European Research Council

ERC Synergy Grant 2020

[redacted]

Understanding and predicting PATHOgen COMmunities

PATHOCOM

Corresponding Principal Investigator	Detlef Weigel
Host institution	Max-Planck-Gesellschaft zur Förderung der Wissenschaften
Other Principal Investigator	Fabrice Roux
Host institution	Centre National de la Recherche Scientifique
Other Principal Investigator	Joy Bergelson
Host institution	University of Chicago
Proposal duration	72 months

Infectious disease is often the **major selective agent** in nature, and we cannot understand how populations evolve without understanding their **pathogenic microbes**. Beyond host immunity, an important factor determining the ability of pathogens to invade and proliferate in a host is the resident **microbiota**, but we are only beginning to glimpse its multifarious impacts. We know even less about interactions among pathogenic microbes themselves. Any effort to explain how **pathogen communities**, or **pathobiota**, develop within a host requires knowledge about the extent to which pathogens engage in **competition**, **commensalism** and **cooperation**, both with other pathogens and the rest of the microbiota. To date, studies of plant-pathogen interactions in the lab and descriptive work in the field have **focused** on **pairwise interactions** between one plant host and one pathogen, leaving a large gap in our understanding of how **different types** of **interactions between microbes**, and especially pathogens, determine the **outcome** of **host-pathogen interactions** in the **real world**. In PATHOCOM, we will implement a program that **integrates large-scale field observations** of microbes in the plant ***Arabidopsis thaliana*** with **ultra-high-throughput experimental tests** of host-dependent interactions among microbes, allowing **experiment-informed modeling** of pathogenic microbe-microbe interactions. These models, which will be improved through an iterative process of data collection with **synthetic communities**, will illuminate how **interactions**, from pairwise to **higher-order**, shape microbial community composition and structure. In the final step, the resulting **models** will be tested against and refined with field data. Together, these efforts will transform the study of plant pathogens by applying deep analyses of microbial interactions in an ecological context to **explain patterns** in **nature**. Ultimately, PATHOCOM will play an essential role in **refashioning plant-pathogen-microbiome studies** into a **predictive science**.

ERC Synergy Grant 2020 Research proposal [Part B2]

- Weigel, Detlef **Max-Planck-Gesellschaft zur Förderung der Wissenschaften**
- Roux, Fabrice **Centre National de la Recherche Scientifique**
- Bergelson, Joy **University of Chicago**

Part B2: *The Project Proposal*

Section a. State of the art and objective

Motivation and goals of PATHOCOM

Infectious disease is often the **major selective agent** in nature, and we cannot understand how natural populations evolve without understanding their pathogenic microbes. Much less is known about pathogens in **natural** compared to **agricultural environments,** but it is generally thought that the two are very different: wild plants are genetically and developmentally less uniform, they typically occur interspersed with other species, and they suffer from noticeable disease more rarely[1–3]. Despite these differences, **diverse pathogen communities** seem to be **common** in both settings[4–11], reminiscent of mutualistic interactions between pathogens in human disease[6,12–14].

In wild plant pathosystems, even very recent invasions harbor genetically distinct invaders[15–17]. Genetic diversity is also seen in more established plant-pathosystems[18–22], where the severity of epidemics is associated with higher levels of co-infection[7,23]. Such a picture is emerging for the plant ***Arabidopsis thaliana***, a **model** for **evolutionary** and **ecological genetics** and **genomics**. In this host, **diversity** in the pathogenic microbiota, or **pathobiota**, is apparent within local host populations, within individual plants, and within individual pathogen taxa[24,25]. Importantly, we have found that the fraction of pathogenic isolates in *A. thaliana* leaves is correlated with the diversity of co-infecting isolates[25]. In other words, as **pathogens overtake** a **community**, they often appear to **do so as a group**[9,26,27]. A related observation is that genetically diverse sublineages of a single *Pseudomonas* lineage have co-occurred for hundreds of thousands of years[24] and jointly dominate *A. thaliana* bacterial communities in the Southwest of Germany. These results suggest a paradigm in plants where one of the greatest risks for severe disease results from reciprocal help among pathogens.

Such recent findings suggest new opportunities for intervention during early phases of host infection, for example, by interfering with cooperative interactions among pathogens. As a first step towards such a long-term goal, we aim to understand the nature of **pathogen-pathogen interactions** within **realistic community contexts**. Our overarching goal is to reveal the **ecological** and **genetic principles** enabling particular pathogens to **invade** a **microbial community,** notably by assessing how **pathogen-pathogen interactions change** as a **function** of **habitat qualities, genetic diversity** of **co-occurring commensal bacterial species,** and **host genetic diversity** in **different geographical locations**.

Arabidopsis thaliana and its leaf microbes are particularly well suited for such an endeavor because natural populations are readily accessible, infections are easily carried out in the lab, and many molecular mechanisms of plant-microbe interactions are known in exquisite detail. We will decipher the **forces** that **shape** the **pathobiota**, within the leaves (known as the **phyllosphere**) of *A. thaliana* by combining **extensive field observations**–determining patterns of co-occurrence among microbial species and pathogenic strains in the real world–with **systematic laboratory studies**–focusing on high throughput phenotyping of microbe-microbe interactions–to inform **models** that can **produce expected patterns** in **pathogen communities**.

Central to our approach is the hypothesis that an understanding of community level processes will enable prediction from smaller to larger scales, both at the level of individual plants and plant communities. Thus, at the core of our strategy are models to be developed through an iterative process including validation based on strategically designed synthetic communities. These models will ultimately be used to **generate broad**, **qualitative patterns** in **natural conditions** that can be used to make sense of observational data.

PATHOCOM addresses several gaps in our understanding of microbe interactions within hosts. In particular, previous work in plants has typically focused on either single microbial strains (alone or in pairwise combinations[28]), or has used complex synthetic communities[29,30], with poorly understood intra-community interactions. In addition, **previous lab work** has been largely, if not entirely, **disconnected from patterns observed** in **natural populations**. With the rich set of experiments that we will perform, we will **pioneer** a **mechanistic understanding** of **observed patterns** of **pathogen-pathogen associations** in **nature** across spatial and temporal scales. PATHOCOM's specific goals are to:

- Characterize *A. thaliana* and its associated **(patho)microbiota** in a **large, hierarchically structured sample** in the native and introduced range, across multiple seasons and years.
- Determine the **spectrum** of **interactions** among large isolate collections from three pathogenic bacteria and the most frequent commensal species, and identify genes underlying these interactions.

- Determine how **environment**, **plant genetics**, and **microbiota genetics** affect pathogen-pathogen interactions, and **identify specific genes** from both **host** and **microbe** that impact these interactions.
- Integrate the **experimental** and **genetic** data and develop a **model** that enables the **testing** of **key ecological** and **genetic drivers** of pathobiota community structure.
- **Compare** output from our experiment-based **model** with **patterns** of association in **natural conditions** as functions of microbial interactions, microbial genetics, plant genetics and the environment, and use this information to improve the model.

In the following, we will introduce our host-microbiota system; discuss experimental evidence for pathogen-pathogen cooperation *in planta*; introduce a recent model allowing us to predict coexistence in communities; provide background on plant control of microbes; and illustrate the power of Genome-Wide Association (GWA) mapping for disentangling the genetic basis of host-pathobiota-microbiota interactions. As appropriate, we will focus on our own past research that has laid the foundation for PATHOCOM.

The *Arabidopsis thaliana* – microbiota system

Arabidopsis thaliana and its associated microbes serve as an outstanding model system for this work. This wild plant occurs naturally throughout much of **Europe, Asia** and **Africa**, where it has maintained persistent populations in both pristine and disturbed habitats for tens of thousands of years[31–34]. In contrast, *A. thaliana* invaded **North America** only ~400 years ago, and there it is restricted to highly disturbed habitats and depauperate in genetic variation[35]. The three Principal Investigators have extensive experience working with *A. thaliana* **in** the **lab**, in **field experiments** and ***in situ***, including leading the *A. thaliana* 1001 Genomes Project[32,33,36] and developing and applying GWA mapping methods in this species[37–47]. We have also generated extensive collections of and information about microbes that colonize *A. thaliana*[24,25,43,48–53], including 1000s of samples, collected from Germany, France and the US, that provide ample material to start experiments immediately. We have learned that although ***Sphingomonas*** sp. do not seem to include pathogens, they nevertheless comprise a dominant group (often >20%) within the bacterial microbiome of *A. thaliana* leaves. The three most abundant pathogenic species (~75% of the pathobiota) in *A. thaliana* leaves are the ***Pseudomonas*** *syringae* complex (including the pathogenic species *P. syringae* and *P. viridiflava*)[24], ***Xanthomonas*** *campestris* and ***Pantoea*** *agglomerans*[25] (although not all strains of all three species are necessarily expected to be pathogenic). Fungi and oomycetes can be found within the microbiota of *A. thaliana*[43,54–56], but we have chosen here to focus on **bacteria** due to (i) their dominance in *A. thaliana* microbiota by number and biomass[56,57], (ii) disease symptoms in natural populations of *A. thaliana* mainly resulting from pathogenic bacteria[25], and (iii) their **genetic tractability**, which will be essential for our ultra-high-throughput infection tests. We will, however, obtain information on eukaryotic microbes in our field sampling and can include such information in our models

