (this ignores matrix pairs that are permutations of other pairs). In order to gain more insight into the structure of these matrices, we compared their contents across the alternative solutions. First, we assessed matrix contents on a very basic level and compared the number of Key and Lock factors inferred by the different solutions. This analysis showed that Lock matrices generally contain fewer factors than Key matrices. Across all solutions, the minimum excess of Key factors across all strains relative to Lock factors was 75. A similar difference was observed when we looked at the numbers of Key and Lock factors inferred for individual strains (the rows of the **K** and **L** matrices). These varied between 4 and 7 in the **K** matrix and between 1 and 4 in the **L** matrix.

We also made comparisons that were more directly concerned with the structure of the matrices, that is, with how consistent particular matrix entries are across solutions. Comparisons of this kind must take into account the fact that columns of the Lock and Key matrices can be switched, so that column *j* in one solution might not correspond to column *j* in another solution. To solve this problem, we systematically ranked columns in the Lock and Key matrices in the same way, on the basis of the location of the corresponding rectangle of 0s in the **C** matrix. With this ranking, the same columns in different solutions can be regarded as "homologous" genes, with identical or very similar functions. We were then able to assign a consensus score to each entry of a solution, calculated as the proportion of solutions sharing the same value for that particular matrix cell. We then averaged these scores over the cells of a matrix and identified the solution with the highest average score (shown in fig. 4). Inspection of this maximum consensus matrix illustrates that there are more Key than Lock factors (118 and 37, respectively).

C matrix	fem A	fem B	fem C	fem D	fem E	fem F
male A	1	0	0	0	0	0
male B	1	1	1	0	0	1
male C	0	0	1	0	0	1
male D	0	0	0	1	1	1
male E	1	0	0	1	1	1
male F	1	0	0	1	1	1

Lock Matrix	Lock 1	Lock 2	Lock 3	Lock 4
strain A	1	0	1	1
strain B	1	0	0	0
strain C	1	1	0	0
strain D	0	1	1	0
strain E	0	0	1	0
strain F	0	0	1	0

Key Matrix	Key 1	Key 2	Key 3	Key 4
strain A	1	0	1	1
strain B	1	0	0	0
strain C	1	1	0	0
strain D	0	1	1	0
strain E	0	1	1	0
strain F	1	1	1	0

Figure 3: Example illustrating the equivalence between the present problem and the Biclique Bipartite Edge Cover Problem, which can be expressed as the problem of finding the minimum number of rectangles of 0s covering all 0s in a matrix (Nor et al. 2012; note that in that work, as opposed to this study, 0s refer to compatible crosses, following a classical algorithmical presentation). Each rectangle of 0s in the **C** matrix (color coded) is explained by at least one column in the **L** and **K** matrices (corresponding color). Notably, (1) 0s can be contained in a single rectangle even if they are not adjacent in the **C** matrix; in other words, permutations of rows and columns in the **C** matrix are allowed; (2) a given 0 can be contained in more than one rectangle, that is, it can be explained by more than one column in the **L** and **K** matrices.

Quantitative Model

We now introduce the potential for quantitative variation of the Lock and Key gene products. Here, the symbiont strains may differ not only in their gene content but also in the amount of Lock and Key gene products from the different loci. As detailed in the "Discussion," this quantitative model provides a flexible framework within which solutions from the recently proposed "goalkeeper" CI

Strain	Lock 1	Lock 2	Lock 3	Lock 4	Lock 5	Lock 6	Lock 7	Lock 8		Key 1	Key 2	Key 3	Key 4	Key 5	Key 6	Key 7	Key 8
A				1			1			1		1	1	1		1	
B	1			1						1	1	1	1	1		1	1
C		1		1	1		1				1		1	1		1	
D				1						1	1	1	1	1		1	1
E				1						1	1	1	1	1		1	1
F			1							1	1	1	1	1		1	1
G		1		1			1			1	1	1	1	1		1	1
H					1	1					1	1		1	1	1	1
I							1			1		1	1			1	
J				1						1	1	1	1	1	1		
K							1				1	1	1	1		1	
L				1						1	1		1	1		1	1
M				1						1	1		1	1		1	1
N				1			1	1		1	1	1	1	1		1	1
O		1		1						1	1	1	1	1		1	1
P				1	1						1	1	1	1		1	1
Q		1		1			1			1	1	1	1	1		1	1
R		1		1			1			1	1	1	1	1		1	1
S		1		1			1			1	1	1	1	1		1	1