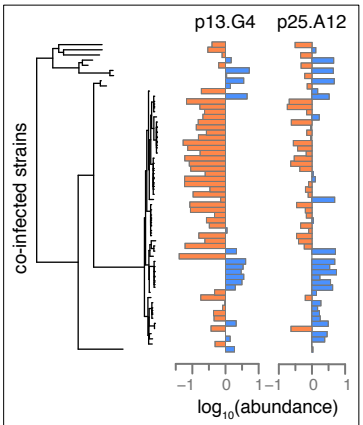


Figure 1. Pairwise co-infections of luciferase-tagged *P. viridiflava* strains isolated from *A. thaliana* and 60 other isolates whose phylogenetic relationship is shown on the left. Abundances of the focal strains in co-infections are expressed relative to single infections.

Detecting positive pairwise and higher-order microbial interactions

Similar to the microbiome in animals and humans, **background microbiota** can diminish the ability of plant pathogens to cause disease, and research on such biocontrol agents has a long history in plants[54,58–64]. At the same time, there is growing appreciation of **synergistic**, **positive pathogen-pathogen interactions**[27], including taxa that occur frequently on *A. thaliana*[8,65,66]. However, with very few exceptions[30,54], there have not been **systematic studies** of the **genetic basis** of **isolate-isolate interactions**, and none *in planta*.

We have developed **ultra-high-throughput tools** to quantify **interactions between bacterial isolates** in leaves of gnotobiotic *A. thaliana*. For a proof of concept, we co-inoculated *A. thaliana* with two *P. viridiflava* strains expressing luciferase and 60 randomly chosen strains from the *P. syringae* complex (Fig. 1). Luciferase activity was measured after 36 hours to quantify abundance of the focal strains. Two key results reveal that **co-infection** strongly impacts pathogen performance in *A. thaliana*: (i) the **mean growth** of the focal strain

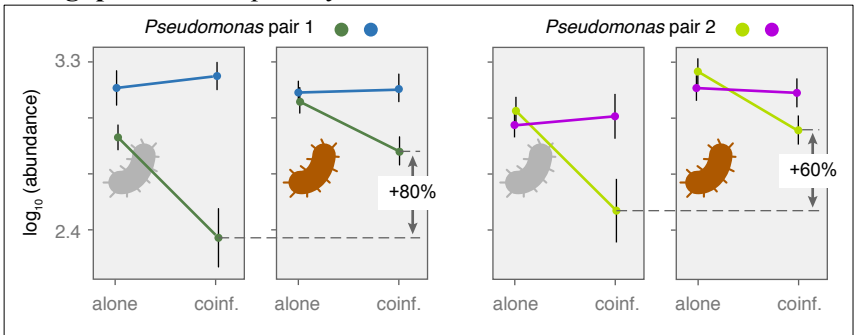


Figure 2. Three-way co-infection reveals a positive impact of a *Sphingomonas* sp. strain (indicated in brown) on the performance of competitively inferior *P. viridiflava* isolates (green) during intra-specific competition. The generality of this phenomenon remains to be tested.

can differ by **two orders of magnitude** between the most and least favorable coinfection combinations; (ii) there are both **costs** and **benefits** of **co-infection**–in more than a third of pairwise coinfections, the focal strain grew to a higher abundance than when singly inoculated, whereas it decreased in abundance in many of the other combinations. Importantly, these effects are **highly consistent,** dwarfing experimental noise. The **identity** of the **co-infecting strains** explained **over 70%** of the **variance** in focal strain abundance. This underscores the importance of accounting for **microbe genotype x microbe genotype interactions**, as we propose to do, when predicting infection outcomes. We have furthermore found that, in two distinct sets of strains, the presence of *Sphingomonas* reduces the strength of competition between *P. viridiflava* isolates (Fig. 2).

Predicting coexistence in communities

Determining how **populations interact**, and predicting how **ecological systems evolve** over time or after perturbations, is a main goal of ecology. Historically, methods to parameterize models of population dynamics have been developed using time-series data[67–69]. But when dealing with experimental communities in laboratory conditions, other approaches are possible[70]. Our Chicago co-investigator Allesina has recently developed a **method** to **parameterize ecological interactions** using measured abundances in a handful of communities[71]. Briefly, the goal is to predict whether a community S consisting of k strains (out of a pool of n) will coexist when co-cultured, and if so to predict the abundance of all coexisting strains. One can write a set of equations relating the abundance of strain i with all the other strains in the community S of size n:

$$x_i^{(S)} = a_i + \sum_{j \in S; j \neq i} b_{ij} x_j^{(S)}$$

where $x_i^{(S)}$ is the abundance of strain i when grown in community S, a_i the average abundance of strain i when grown alone, and b_{ij} the average effect of strain j, if present, on strain i. Importantly, one can write a similar equation for each of the n strains in S, and for each community S that is observed experimentally. This means that (i) if one is able to observe a sufficiently large number of communities, **all n^2 parameters can be fit**, and (ii) **larger communities provide more information** (i.e., allow more equations to be written) than smaller communities. Exploiting this idea, Allesina showed that large communities can be parameterized using a small number of experiments, provided that each experiment probes a different community[71]. Under these conditions, approximately $2n$ experiments are sufficient to parameterize the model when n strains are considered.

A

community	Ea	Pa	Pp	Pv
Ea	71			
Pa		241.9		
Pv				72.7
Ea–Pp	88.8		8.2	
Ea–Pv	83.6			19.5
Pa–Pp		140.2	5.8	
Pa–Pv		182.4		29.5
Pp–Pv			29.9	20.8
Ea–Pa–Pv	7	39		56
Pa–Pp–Pv		171.8	6.3	27

B

	Ea	Pa	Pp	Pv
Ea	–1.25	–8.75	0.91	0.65
Pa	–10.17	–1.17	–1.14	–0.01
Pp	–0.72	–1.21	–1.01	–0.01
Pv	–0.77	–0.70	–0.71	–1.03

C

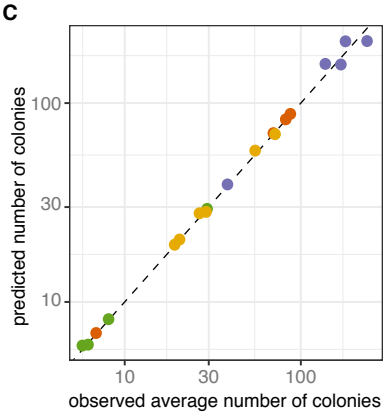


Figure 3. Application of our model[71] to published data from ref. 72. Different combinations of *Enterobacter aerogenes* (Ea), *Pseudomonas aurantiaca* (Pa), *P. putida* (Pp), and *P. veronii* (Pv) were cultured in batch cultures with serial dilutions (here, 1000x dilution). (A) Average colony count, a proxy for abundance, for each experiment. (B) The data from (A) data were fit using our method[71], producing the matrix of parameters shown here, encoding the outcomes of all possible experiments. (C) The predictednumber of colonies for all experiments closely matched the observed data. Note that the model also makes predictions for experiments that were not used to fit the model. For example, the model predicts that Ea and Pa cannot coexist, a prediction that agrees with the experimental data[66].

Allesina applied the method to three data sets from the literature, all from natural systems (albeit not microbial), with excellent fit[71]. An additional example, for bacterial communities, is given in Fig. 3, where we used experimental data from a published paper[72] to fit our model. The parameters of the model were organized in a matrix that summarizes the relationships between the abundance of the strains in all combinations. The **fit** was also **excellent**. Crucially, the matrix can be used to predict the coexistence and abundance of communities that were not used to fit the model–and in turn these predictions can be tested experimentally.

Host control of the phyllosphere

Microbes can be found both on leaf surfaces and inside leaves; together, the **epiphytic** and **endophytic compartments** constitute the **phyllosphere**. The endophytic compartment is more favorable, as microbes are protected from UV and can more easily access nutrients from the host. Microbes including pathogens can enter the endophytic compartment either by mechanical means, or through stomata (the leaf valves for air exchange) and hydathodes (water secreting pores in leaf blade or margin). While microbes can manipulate their hosts to gain entry, there is strong evidence that *A. thaliana* consistently **filters microbes** from the soil, a major source of phyllosphere microbes[73]. We have furthermore found strong differentiation among microbiomes resident in leaves, fruits, roots and stems–another indication that the host influences the microbes that reside within it.