Figure 4: Best confidence solution of the binary model. The value in each cell indicates whether we infer presence (1) or absence (empty cell) of this factor in this particular *Wolbachia* strain. Gray cells contain a value inferred in at least 90% of the solutions of minimum size.

model (Bossan et al. 2011) can also be interpreted. We incorporate quantitative variation into the model by allowing the entries in the **L** and **K** matrices to take nonnegative integer values (L_{ij}, $K_{ij} \in \{0, 1, 2, 3, ...\}$). For a cross between a male carrying strain i and a female carrying strain j to be compatible ($\mathbf{C}_{ij} = 1$), all entries in the **K** matrix of strain j must be equal to or greater than the corresponding entries in the **L** matrix of strain i (i.e., $K_{ik} \geq L_{jk} \; \forall \; k \in \{1, 2, \ldots, f\}$ must hold). If this is not the case, then the cross is incompatible ($\mathbf{C}_{ij} = 0$).

As with the binary case, inferring the minimum number of Lock and Key factors required to explain observed incompatibilities involves rearrangements of the **C** matrix and the detection of 0s clustering into particular shapes. Here we are looking for groups of rows in the **C** matrix that satisfy the following condition: for any pair of rows i and j in this group, either all 0s from row i are also 0s in the corresponding positions in row j (0s from row i are included in row j) or the reciprocal is true: 0s from row j are included in row i. In what follows, we refer to such particular clusters of 0s as "quantitative shapes." Every quantitative shape in **C** can be explained by using a single Lock-Key pair with quantitative variation in the Lock and Key products (see fig. 5 for a simple example and "Methods: Solving the Quantitative Problem" for more details).

Applying the quantitative model to the *C. pipiens* data, we infer that five Lock-Key pairs are required to explain the observed incompatibility patterns. The significant reduction compared to the binary model implies that quantitative variation in Lock and Key products potentially accounts for a significant part of the incompatibility relationships in *C. pipiens*. Figure 6 shows an actual solution, that is, a plausible set of genotypes of the 19 *Wolbachia* strains under this model. Unfortunately, identification of the solutions for the quantitative model cannot be automated for now. As a consequence, we were unable to assess the confidence scores of the solution shown in figure 6. We note, however, that this solution for the quantitative model features strains that again carry only a few (one or two) Lock factors but at least three and up to five Key factors, thus mirroring the enrichment in Key factors observed in the binary model.

Discussion

In this article, we have extended and applied a new method to investigate patterns of symbiont-induced cytoplasmic incompatibility. We analyzed incompatibility data from a large number of reciprocal crosses between lines of the mosquito *Culex pipiens* infected with different *Wolbachia* strains, a prime example of complex incompatibility patterns. Using our algorithm, we were able to make inferences about the genetics underlying the observed patterns of incompatibility between those lines. Under a compu-

C matrix	fem A	fem B	fem C	fem D	fem E
male A	1	0	1	1	0
male B	1	1	1	1	0
male C	1	0	1	1	1
male D	1	0	0	1	1
male E	0	0	0	0	1

Lock Matrix	Lock 1	Lock 2
strain A	2	0
strain B	1	0
strain C	0	1
strain D	0	2
strain E	0	3

Key Matrix	Key 1	Key 2
strain A	2	2
strain B	1	0
strain C	2	1
strain D	2	2
strain E	0	3

Figure 5: Example leading to the inference of quantitative variation in Lock and Key products. In this **C** matrix, all 0s in rows A and B can be included in the blue quantitative shape (defined in the text), and all 0s in rows C–E can be included in the yellow quantitative shape. Accordingly, only two genes are required in the Lock and Key matrices to explain the data, with quantitative variation.

tationally tractable binary Lock-Key model (i.e., where gene products are either present or absent), we were able to analyze the contents and structure of all matrices of minimum size that satisfy the *C. pipiens* incompatibility relationships. This analysis showed that under this simple model, eight distinct Lock-Key pairs are required to account for all observed incompatibilities. There were, however, a large number of alternative **L**-**K** matrix pairs with this minimal number that provided equivalent solutions to the incompatibility matrix **C**. Exploring these solutions in further detail, we found that the inferred solutions always required symbiont strains to carry a significantly larger number of Key factors than Lock factors.