While it is clear that the host leaf environment affects phyllosphere composition, the relative importance of *A. thaliana* genetics in shaping natural variation in microbial communities has been a matter of debate[74]. Although host genetics was first considered of minor importance in shaping natural variation in microbial communities in *A. thaliana*[75,76], we recently found **high heritability** estimates for the **success of individual members** of the **phyllosphere microbiota** as well as for **microbial community traits**[43,77,78]. Finally, it is well established that variation in host disease resistance genes has very clear effects on serious pathogen infection in crops[79,80]. While selection is almost certainly a stronger force acting on specialized pathogens attacking crops grown in monoculture than on the generalists common on *A. thaliana,* these findings nonetheless suggest that host genetics shapes the phyllosphere (patho)microbiota.

Mapping genetic factors shaping microbial communities

Genome-Wide Association (GWA) is a powerful tool for discovering genomic regions associated with natural variation of disease resistance in both wild and cultivated plants[81]. Our team has **pioneered** the **successful application** of **GWA** studies in *A. thaliana* by (i) developing mapping populations at various geographical scales[31–33,38,45]; (ii) developing new statistical methods of GWA mapping including one for simultaneous GWA mapping on two interacting species[41,47,53]; (iii) fine-mapping genomic regions associated with disease resistance in both growth chamber/greenhouse and field conditions[37,38,41,42,46,82,83]; and (iv) functionally validating candidate genes[40–42,82–84], thereby revealing diverse molecular mechanisms underlying disease resistance beyond typical NLR immune receptors[74,85]. While informative, most of these GWA studies have been conducted in the context of **'one host–one pathogen strain'**, which ignores the fact that plants simultaneously interact with many microbes.

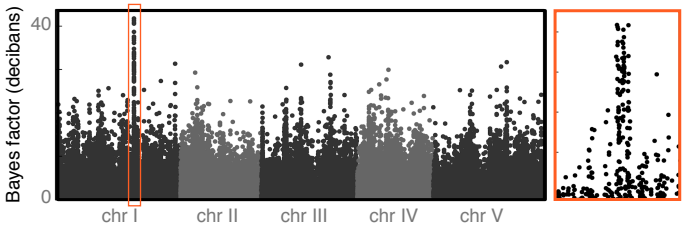


Figure 4. GWA for pathobiota richness in leaves of 168 *A. thaliana* strains. Left, Manhattan plot of all 5 chromosomes; significance estimated with a Bayesian hierarchical model implemented in BayPass. Right, close-up of region on chromosome 1. The top 40 SNPs are spread over 10 kb containing three genes.

We recently found that the **genetic architecture** of **disease resistance differs** when plants are **co-infected** with **multiple pathogens** instead of a single pathogen. In one specific case, the plant response to co-infection could not be accurately predicted from the response to infection by each of two single *X. campestris* strains, and fewer than 50% of the top SNPs identified with the mono-infections were shared with those found after co-infection. In another example, when we characterized the bacterial pathobiota of 168 natural strains of *A. thaliana* from near Toulouse[25], we detected well defined association peaks for pathobiota richness in leaves, but not for individual pathogens, supporting the importance of co-infection in shaping genetically encoded host responses (Fig. 4).

Not only is host genetics important in shaping microbial communities, but **microbial genetics plays** a **role** as well. GWA mapping in bacteria is gaining popularity[81], as it provides a rapid means of identifying genes important in adaptation. We are particularly interested in genes underlying microbe-microbe interactions within and between *Sphingomonas* sp., *Pantoea agglomerans,* the *Pseudomonas syringae* complex and *Xanthomonas campestris.* We have therefore analyzed 87 to 483 whole genome sequences per group to calculate the extent of linkage disequilibrium (LD), an important determinant of the power of GWA. In all cases, LD decays to $r^2 = 0.2$ within an average of 1 kb, which will easily support fine mapping of genes. In an initial mapping experiment, we sought genes underlying the impact of 220 *P. viridiflava* strains on a focal *P. viridiflava* strain. A variety of strong candidate genes could be identified, of which two are shown in Fig. 5.

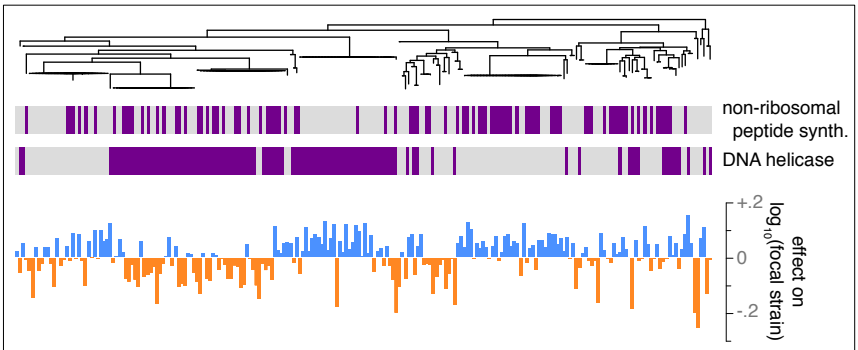


Figure 5. GWA for the effect of co-infection on the growth of a focal (luciferase-labelled) *P. viridiflava* isolate. Note that the candidate encoding a DNA helicase is associated with a strong phylogenetic signal, but the other candidate is not. Purple indicates presence of each of the candidate genes in the genomes of strains from the phylogeny at the top.

The next step

Altogether, our previous work puts us in a strong position to identify and understand the **ecological** and **genetic mechanisms** that determine **how microbial pathogens interact** when infecting plants **in natural conditions**. Over the last 25 years, we have learned much about the molecular machinery with which plants can detect

pathogens, and about the downstream defense reactions plants mount in response. At the same time, we have obtained a deep understanding of a subset of the pathogen molecules (from a few species) that trigger these defenses, as well as the molecules that pathogens use to suppress host responses. This body of work provides **rich context** for investigating indirect interactions among microbes that are mediated by the host. In parallel, we have found that microbial isolates interact strongly within *A. thaliana*, with both positive and negative outcomes. Furthermore, there is evidence for **pairwise**, **three-way**, and likely even **higher-order interactions**.

Progress on multiple fronts notwithstanding, there are **important gaps** in our knowledge of plant-pathogen interactions, **especially** with **endemic pathogens**. We not only have limited insight into how molecules identified in the lab are deployed in natural conditions, but we also do not understand well how different types of interactions between pathogens determine the **outcomes** of **infection** in the **real world**. 16S rDNA and similar amplicon surveys of background microbiota have mostly led to the conclusion that "things are complicated". Here, we propose to fill these gaps in our understanding of real-world plant-pathogen interactions, in an ambitious program with the following **Specific Aims**:

1. Geographically **structured characterization** of *A. thaliana* and its complex **(patho)microbiota**
2. **Experimental** characterization of a spectrum of simplified pathogen-pathogen **interactions** *in planta*
3. Building a **model** of **persistent communities** from empirical pathogen-pathogen interactions
4. Experimental characterization of **(a)biotic factors modulating** pathogen-pathogen **interactions**
5. Applying the **model** in an **ecological genomics framework**

PATHOCOM relies on an **integration** of **large-scale field observations**; **ultra-high-throughput experimental tests** of interactions among microbes; **experiment-informed modeling** of (microbe)^n-host interactions; and **comparing** the output of these models with **real-world data**. Importantly, this integration will be conducted at both **ecological** and **genetic levels**. Together, our efforts will transform the field by transitioning from the study of small-scale networks to deeper analysis of **higher-order interactions,** and by extrapolating from local interactions to **broad community-level patterns**. We envision that PATHOCOM will ultimately become a catalyst for **refashioning plant-pathogen-microbiome studies** into a **predictive science**.

The Need for Synergy

The idea for an ERC Synergy grant was born from the exemplary complementarity of the three main partners: Bergelson trained as an **ecologist** and later on moved into plant genetics. She pioneered the study of *A. thaliana*-microbe interactions in the field, especially with the use of organisms designed to test the genetic basis of species interactions. Roux trained as an **evolutionary and quantitative geneticist**. He has pioneered GWA experiments in semi-natural and natural conditions (which often produced very different results in terms of causal genes and genetic architecture than comparable indoor experiments). He contributes a strong community ecology emphasis to this proposal, an aspect that has largely been ignored by conventional microbiome studies. Weigel trained as a **molecular geneticist**. He has greatly advanced the plant genetics field through his increasingly detailed efforts to describe and understand the causes and consequences of **genomic variation** in *A. thaliana*. Weigel's efforts in turn have been the basis for the **innovative microbiome GWA** approaches developed by Bergelson and Roux. We are united by our interests in **natural host-microbe interactions**, coming either from the host/microbiome (Bergelson), the community/microbiome (Roux), or the pathogen/microbiome (Weigel) side. Starting from different avenues, we have **converged** on **similar questions** that **cannot** be **answered** by a **conventional research team alone** because of the **required scale** and **integration** of **efforts**. Central to this is the **comparative approach**, studying different geographic regions and different sets of microbes, an approach that is urgently needed to overcome a major weakness of much of the plant microbiome field, which has almost always focused on temporally and spatially limited data, making it impossible to arrive at generalizable conclusions.