We also analyzed a more sophisticated model for incompatibility that allowed for quantitative differences in the efficacy of CI genes. Accounting for quantitative effects significantly reduced the number of Lock-Key pairs required to explain incompatibility in *C. pipiens*, the minimum number inferred dropping to five. This result suggests that effects of dosage potentially account for a good proportion of the incompatibility relationships we observe, at least in *C. pipiens*.

As with the binary model, the Lock matrices inferred under the quantitative model contain many more empty cells than do the Key matrices (figs. 4, 6). Specifically, we infer under the quantitative model that most bacterial strains carry a single Lock factor (all strains except in lines B and H), while all strains carry at least three Key factors (fig. 6). A plausible interpretation of this pattern is that a large part of the Lock variability among strains is allelic variation, with different strains carrying different alleles at a given Lock locus (see fig. 7 for an illustration). The presence of several Key factors in each strain, on the other hand, is not compatible with allelic variation and must be interpreted as being (at least partly) the result of variation in gene content. Interestingly, this interpretation is compatible with current theory for the evolution of CI types (Charlat et al. 2001, 2005; Engelstädter et al. 2006) and a model of evolution in which new Lock or Key types arise by point mutations, whereas new genes are created by duplication events. In order to understand the evolution of incompatibility under this model, consider a randomly mating host population fixed for CI bacteria containing a single pair of Lock and Key loci $\{Lock_1, Key_1\}$. In this situation, any point mutation in a symbiont's Key locus, producing a $\{Lock_1, Key_2\}$ strain, would render host females incompatible with all males present in the population and thus be quickly eliminated by purifying selection (see fig. 8*A*). In contrast, point mutations in the Lock gene producing a $\{Lock_2, Key_1\}$ strain do not alter the compatibility patterns of females carrying this new symbiont and thus are neutral in randomly mating populations (fig. 8*B*). The presence of such neutral mutants in the population then paves the way for the evolution of a matching Key factor, Key_2, which could arise through duplication of the existing gene Key_1, followed by divergence through point mutation (fig. 8*C*). Importantly, this duplication and diversification of Key genes is favored by selection, because any strain that is compatible with both Locks segregating in the population renders carrier host females compatible with all males and hence confers a fecundity advantage to its hosts. These arguments show that allelic variation is not expected to occur at Key loci. Variation at Lock loci, in contrast, is possible and can then trigger the diversification of Key genes. These conclusions are consistent with the inferred structure of the Key and Lock matrices presented here.

Recently, a new theoretical interpretation of CI, called the "goalkeeper model," has been proposed as an alternative to the Lock-Key model (Bossan et al. 2011). The goalkeeper model assumes that each symbiont strain contributes two distinct factors in the same quantity in both eggs and sperm. In addition, the female host provides a certain quantity of each factor in the eggs. A cross is then assumed to be incompatible if the amount of at least one factor produced in the sperm exceeds the amount present in the egg. This model was shown to be consistent with a number of findings concerning CI and could explain the

Strain	Lock 1	Lock 2	Lock 3	Lock 4	Lock 5		Key 1	Key 2	Key 3	Key 4	Key 5
A		**						**	**	**	*
B			*		*		*****	**	**	*	*
C	****						****	**	**		
D			*				*****	**	**	*	*
E			*				*****	**	**	*	*
F				*			****	**	**	*	*
G	****						*****	**	**	*	*
H	**			**			**	*		**	*
I		*					*	**	*	**	*
J			*				***		**	**	*
K		*					****	**	**	*	
L			*				*****	**	**		*
M			*				*****	**	**		*
N	*****						*****	**	**	*	*
O	***						*****	**	**	*	*
P			**				*****	**	**	*	
Q	****						*****	**	**	*	*
R	****						*****	**	**	*	*
S	*						*****	**	**	**	*

Figure 6: An output of the quantitative model. For clarity, relative amounts of gene products are symbolized by the number of asterisks in cells. Empty cells indicate that no gene product is inferred from the analysis. Each color symbolizes a single Lock-Key pair.

compatibility relationships between six *Wolbachia* strains in *Drosophila simulans* (i.e., the three native and three artificially introduced strains). Like our quantitative Lock-Key model, the goalkeeper model assumes that CI patterns partially rely on quantitative variation among strains. Moreover, numerical solutions from the goalkeeper model (e.g., see fig. 4 in Bossan et al. 2011) can be interpreted as solutions of our quantitative model. The x_a and y_a parameters from Bossan et al. (2011) are equivalent to the amounts of Lock factors that would be expressed by two Lock genes; the combined contributions of *Wolbachia* and the hosts in females ($x_h + x_a$, $y_h + y_a$) are equivalent to the amounts of Key factors expressed by two Key genes. Notably, the difference between the amounts of Lock and Key gene products in our model is thus equivalent to the host contribution (x_h and y_h) in the goalkeeper model. The x_h and y_h parameters represent host properties and are therefore constant, by definition, across all *Wolbachia* strains. Thus, the goalkeeper model can be seen as a special case of our more general quantitative Lock-Key model. Accordingly, any solution from the goalkeeper model is equivalent to a solution of the quantitative model, but the reverse is not true: solutions from the quantitative Lock-Key model can be translated into goalkeeper solutions only if, at each locus, the difference in the amounts of Lock and Key gene products is fixed across all strains (a condition hereafter referred to as the "goalkeeper condition").

This partial equivalence between the models can be exemplified by running our quantitative model on the *D. simulans* data set, which can produce a variety of solutions with two Lock-Key pairs, among which some, but not all, fulfill the goalkeeper condition (data not shown). Our results thus generally concur with the goalkeeper model to suggest that quantitative differences in gene products can explain a good part of the variation in compatibility. However, our analysis also suggests that more than two Lock-Key pairs are required to account for the *C. pipiens* data, implying that this model in its present formulation would fail to account for such complex incompatibility patterns.

One critical assumption of our analysis is that a single Key matches a single Lock, whether the latter is encoded by different genes or by different alleles of a single gene. Alternatively, some Key factors could have a larger spectrum, that is, match more than one Lock, as previously envisaged (Werren 1998). Such "Master Keys" should be favored by natural selection, although trade-offs between efficiency and spectrum width might constrain the process. Integrating such variation in the spectrum of Key and Lock factors could potentially reduce the number of factors required to explain a CI data set. Addressing this issue will require the development of new algorithms, since the problem, under this slightly different formulation, might radically differ mathematically.

Although our approach was specifically designed with

Strain	Lock 1	Lock 2
A	**	
B	*	*
C	****	
D	*	
E	*	
F		*
G	****	
H	**	**
I	*	
J	*	
K	*	
L	*	
M	*	
N	*****	
O	***	
P	**	
Q	****	
R	****	
S	*	

Figure 7: Allelic variation in Lock factors. Here we assume that each Lock locus can carry more than one allele and that different alleles (different colors) can be matched by different Key loci (matching colors). The figure shows the resulting reduction of the quantitative model solution shown in figure 6. Columns 1–3 from the Lock matrix in figure 6 can be combined in a single locus, since these Lock factors never coexist within a strain. Similarly, columns 4 and 5 can be combined in a single locus, so that only two Lock loci are required. Conversely, the Key matrix remains as in figure 6, because columns 1–5 from the Key matrix cannot be combined.

cytoplasmic incompatibility in mind, our algorithm could in principle also be applied to other toxin-antitoxin interactions. One area where our approach might prove fruitful is the analysis of host-parasite interactions under the so-called gene-for-gene model (Flor 1955). This type of interaction has been reported in a number of plant-pathogen systems (Burdon 1987; Thompson and Burdon 1992) and was the subject of several theoretical studies of host-parasite coevolution (e.g., Parker 1994; Otto and Nuismer 2004; Salathé et al. 2005). In the gene-for-gene model, both parasites and hosts are characterized by a number of loci that determine whether a given parasite genotype is able to infect a given host genotype. At each locus, hosts can have either a resistance allele or a nonresistance allele. Parasites can also have two alleles at each locus: a virulence allele and an avirulence allele (following the plant pathogen terminology, virulence is defined here as the ability of the pathogen to develop in a host, regardless of its cost to this host). It is then assumed that a parasite can infect a host if and only if none of its avirulence alleles is matched by a resistance allele at a corresponding locus in the host. This implies that a parasite carrying only virulence alleles can infect all hosts and that hosts carrying only nonresistance alleles can be infected by all parasites.