PATHOCOM is not only of a scale that goes beyond the capability of an individual research group, but it also requires integration of a **team** that **combines diverse expertise** with **access** to **specialized facilities** and **well-understood field sites**. A key element is the effort to link diversity of important foliar microbes with diversity of their plant host and associated ecological variables in three distinct geographic regions, near each of the Principal Investigator's home institutions in France, Germany and the US. We have established **networks of natural sites** with ***A. thaliana* populations**, some of which we have followed for over 20 years. Through this foundational work, the three Principal Investigators have extensive experience with the behavior of *A. thaliana* in the field in different years and seasons. To obtain field data that are comparable between sites, all populations have to be closely monitored; to do this for several seasons would be infeasible at sites that are not within easy driving distance of the involved labs.

The Principal Investigators have also established **large microbe collections** from local *A. thaliana* populations, with an emphasis on **sites** with **well-understood host genetics**; these efforts set our work apart from others in the plant microbiome field. The **microbial interaction data** that will be generated are of a **scale** that

is beyond the capability of any single academic group. To ensure that data generated in a distributed fashion are directly comparable and thus **suitable** for **extensive mathematical modeling**, it is essential that **experimental methods** and **sampling strategies** are applied in an **identical manner** in the three groups, which can only be achieved by tight integration of experimental practices via multiple mutual, extended visits and secondments of students and postdocs over substantial periods of time.

Finally, it needs to be emphasized that the proposed effort would not be possible without integration of the **US-based Principal Investigator**. The US *A. thaliana* populations play a key role in the project because of their unique genetics, and Bergelson has also a dedicated field station in Michigan for **true field experiments** in which genetically modified organisms are grown and microbe infections carried out outdoors, in a habitat where naturalized *A. thaliana* populations occur. Similarly important are Bergelson's local **co-investigators**, McPeek and Allesina, who are an integral part of our team; they have committed their time and expertise in supervising three dedicated postdoctoral fellows, two of whom who will begin work in Chicago and then continue in Europe. This strategy provides us with **specialized modeling expertise**, generating further synergy. Transfer of such expertise to the European team members provides significant added value.

Section b. Methodology

At the heart of our project is an ambitious attempt to understand and explain **patterns of association among diverse pathogens** in the phyllosphere of wild plants. Several **unique features** distinguish our project from other efforts in the field. First, we will generate data that provide an exquisitely nuanced picture of the *A. thaliana* pathobiota, across **temporal** as well as **local**, **regional** and **continental scales**, at a resolution that enables us to track not only different species, but also genetic diversity among isolates. We will furthermore test **hundreds** of **thousands** of **interactions**, both within and between microbial species, to determine the extent of competition and cooperation/facilitation, and the factors favoring either of them. We will also deeply probe the importance of **higher-order interactions** among microbial species. In all of this work, our approach will be **two-tiered** (Fig. 6). First, we will consider **ecological drivers** of microbial communities. Second, genome sequences as well as GWA mapping results will allow us to examine the structure of the same microbial communities from a **genetic perspective**. **Unifying** these **two approaches** is an overarching goal of our project. Finally, the close interplay between model and data collection–as summarized below–will not only help us to constantly challenge our understanding of pathobiota interactions, but also enable us to work towards the ability to **predict community level patterns** starting with knowledge of local interactions.

In brief, PATHOCOM contains **three major elements** (Fig. 6), which we will address with five Specific Aims. First, we will characterize **natural patterns of (patho)microbiota** within a structured set of *A. thaliana* plants from 20 populations each in France, Germany and the US across 'three year x two season' combinations. A wide variety of environmental variables will be collected at these locations as well (Aim 1). In parallel, we will experimentally study **interactions** among a diverse range of **microbial isolates**, to determine which behave **cooperatively** versus **competitively** using **ultra-high-throughput infection assays**. Because there is much greater power to understand interaction strengths when considering absolute rather than relative abundances, we will complete these tests on **genetically barcoded plants**, a **major innovation** that increases the throughput of infection assays by **two orders** of **magnitude** (Aim 2). We will determine the genetic basis of pairwise microbe-microbe interactions using **joint GWA analysis,** and confirm candidate microbial genes. This characterization of microbe-microbe interactions at both the ecological and genetic levels will provide data for the first step in building a **community-scale model** of microbial associations in the host (Aim 3), and we will refine the model to include appropriate higher-order interactions through a targeted series of experiments using **microbial communities** of **increasing complexity**. To quantify the robustness of particular pairwise microbe interactions, we will then test the impacts of **environmental quality**, **genetic variation** of the ***A. thaliana*** host, and genetic variation in **other microbial associates** (Aim 4). In the final step, we will **generate** the **set** of **persistent communities** expected under our model assumptions, and compare the model output to **field data** from Aim 1. Such a comparison will provide a critical test of our understanding of the **key ecological** and **genetic drivers** of **pathobiota structure** (Aim 5).

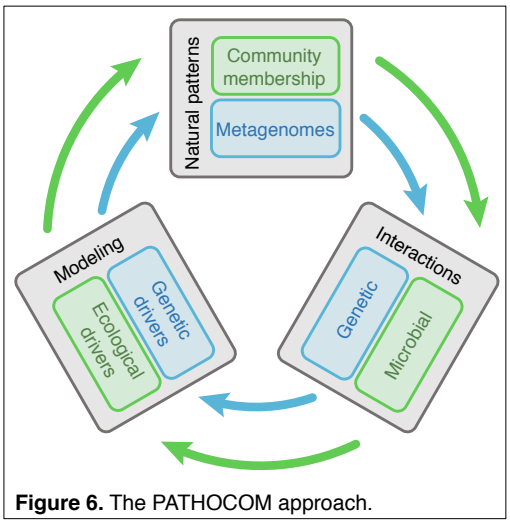


Figure 6. The PATHOCOM approach.

Aim 1: Geographically structured characterization of *A. thaliana* and its complex (patho)microbiota

Rationale: *Under Aim 1, we will generate* ***foundational data****, characterizing the intra- and interspecific diversity as well as abundance of* ***pathobiota*** *and* ***commensal microbiota*** *across* ***multiple populations*** *of* A. thaliana *in three different geographic regions, together with information on environmental variables and host genotypes. We will do this in sufficient scale and with enough detail that it allows inferences about broad patterns, such as presence and absence of large taxonomic groups[5,54], as well as interactions between genomic variants. Importantly, we will generate information both about* ***absolute microbial load*** *and the* ***presence of specific pathogen genes,*** *allowing us to link ecological and genetic drivers of community structure.*

Sampling the intra- and interspecific diversity of *A. thaliana* (patho)microbiota: A substantial weakness of most **prior studies** on microbiota diversity in crop and wild plants has been their inclusion of only **a limited number of individuals** and/or a **limited number of geographic sites**, with few exceptions[24,25,86,87]. Here, we will sample ***A. thaliana*** **metapopulations** in **three regions** that provide **geographic** and **genetic contrasts**. Near Toulouse, France, there is high genetic diversity in both the host[88] and bacterial pathobiota, with *Xanthomonas*, *Pseudomonas* and *Pantoea* (closely related to *Erwinia*[89]) being common[25]. The region around Tübingen, Germany, is characterized by an intermediate host genetic diversity and a pathobiota that is dominated by *Pseudomonas*[24,90]. Finally, in Michigan, USA, *A. thaliana* was introduced only after the year 1600; there is very low host genetic diversity[35] and the pathobiota is dominated by *Pseudomonas* and *Xanthomonas*[48,91].

We will sample 20 sites in each of the 3 regions over 6 consecutive seasons (fall/spring), recording the following environmental parameters: local **weather** (with data loggers); **soil** agronomic properties (including pH, water holding capacity, concentration of N_2 and main mineral nutrients[92]); **density** and **size variation** of ***A. thaliana*** plants; **density** and **total mass of vegetation**; taxonomy of **neighboring plants** (with a metabarcoding approach[93–95]). We will also characterize the microbiome environment by harvesting and pooling leaf punches from 50 companion plants; these will be processed for bacterial 16S and eukaryotic ITS1 rDNA amplicon sequencing, to qualitatively assess the bacterial, fungal and oomycete phyllosphere community. We will similarly assess the soil microbiome at our sites. We will sample the epiphytic and endophytic compartments of a total of **3,600 *A. thaliana* plants** by collecting 10 individuals for each of the 360 site/season combinations. We will collect plants before the onset of flowering by randomly selecting plants within a specific size range, recording signs of infection such as chlorosis, water soaking, necrotic spots, oomycete and fungal reproductive structures.