CI and the gene-for-gene model are equivalent in the following way. Incompatible crosses in CI correspond to failures of parasites to infect a host, Lock alleles correspond to resistance alleles in hosts, and Key alleles correspond to virulence alleles in parasites. Thus, the algorithm presented in this article can in principle be used to obtain information about the number of genes involved in host-parasite interactions that are assumed to follow the as-

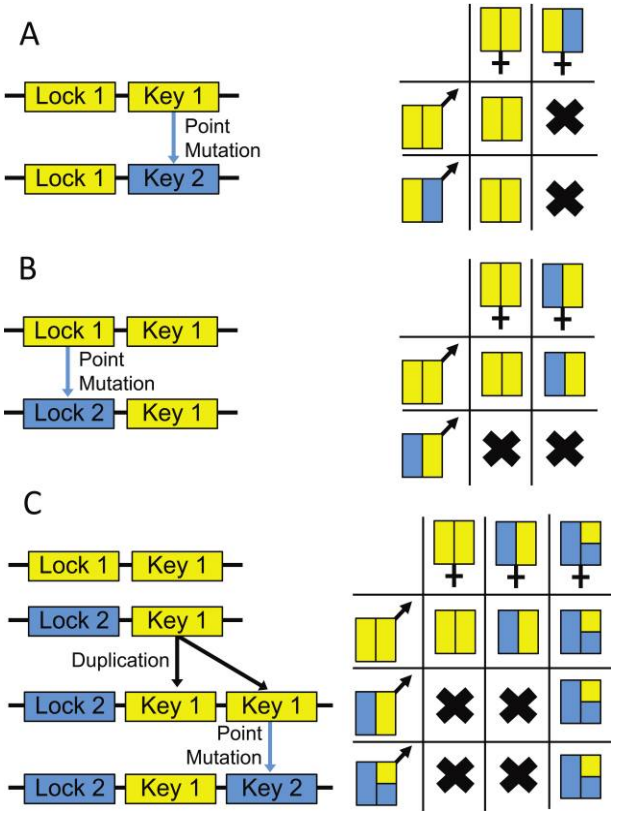

Figure 8: Fate of point mutations and duplications affecting the Lock and Key loci. The left-hand part of the figures shows the mutational events under consideration. The right-hand part shows the crossings occurring in the population following the emergence of the mutants. The Lock and Key factors are color coded: yellow for Lock 1 and Key 1, blue for Lock 2 and Key 2. The Lock and Key properties of the males and females present in the populations are displayed on the left- and right-hand parts, respectively, of the male and female symbols. *A*, Point mutations affecting the Key are deleterious and therefore eliminated by purifying selection. *B*, Point mutations affecting the Lock are neutral. *C*, Duplication of a Key locus followed by point mutation can produce strains carrying Keys for more than one Lock; this can be selected if the population is polymorphic for the Lock.

sumptions of the gene-for-gene model. Traditionally, this question has been addressed by quantitative trait locus analyses that map the genes involved onto specific chromosome locations (reviewed in Wilfert and Schmid-Hempel 2008); the number of genes involved in the interaction is then a by-product of a much more detailed set of information concerning the genetic architecture of the interaction. As the method presented here is based on phenotypic data only (i.e., which parasite strains infect which host strains), it potentially provides a quicker and less expensive first notion of how many genes are (at least) involved in the host-parasite interaction under study.

In the absence of adequate genetic tools to transform *Wolbachia* genomes and thereby assess gene functions, comparative genomics offers the best alternative to identify the genes involved in CI. The *C. pipiens* system, where closely related *Wolbachia* strains differ in their CI properties, seems ideal in this context: genomic regions differing between these strains would represent candidate CI genes. The analysis presented here, which generates explicit predictions regarding the Lock-Key profile of the different strains, would provide explicit guidance in this process. More generally, we believe that our approach of predicting the genetic architecture of toxin-antitoxin interactions from phenotypic data represents a valuable complement to comparative genomics to identify the genetic basis of such phenomena.

Acknowledgments

We thank the CNRS Institut écologie et environnement (INEE; Action Thématique et Incitative sur Programme [ATIP] grant SymbioCode held by S.C.); the Natural Environment Research Council UK (grants NE/D009189/1 and NE/G019452/1, held by M.R.); the Agence Nationale de la Recherche (ANR; project MIRI BLAN08-1335497, held by M.-F.S.); the European Research Council (ERC), under the European Community's Seventh Framework Programme (FP7/2007-2013/ERC grant agreement [247073]10, held by M.-F.S.); and the Swiss National Science Foundation (grant PZ00P3_132934, held by J.E.).