Discovering (patho)microbiota variation by WGS and amplicon sequencing: To identify microbial taxa most likely to be pathogenic, one needs to know not only which taxa are most common in a sample, but also whether they are abundant in absolute terms. Notwithstanding the effects of rare taxa, plant pathogens typically only affect the plant if they accumulate to appreciable levels[96]. The absolute quantification of individual microbial taxa greatly improves the inference of networks, with more correlations among genera becoming detectable. Similarly, causal effects of the gut microbiome on the human host can only be detected with knowledge of **absolute microbe levels**[97]. There have been few attempts to **quantify pathogen levels** *in planta* while simultaneously characterizing the **background microbiota**. In one of our recent studies, for example, absolute microbial load in wild *A. thaliana* plants could be directly predicted by the presence of a specific taxonomic group of *Pseudomonas* strains[24].

For these reasons, we will prepare not only **16S** and **ITS1** rDNA amplicons, but also **whole-genome shotgun (WGS) sequencing libraries** of **plants** and their **endophytic communities** using an in-house transpososomes-on-bead method. WGS reads will first be mapped against the *A. thaliana* reference genome along with PacBio reference genomes for local populations and our database of nearly complete NLR catalogs from over 60 strains[98] (providing host genotype information and allowing us to normalize sample amount). We will taxonomically assign the remaining reads with a pipeline we have established[55]. Pilot experiments in Germany and Sweden indicate that **most plants produce ~5% microbial reads**, and 1.2 Gb sequence provides excellent power for **species-level assignments**[55] (Fig. 7). Absolute microbial load will be determined as the ratio of bacterial to plant reads, and load for each taxon will be estimated from the ratio of their reads relative to total microbial reads, scaled by average genome size of each taxonomic group. We have also developed methods to

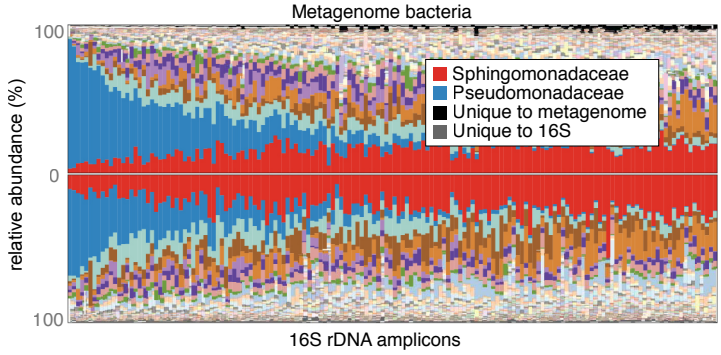


Figure 7. Agreement of taxonomic assignments based on whole-genome shotgun sequencing (top) with taxonomic assignments based on 16S rDNA. The two most common taxa across this data set, Sphingomonadaceae and Pseudomonadaceae, are highlighted.

use the WGS data for estimating absolute abundance of taxa detected only in the 16S/ITS1 rDNA amplicon data[55]. For the epiphytic communities, we will perform only amplicon analyses, since the goal here is merely to determine how much more predictive they are for endophytic colonization in comparison with microbiota from neighboring plants. Finally, we will analyze samples from the first two seasons (1,200) by shallow RNA-seq to assess how well microbial taxonomy, genetics and load predict the induction of markers for PAMP- and effector-triggered immunity, the two main layers of the plant immune system[99].

Discovering (patho)microbiota variation by PEN-seq: To generate a deep understanding of variation in our three pathogenic taxa (*Pseudomonas*, *Xanthomonas* and *Pantoea*) and *Sphingomonas* to discover potential genetic variants responsible for interactions within and between them, we will use **pathogen enrichment sequencing, PEN-seq**[11,100]. *Sphingomonas* will be included, because this bacterial genus is not only often one of the most abundant bacterial taxa in the phyllosphere of wild *A. thaliana*[55,56], but it can also protect *A. thaliana* against *Pseudomonas* induced disease[58], and it can modify *Pseudomonas* competitive interactions (see *Section a*). Based on our complete genome sequences of local isolates of the four taxa, we will build corresponding **baits** based on consensus sequences of the core genomes and genetic elements such as effectors and secretion systems that are likely to be important for pathogen success, targeting about **1.5 Mb** per **genus**[24,89,101–103]. We will enrich DNA of the four focal bacterial taxa from pools of barcoded WGS libraries, aiming for 80x read depth per sample. Reads will be mapped against the genomic components represented by the baits, to determine both SNP and presence/absence (P/A) variation. We will infer linkage between variants for each species within each sample using polymorphism frequencies[104], to ascertain not only the **presence** and **relative levels** of different **genes** and genetic **variants**, but to estimate also the **diversity at the strain level** for our focal taxa.

Discovering factors affecting pathobiota diversity: Using bilinear factors models such as Partial-Least Square Regression, we will determine how **abiotic variables** as well as richness and composition of **background microbiota** and **plant communities** affect both absolute levels and genetic diversity of our focal microbes, at the level of individual plants, sites, seasons and geographical regions. Finally, we will infer microbial **networks**[105,106] and test for **more complex effects** of pathobiota-associated microbial communities. The goal of this effort is to identify the **major axes** of **(a)biotic variation** that affect our focal microbes, which we then will vary experimentally in the infection trials described in Aim 4.

Anticipated knowledge gained: (i) Rich insights into the **pathobiota**, **microbiota** and **environment** of our host species *A. thaliana*; **(ii) differences** in the microbiota and pathobiota in **native** and **introduced ranges** of the host; **(iii) differences** in microbiota and pathobiota structures when assessed at the level of **genes** versus **taxa**; **(iv) host resistance genes** and **bacterial effectors** affecting patterns of association–perhaps identifying drivers of co-evolution; **(v)** knowledge of the **robustness** of **pathobiota community structure**; and **(vi)** discovery of key **abiotic** and **biotic factors** affecting **pathobiota diversity**.

Contingency planning: The proposed methods are already in place, and we have long experience with the field sites that we will sample. The greatest risk is that individual A. thaliana *populations disappear or sites are destroyed by development, but in each of the regions, we can easily choose from additional sites. We estimate that over the course of three years, fewer than 4 of the 20 sites in each region are at risk. A final risk is that our PEN-seq enrichment baits do not include all causal genes, but core genome relatedness should still capture at least in part patterns of sharing of causal non-core genome genes, plus we can resort to the WGS data for additional gene discovery (although these will be relatively low coverage).*

Aim 2: Experimental characterization of the spectrum of microbe-microbe interactions

Rationale: *In any ecological network,* ***highly diverse interactions*** *can be observed at both intra- and inter-specific levels. In Aim 2, we seek to characterize the full matrix of interactions among 600 focal isolates. There are two motivations for this ambitious experiment. First, we know very little about the relative frequency of* ***competition*** *(both partners are harmed),* ***commensalism*** *(one partner benefits, the other is not impacted),* ***asymmetry*** *(one partner benefits, the other is negatively impacted) and* ***reciprocal help*** *(both partners benefit) in any system; we are particularly interested in the* ***prevalence*** *of* ***mutually beneficial interactions****. Second, this matrix of interactions provides the first building block for predicting which assemblages of isolates should persist, under the assumption that microbial interactions are the main driver of community structure. We think of these predictions as a* ***useful null distribution*** *for comparison to* ***natural communities*** *measured in Aim 1. We will estimate this matrix in two ways, once measuring pairwise interactions in the absence of other isolates and once in their presence; the comparison of these matrices will guide our search for higher-order interactions in Aim 3. A major innovation of our work will be that we will not only use genetically barcoded microbial strains, but also* ***genetically barcoded plants****, which will support the analysis of an unprecedented number of parallel infections.*

A system for discovery of microbe-microbe interactions: We have established extensive collections of bacterial strains from *A. thaliana* in France, Germany and the US over the past decade[24,25,42,43,48,49,51,53] (including the 3 x 20 sites of Aim 1), and have tested many strains for their ability to cause disease in our gnotobiotic system. We will test interactions among **150 strains** each of *P. agglomerans*, *X. campestris, P. syringae* complex, *Sphingomonas* sp., with one third originating from each geographic region. (The number of 150 strains is based on power estimates for subsequent GWA studies[53,107]). As laid out above, we will include *Sphingomonas*, both because it is often a dominant taxon in wild *A. thaliana*[55,56], and because they can affect competitive interactions between *Pseudomonas* strains (see *Section a*). ***Sphingomonas*** can thus be thought of as a **proxy** for the **non-pathogenic background microbiota**. Note also that only *X. campestris* is not expected to comprise non-pathogenic strains. From each geographic region, 50 strains of each taxon will be chosen based on **phylogenetic** diversity[25,108,109] and diversity of **habitats** from which they derive, prioritizing sites in which the four species naturally **co-occur**.