APPENDIX

Methods: Solving the Quantitative Problem

Under the quantitative model, finding the smallest number of Lock and Key factors required to explain a given **C** matrix is equivalent to finding the smallest number of quantitative shapes (defined in "Quantitative Model") to cover all 0s in the **C** matrix. We applied an algorithm analogous to the previously used "isolated locations" (Nor et al. 2012), allowing us to determine the lower bound of the solution. This requires determining the maximum number of 0s in the **C** matrix that are "strictly isolated," that is, that cannot be included in the same quantitative shape. Two 0s in positions $C_{i1,j1}$ and $C_{i2,j2}$ are strictly isolated if and only if we have 1 at $C_{i1,j2}$ and $C_{i2,j1}$. For example, in the *C. pipiens* case, $C_{3,1}$ and $C_{2,3}$ are strictly isolated, while $C_{3,1}$ and $C_{8,2}$ are not. To identify not only the lower bound but also actual **L** and **K** solutions to the problem, we constructed quantitative shapes initiated from strictly isolated 0s until all 0s were covered; specifically, the base of each quantitative shape contains one strictly isolated 0.

Literature Cited

Atyame, C. M., O. Duron, P. Tortosa, N. Pasteur, P. Fort, and M. Weill. 2011. Multiple *Wolbachia* determinants control the evolution of cytoplasmic incompatibilities in *Culex pipiens* mosquito populations. Molecular Ecology 20:286–298.

Barr, A. 1966. Cytoplasmic incompatibility as a means of eradication of *Culex pipiens* L. Proceedings and Papers of the California Mosquito Control Association 34:32–35.

Bossan, B., A. Koehncke, and P. Hammerstein. 2011. A new model and method for understanding *Wolbachia*-induced cytoplasmic incompatibility. PLoS ONE 6:e19757.

Bourtzis, K., H. R. Braig, and T. L. Karr. 2003. Cytoplasmic incompatibility. Pages 217–246 *in* K. Bourtzis and T. Miller, eds. Insect symbiosis. CRC, Boca Raton, FL.

Burdon, J. J. 1987. Diseases and plant population biology. Cambridge Studies in Ecology. Cambridge: Cambridge University Press.

Charlat, S., C. Calmet, O. Andrieu, and H. Merçot. 2005. Exploring the evolution of *Wolbachia* compatibility types: a simulation approach. Genetics 170:495–507.

Charlat, S., C. Calmet, and H. Merçot. 2001. On the *mod resc* model and the evolution of *Wolbachia* compatibility types. Genetics 159: 1415–1422.

Duron, O., C. Bernard, S. Unal, A. Berthomieu, C. Berticat, and M. Weill. 2006. Tracking factors modulating cytoplasmic incompatibilities in the mosquito *Culex pipiens*. Molecular Ecology 15:3061–3071.

Duron, O., J. Bernard, C. M. Atyame, E. Dumas, and M. Weill. 2012. Rapid evolution of *Wolbachia* incompatibility types. Proceedings of the Royal Society B: Biological Sciences 279:4473–4480.

Duron, O., A. Boureux, P. Echaubard, A. Berthomieu, C. Berticat, P. Fort, and M. Weill. 2007. Variability and expression of ankyrin domain genes in *Wolbachia* variants infecting the mosquito *Culex pipiens*. Journal of Bacteriology 189:4442–4448.

Engelstädter, J., S. Charlat, A. Pomiankowski, and G. D. D. Hurst. 2006. The evolution of cytoplasmic incompatibility types: integrating segregation, inbreeding and outbreeding. Genetics 172: 2601–2611.

Engelstädter, J., and A. Telschow. 2009. Cytoplasmic incompatibility and host population structure. Heredity 103:196–207.

Flor, H. 1955. Host-parasite interaction in flax rust: its genetics and other implications. Phytopathology 45:680–685.

Ghelelovitch, S. 1952. Sur le déterminisme génétique de la stérilité

dans les croisements entre différentes souches de *Culex* autogenicus Roubaud. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences 234:2386–2388.

Hertig, M., and S. B. Wolbach. 1924. Studies on *Rickettsia*-like microorganisms in insects. Journal of Medical Research 44:329–374.

Hunter, M. S., S. J. Perlman, and S. E. Kelly. 2003. A bacterial symbiont in the *Bacteroidetes* induces cytoplasmic incompatibility in the parasitoid wasp *Encarsia pergandiella*. Proceedings of the Royal Society B: Biological Sciences 270:2185–2190.