Ultimately, we wish to learn which of the almost 2^{600} possible multi-member communities (each strain can be present or absent) can persist. To this end, we will generate empirical data, by infecting gnotobiotic plants with barcoded bacterial isolates, as described in detail below. We will approach the universe of possible interactions in two ways. First, in a "bottom-up" approach, we will test **all pairwise interactions** among our **600 strains**, and infect plants with each of the 179,700 possible pairs (including pairs of differently barcoded derivatives of the same strains; 4 replicates per pair). In addition, all single-strain infections will be tested with 10 replicates. Second, we will take advantage of recent work by our co-investigator Allesina, who has devised a highly efficient way of calculating the complete matrix of interactions among strains. In this "top-down" approach, one begins with **complex mixtures** that are **free to collapse** to their endpoints[71]. We will adopt this strategy by infecting (i) 12 replicates with three 200-member regional communities (containing 50 strains from each of the four bacterial taxa); (ii) 12 replicates with four 150-member species communities (containing 50 strains of the same bacterial taxon from each of the three geographic regions); and (iii) 500 unique, 150-member communities randomly drawn from all 600 isolates, tested without replicates, as the replication here comes from each strain being present in on average 125 communities. In total, we will infect over 700,000 plants in this set of experiments.

While it might be tempting to consider only the more efficient **"top-down" strategy** starting with complex mixtures, we need the results from the **pairwise tests** for **GWA mapping** of genes underlying microbe-microbe interactions. Note also that measurement of the complex mixtures is considerably less efficient than that of pairs because of the much higher number of barcodes present in each starting mixture. A comparison between the matrices from these two strategies will nevertheless be invaluable in **guiding** our exploration of **higher-order interactions** required for generating a null distribution of feasible communities in Aim 3.

Barcoding bacterial isolates and plants: To achieve this massive phenotyping, which will be divided among the three groups, we will use an **ultra-high-throughput method** that we have recently developed to measure the **interactions** between **bacterial isolates** within **gnotobiotic** *A. thaliana.* The bacterial strains will be marked with "barcodes", which are short, unique DNA sequences. Similarly, we will construct genetically **barcoded plants**. Barcodes can be PCR amplified via flanking primer binding sequences that are the same in all strains, and the amplicons counted after Illumina sequencing. The ratios of bacterial to plant barcode sequencing reads then provide information on absolute microbe abundances scaled to the number of plant cells in each infection.

The **mini-Tn7 system**[110–112] has successfully yielded site-specific chromosomal integrations in all four of our bacterial taxa[113–115], and we have already transformed three of them in our own labs (Fig. 8). Sequences integrated downstream of the *glmS* gene are stably maintained and do not impose fitness costs[110]. We will use a published high-throughput method[116] to insert the barcodes, and we will **whole-genome sequence** the transformed strains on the **PacBio** platform (the genome sequences will also be used for joint GWA mapping, see below).

A **major innovation** of our work is that infections will be carried out with a collection of *A. thaliana* **plants** that are genetically identical except for **unique barcode sequences** that have been integrated into their genomes, using a similar principle as for the bacterial isolates (Fig. 8). We will use the HPG1 strain, the most common strain among *A. thaliana* populations in the US[35], to facilitate the use of a local strain for US-based field tests in Aim 4;

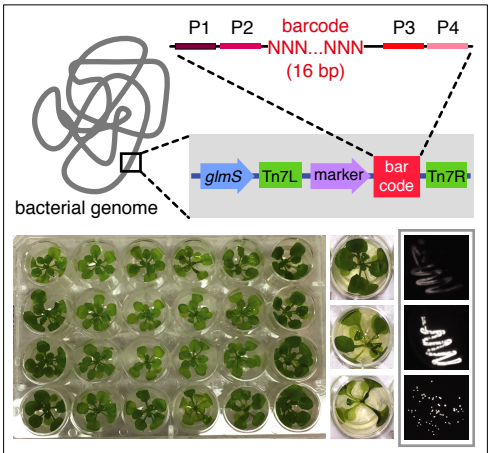


Figure 8. Top, barcoding strategy for microbes. P1/2/3/4 indicate PCR primers flanking barcodes integrated at the *glmS* locus (after ref. 110). Bottom left, gnotobiotic assay system. Bottom middle, close-ups of lightly, moderately and heavily diseased plants 36 hours after infection with different *P. viridiflava* strains. Bottom right, Examples of *X. campestris*, *Sphingomonas* sp. and *P. syringae* isolated from *A. thaliana* and transformed with a luciferase construct with the same backbone as planned for barcoding.

we note that this strain belongs to a genetic group common in Western Europe[33]. At least 192 barcodes will be integrated into a single chromosomal position that does not affect plant fitness, using an improved CRISPR/Cas9 system[117]. These lines are currently being constructed in collaboration with Daisuke Miki (Shanghai) and Michael Desai (Harvard).

Ultra-high-throughput infections: We have developed an **automated protocol for infection** using a pipetting robot with which we can infect >1,500 plants/day (requiring ~9 months' work for each of the three groups). Amplicon sequencing will be used to quantify the barcodes in each infection. Because the barcodes are present in a single copy per genome, the **ratio** of **bacterial** to **plant barcode counts** gives a direct measure of the absolute abundance of each strain (in units of bacterial genomes per host plant genomes). Because every plant line and bacterial strain is uniquely barcoded, plant material from 192 infections can be **pooled for DNA extraction** and **PCR** and converted into a **single sequencing library**, thereby increasing the throughput of infection assays by almost **two orders of magnitude**. Libraries can in turn be pooled for sequencing. We will aim for ~16,000 reads per infected plant for the pairwise infections, and correspondingly more reads for the more complex infections, which should provide excellent precision (most reads will be from the host plant).

The results from barcode counting will be classified qualitatively as one of six outcomes: competition (--), reciprocal help (++), independence (00), asymmetry (+-), commensalism (0+), and amensalism (0-). We will also calculate a competitive score s_i of each strain i as its mean fraction f_{ij} after co-inoculation with each of the n-1 interacting strains. The relevance for **whole-plant outcomes** will be confirmed by assessing disease state of infected plants for the most dramatic interactions using imaging. We will extract rosette surface area as a proxy for plant biomass from the images, with RGB color analysis revealing the visible disease state of infected plants. From this, we will determine whether potential differences in disease between mono- and pairwise infections outcome align with interactions inferred from barcode counting.

Predicting interactions from bacterial genetics: To ascertain the genetics underpinning strain interactions, one needs to **map** a **trait** of **interest**, such as absolute/relative abundance of each strain or ability to cause disease, to **two genomes simultaneously**. We have published a **method**, **ATOMM**, for doing precisely that when considering a pathogen and its host[53]. ATOMM uses a **two-way mixed-effect model** to test for genetic associations and cross-species genetic interactions while **accounting** for **sample structure** including interactions between the genetic backgrounds of the two organisms. It furthermore has the capacity to consider both **SNPs** and **presence/absence polymorphisms**, a feature that is necessary for effective mapping within microbial species[118–121]. By pairing 130 whole-genome sequenced *A. thaliana* strains with only 22 whole-genome sequenced *Xanthomonas* strains, ATOMM allowed us to fine map very small *Xanthomonas* genomic regions (<50 bp) involved in cross-species interactions[53]. The method is also very powerful in finding rare genetic variants. Our Chicago co-investigator Mary Sara McPeek together with a PATHOCOM-supported postdoc will adapt ATOMM for mapping microbe-microbe interactions in which interacting species both have a **large dispensable genome** and, in case of within-species interactions, **population structure** is shared.

Horizontal gene transfer (HGT) can promote cooperation by increasing genetic relatedness at loci on **mobile genetic elements (MGEs)**[122], and many genes involved in social interactions between bacteria are indeed carried by MGEs, in particular plasmids[123–125]. As mentioned, we will have ***de novo*** **assemblies** of **genomes** of our 600 bacterial strains, including complete MGE information. Preliminary analyses of a subset of our strains indicate that the frequency of plasmid-bearing strains spans a wide range across our four focal bacterial species (from ~5% for *X. campestris* to 100% for *P. agglomerans*). We expect HGT to be only a **minor factor** in our system, given the short time course of our experiments (36 hours). Nonetheless, we will estimate the **upper bound** of **HGT** in our system by analyzing at least 100 instances where replicate infections differ the most in their endpoints (an expectation is that the stochasticity of HGT events will lead to higher variance). For each of these, we will use proximity ligation based sequencing methods[126,127] to empirically determine the extent of HGT. If warranted, we will incorporate such information in our models in Aim 3.

To **confirm** the causal role of **candidate genes** identified by joint GWA mapping, we will functionally validate up to 50 candidates using classical molecular genetics (e.g., marker-exchange deletion, complementation by chromosomal insertion etc.). Candidates will be chosen according to criteria such as type of interactions (with a preference for positive interactions), intra- vs. interspecific interactions (with a preference for pathogen-pathogen interactions), and percentage of variance explained by GWA hits. These efforts will directly allow for estimating false positives in the GWA analyses.

Anticipated knowledge gained: (i) Relative **frequency** of different **types** of microbe-microbe **interactions**, in both pathogenic and non-pathogenic strains; **(ii) effects** of various **categorical variables** (e.g., regions, bacterial species, presence of MGEs), in particular whether cooperation/facilitation is more common within/between species or geographical regions[128]; **(iii) relationship** between **virulence** and **competitive** ability, which is critical for epidemiology, yet poorly understood; **(iv)** importance of **higher-order interactions**; **(v) spectrum of genes** involved in bacterial microbe-microbe interactions; and **(vi)** roles of **major genetic**

mechanisms hypothesized to underlie positive interactions, such as kin recognition, Greenbeard effects and compatibility genes[129].

Contingency planning: One risk is that the direct integration of barcodes into plant genomes might turn out to be too inefficient. In this case, we will adopt the GESTALT method for generating barcodes in a special transgene[130]. Regarding GWA mapping, the effect of population structure is always highly dependent on the trait considered[38]. If population structure leads to an inflation of false negatives[39,131], we will reduce population structure by performing GWA mapping separately within each region (France, Germany, USA). The genetic architecture underlying natural variation of microbe-microbe interactions can be unpredictable as well. If allelic effects of candidate genes are marginal[132], we will create lines that contain multiple mutations/overexpress multiple genes.

Aim 3: Building a model of persistent communities from empirical microbe-microbe interactions

Rationale: *The glue that connects our diverse datasets will be a* ***model*** *that is* ***developed*** *and* ***refined*** *in an* ***iterative process*** *with* ***data collection****. By tailoring our model structure to encompass two-way, three-way, and increasingly more complex interactions between strains, we will refine it until it captures the key features of microbial interactions* in planta*, as measured with increasingly complex synthetic communities.*

Parameterizing ecological interactions from experimental data: The modeling is based on the recent work by our co-Investigator Allesina[71], who devised a simple statistical method to predict coexistence and expected abundance in a pool of known species, using data produced by a limited number of experiments (see *Section a*). Once the model is fit, and its quality assessed by performing out-of-fit predictions, it can be used to **predict** the **coexistence** and **abundance** of **individual strains** in **any community** that can be formed by combining any number of strains from our 600 strains studied in Aim 2. The model can be made more complex by incorporating higher-order interaction (HOI)[133] terms (the model described in *Section a* maps into a linear regression; including HOIs amounts to performing polynomial regressions):

$$x_i^{(S)} = a_i + \sum_{j \in S; j \neq i} b_{ij} x_j^{(S)} + \sum_{j,k \in S; j \neq k \neq i} c_{ijk} x_j^{(S)} x_k^{(S)}$$

Despite the fact that the model now contains a larger number of parameters, fitting only requires marginally more data. This means that one can build **progressively more complex models**, until out-of-fit predictions match experimental data quite closely. Whenever the model containing HOIs fits data better than the simpler model, this indicates the presence of **interaction modification**: the presence of strain *k* modifies the interaction between strains *i* and *j* (e.g., by modifying the environment, or because of cross-feeding).

Building the model through an iterative process with data collection: As specified in Aim 2, we will measure **abundances** of all strains in **isolation**, in **pairs** and in more **complex** communities. Whenever two strains grow on their own and also coexist, we can fit the model, and by collating all two-strain models together, we can build a **large matrix** of all **pairwise relationships**. Note that (under ideal conditions) this matrix contains information on all **possible communities** that can be formed from the entire set of strains, not only those communities that we have directly measured. However, because the model is built with information on monocultures and pairs, it necessarily neglects scenarios in which a third species influences the interaction between the other two. We can therefore exploit the **150- and 200-member communities** measured in Aim 2 to build a **"top-down" model** in which only information on these larger communities is used to parameterize interactions. We can then contrast the parameters found in the bottom-up vs. top-down model, thereby highlighting discrepancies that would suggest the **presence of higher-order interactions**. We will use this comparison to guide subsequent experiments that will tackle communities of various complexity, and we will combine the top-down and bottom-up models to produce an even more accurate measure of interactions. The combined model will predict **microbial triplets, quadruplets, etc.** that are **most likely** to (i) contain **non-pairwise interactions** (via comparison of the bottom-up and top-down models), and (ii) **coexist robustly** (via the combined model). We expect the two models to be sufficient to describe the data; to this end, we are performing a Taylor expansion for surface embedding of all the points describing abundances of the isolates in all possible communities. Having identified potential higher-order interactions, we will measure the relevant communities, assess the goodness of the predictions, and refine our estimates by incorporating the new data. In this way, we can set up an **iterative experimental design** that chooses the next experiment to perform in a way that maximizes the improvement of the quality of fit, while at the same time minimizing the number of experiments required for **improvement** of the **model**.

Comparison of the model output with **patterns** of **association** at the **species**, **isolate**, and even **gene level**, as described in Aim 5, will reveal the extent to which dynamics of the pathobiota can be explained based on a characterization of strain/species interactions. Other factors, such as the abiotic environment, plant genetic

variation, and the presence of other members of the microbiota, will be considered in Aim 4, and additionally incorporated in Aim 5.

Anticipated knowledge gained: This analysis will **(i)** define the **importance** of **higher-order interactions** in our microbial communities; characterization of such interactions is an open problem in ecology, and our analysis has therefore the potential to **(ii)** impact the discipline generally by providing **high-quality**, **replicable** and **documented cases of higher-order interactions** that can be dissected and studied in laboratory conditions.

Contingency planning: Few risks are foreseen for the implementation of this Aim, as the basic framework for the model that we want to construct is already in place. The major risk is that we lag with the generation of the experimental data in Aim 2. However, model construction can begin even before all the data are in hand.

Aim 4: Experimental characterization of (a)biotic factors modulating pathogen-pathogen interactions

Rationale: Environmental factors *influence disease development and transmission dynamics*[134,135]*. Aerial (e.g. temperature, precipitation) and soil parameters (e.g. pH, nutrient content) affect the* ***physiology*** *of the* ***plant****, which may alter pathogen resistance*[135]*. Similarly,* ***plant genetics*** *can greatly impact growth of a single pathogenic strain*[74]*, and genetic diversity is known to exist among bacterial biocontrol agents*[136]*. What remains unknown is how these* ***abiotic*** *and* ***biotic factors*** *influence* ***pathogen-pathogen interactions****. In this Aim, we will test the effects of abiotic conditions,* Sphingomonas *genetics, and* A. thaliana *genetics on a* ***core matrix of pathogen-pathogen interactions*** *that will be representative of the full matrix of pairwise interactions among the 450 pathogenic strains. We predict that a* ***majority*** *of* ***interactions*** *will be* ***affected*** *by* ***abiotic conditions*** *due to qualitative and quantitative modification of host-produced resources. On the other hand, the genetics of* Sphingomonas *and* A. thaliana *should have more restricted effects, due to both* ***host-strain***[38,53,137] *and* ***strain-strain*** *specificity*[107]*. We will test these hypotheses. In all experiments, bacterial growth will be quantified with our* ***high-throughput barcode method*** *from Aim 2.*

Establishing a core matrix of pathogen-pathogen interactions: We will sample *in silico* one million combinations (i.e., submatrices) of **18 pathogenic strains** (six each from *Pantoea*, *Xanthomonas* and *Pseudomonas*, including two per geographic region). From our empirical data, we will know the frequencies of different interactions (competition, reciprocal help etc.) as well as the distribution of competitive scores among the 153 pairs of 18 strains in each of these randomly chosen combinations. We will identify the combination of 18 pathogenic strains that best reflects the **distribution** of interactions observed in the **full matrix** of pairwise interactions among the 450 pathogenic strains. If several submatrices have equally good fits, we will choose the submatrix with the highest variance of competitive scores.