Irving-Bell, R. 1983. Cytoplasmic incompatibility within and between *Culex molestus* and *Cx. quinquefasciatus* (Diptera: Culcidae). Journal of Medical Entomology 20:44–48.

Laven, H. 1953. Reziprok unterschiedliche Kreuzbarkeit von Stechmücken (*Culicidae*) und ihre Deutung als plasmatische Vererbung. Zeitschrift für Induktive Abstammungs- und Vererbungslehre 85 118–136.

———. 1957. Vererbung durch Kerngene und das Problem der ausserkaryotischen Vererbung bei *Culex pipiens*. Zeitschrift für Induktive Abstammungs- und Vererbungslehre 88:478–516.

———. 1967. Speciation and evolution in *Culex pipiens*. Pages 251–275 *in* J. Wright and R. Pal, eds. Genetics of insect vectors of disease. Amsterdam: Elsevier.

Merçot, H., and S. Charlat. 2004. *Wolbachia* infections in *Drosophila melanogaster* and *D. simulans*: polymorphism and levels of cytoplasmic incompatibility. Genetica 120:51–59.

Merçot, H., and D. Poinsot. 1998. ...and discovered on Mount Kilimanjaro. Nature 391:853.

Nor, I., D. Hermelin, S. Charlat, J. Engelstädter, M. Reuter, O. Duron, and M.-F. Sagot. 2012. Mod/Resc parsimony inference: theory and application. Information and Computation 213:23–32.

Otto, S. P., and S. L. Nuismer. 2004. Species interactions and the evolution of sex. Science 304:1018–1020.

Parker, M. A. 1994. Pathogens and sex in plants. Evolutionary Ecology 8:560–584.

Penz, T., S. Schmitz-Esser, S. E. Kelly, B. N. Cass, A. Müller, T. Woyke, S. A. Malfatti, M. S. Hunter, and M. Horn. 2012. Comparative genomics suggests an independent origin of cytoplasmic incompatibility in *Cardinium hertigii*. PLoS Genetics 8:e1003012.

Poinsot, D., S. Charlat, and H. Merçot. 2003. On the mechanism of *Wolbachia*-induced cytoplasmic incompatibility: confronting the models with the facts. Bioessays 25:259–265.

Rousset, F., M. Raymond, and F. Kjellberg. 1991. Cytoplasmic incompatibility in the mosquito *Culex pipiens*: how to explain a cytotype polymorphism? Journal of Evolutionary Biology 4:69–81.

Salathé, M., A. Scherer, and S. Bonhoeffer. 2005. Neutral drift and polymorphism in gene-for-gene systems. Ecology Letters 8:925–932.

Serbus, L. R., C. Casper-Lindley, F. Landmann, and W. Sullivan. 2008. The genetics and cell biology of *Wolbachia*-host interactions. Annual Review of Genetics 42:683–707.

Sinkins, S. P., T. Walker, A. R. Lynd, A. R. Steven, B. L. Makepeace, H. C. Godfray, and J. Parkhill. 2005. *Wolbachia* variability and host effects on crossing type in *Culex* mosquitoes. Nature 436:257–260.

Thompson, J., and J. J. Burdon. 1992. Gene-for-gene coevolution between plants and parasites. Nature 360:121–125.

Walker, T., S. Song, and S. P. Sinkins. 2009. *Wolbachia* in the *Culex pipiens* group mosquitoes: introgression and superinfection. Journal of Heredity 100:192–196.

Werren, J. H. 1997. Biology of *Wolbachia*. Annual Review of Entomology 42:587–609.

———. 1998. *Wolbachia* and speciation. Pages 245–260 *in* D. Howard and S. Berlocher, eds. Endless forms: species and speciation. Oxford: Oxford University Press.

Wilfert, L., and P. Schmid-Hempel. 2008. The genetic architecture of susceptibility to parasites. BMC Evolutionary Biology 8:187–194.

Yen, J. H., and A. R. Barr. 1971. New hypothesis of the cause of cytoplasmic incompatibility in *Culex pipiens*. Nature 232:657–658.

Associate Editor: Kimberly A. Hughes
Editor: Troy Day

Hatching *Culex pipiens* eggs resulting from a compatible cross. Photograph by Olivier Duron.