Robustness of pathogen-pathogen interactions to abiotic variation: Two sets of experiments will be carried out. First, we will **alter** the **growth conditions** of the gnotobiotic *A. thaliana* plants for infections. To this end, we will employ five growth conditions that best reflect **major axes** of **variation** in **climate/weather** and **soil parameters** affecting pathobiota variation in wild *A. thaliana* populations in year 1 (Aim 1). Plants will be infected with each of the 153 microbial pairs (10 replicates) or single strains (20 replicates) in five conditions, for a total of 9,450 plants.

Second, we will grow our standard *A. thaliana* HPG1 strain (which is from the USA) **outdoors** at the Bergelson **field station** in southwest Michigan, where field experiments with genetically modified organisms can be carried out. To take seasonal and year-to-year variation in climate into account, we will use **two sowing dates** that match the main germination cohorts in southwest Michigan (mid-October and mid-March) over two years. Local **weather** (including temperature and humidity) will be recorded with data loggers. We will add different amounts of the three major nutrients to the native soil, to **manipulate nutrient status** along the axes identified as important in Aim 1. Each nutrient level will include 153 pairwise infections (10 replicates), plus single-strain infections (20 replicates), for a total of 5,670 plants per 'year x season' combination. Seeds will be **directly sown** in the field and seedlings at the five-to-six leaf stage will be sprayed with bacterial suspensions. In addition, we will monitor the **background microbiota** in 10 pools of 10 **control plants before infection** for each of the 12 'year x season x nutrient status' combinations using 16S rDNA and ITS1 profiling. These experiments will reveal which abiotic factors have the greatest impact on pathogen interactions in our system and they will provide additional data for our modeling efforts in Aim 5.

Robustness of pathogen-pathogen interactions to the genetics of *Sphingomonas*: We have shown that some *Sphingomonas* strains can **modify pathogen-pathogen interactions** (see *Section a*). To test how genetic diversity within *Sphingomonas* affects pathogen-pathogen interactions, we will grow gnotobiotic barcoded *A. thaliana* plants, and individually test the response of each of the 153 pathogen pairs from above to our 150 *Sphingomonas* strains (3 replicates, 68,850 plants total; individual interactions between *Sphingomonas* and each of the 18 pathogens will be known from Aim 2).

For each of the 153 pathogen pairs, GWA mapping in *Sphingomonas* will be carried out using a modified version of EMMAX[138] that, like ATOMM[53], allows for both SNP and insertion-deletion polymorphisms. The GWA hits will start to establish a **genomic landscape** in *Sphingomonas* associated with **community-wide interactions** in the context of **pathogenicity**. Up to 50 candidate genes will be functionally validated using classical molecular genetics (e.g. marker-exchange deletion, complementation etc.). The candidate genes will be chosen according to several criteria: (i) candidates that shift interactions from cooperation/facilitation to competition; (ii) the number of pathogen pairs affected (i.e., level of environmental pleiotropy), with a preference for generalist candidates; and (iii) fraction of variance explained by GWA hits. These efforts will directly allow for estimating false positives in the GWA analyses.

Robustness of pathogen-pathogen interactions to host genetics: It is well established that the host immune system can greatly affect pathogen proliferation, but we know little about the effects a given host has on interactions among pathogens. To address this question, we will first perform gnotobiotic infections of eight *A. thaliana* strains from each of our three geographic regions. Plants will be inoculated with 153 pathogen pairs and 18 single pathogens (5 replicates, 20,520 plants total). Based on the results, we will select a submatrix of 4x4 pathogens that best capture the range of responses with the 153 (18x18) pairs across these 24 *A. thaliana* strains. We will perform gnotobiotic infections of **300 whole-genome sequenced *A. thaliana* strains** (from near Toulouse) inoculated with the **6 pathogen pairs** and **4 single pathogens** (5 replicates, 15,000 plants total). Such a set of *A. thaliana* strains has ample power to fine-map genomic regions associated with pathobiota descriptors (Fig. 4). Because the genetic architecture of pathogen resistance may be different in a more **ecologically realistic environment**[46,81], we will repeat the experiment in **semi-natural conditions** by growing 30,000 plants over two years at our field station in Michigan. Seeds will be directly sown in the field and seedlings at the five-to-six leaf stage sprayed with bacterial suspensions. As in the growth chamber experiments, we will image plants to quantify the fitness impact of co-infection on growth and disease symptoms. **GWA fine-mapping** of **genomic regions** associated with disease descriptors will be run using a Bayesian hierarchical model that explicitly accounts for the scaled covariance matrix of population allele frequencies, which makes the analyses robust to complex demographic histories and allows permutation of phenotypes among *A. thaliana* strains[139]. From both the gnotobiotic and field GWA experiments, up to ten **candidate genes** will be functionally **validated** with CRISPR/Cas9 mutants in the appropriate backgrounds. Finally, we will analyze a subset of 768 host x pathogen combinations samples by shallow RNA-seq to assess how well differences in **disease symptoms** and **pathogen proliferation** align with altered PAMP- and effector-triggered **immunity**[99].

Anticipated knowledge gained: (i) A **rigorous test** of the prediction that **nutrients** should broadly impact bacterial interactions, whereas **host** and ***Sphingomonas* genetics** should have narrower effects; **(ii)** identification of **genes** underlying *Sphingomonas* effects on **pathogen-pathogen interactions**, thereby providing a glimpse of the types of genes and pathways that can have such effects; **(iii)** identification of **genes** underlying **host effects** on **pathogen-pathogen interactions**, thereby providing a glimpse of the types of genes and pathways that can have such effects in particular in **ecologically realistic conditions**, enabled by our field station in Michigan.

Contingency planning: If first results indicate that pathogen-pathogen interactions are not influenced by one of the three tested factors, we will concentrate our efforts on the other two factors. Regarding field experiments, we are very experienced in the management of thousands of plants[39,42,77,140]. If a field experiment nevertheless fails, we will conduct additional experiments in growth chambers with more pairs of pathogenic strains. For GWA mapping in Sphingomonas*, if the rate of false negatives appears too high when considering the 150 strains, we will perform GWA separately with strains from each geographic region. For* A. thaliana*, the effect of population structure is small in our set of 300 local strains[88], but the genetic architecture underlying trait variation cannot be known beforehand. If allelic effects of candidate genes are marginal, we will create lines with multiple mutations/overexpress multiple genes.*

Aim 5: Applying the model in an ecological genomics framework

Rationale: *Our ultimate goal is to compare **real-world patterns** in the pathobiota to **predictions** generated from our models, in order to decipher the drivers of community structure, based on the following logic: given a characterization of (microbe)^n interactions, there is a method[71] to determine which combinations of microbes form communities that can persist. We will employ this technique in a novel way by defining the **universe** of **persistent communities**, given a defined characterization of microbial interactions and a strong assumption that these interactions determine community structure. Patterns emerging from this set of persistent communities will define a **null distribution** of **expected patterns**, provided that microbial interactions are sufficient to explain community structure. We will compare our survey data from Aim 1 to these null distributions.*

Confirming key abiotic and biotic drivers of pathobiota composition: We will begin by utilizing the models developed in Aim 3 that are based only on the laboratory data generated in Aim 2. Of course, we do not expect that these models will work off the shelf to explain attributes of real-world pathobiota such as their richness, isolate relatedness etc. However, we will investigate **model fit** across our hierarchical field collections to identify **biotic** and **abiotic factors** that impact microbe-microbe interactions. We will have found candidate biotic and abiotic drivers in Aim 4, and we will seek confirmation of their importance. These explorations can be completed at a **variety** of **scales** with respect to the organization of the microbial communities. For example, it would be unrealistic to imagine that we can predict the abundance of all isolates in our regional sets from Aim 2 (even if we can find exact matches in our real-world data). We will instead investigate whether **some isolates** are **ecologically equivalent** by comparing the strength and direction of their interactions with all other strains, and collapse isolates that behave similarly. This will reduce the challenge, and should provide greater chances of successful predictions. For similar reasons, we may find a **better match** between our **model predictions** and **survey data**, if we make those comparisons at a **genetic** rather **than taxonomic level**. We posit that genetic comparisons may be more successful because of the extensive overlap in gene content across isolates.

Extending the model: Once we identify important abiotic and/or biotic drivers, we can extend the model to make **microbial interactions** a **function** of these **drivers**. Data from Aim 4, for example, will allow us to estimate simple functions for nutrients, as well as other factors that we have tested. It is important to note that abiotic and biotic factors that impact all microbe-microbe interactions in the same way will not modify community structure, and therefore will not pose a problem for our approach (for the effect of temperature, see Fig. 4 of ref. 71). With nine postdoc years devoted to model development under the supervision of co-investigator Allesina, we are allowing sufficient room for **model improvement**, but until we can explore the data, it is premature to anticipate which avenues will be most fruitful. Some possibilities include the **roles** of (i) **spatial structure**, especially if structure is revealed in our real-world data (Aim 1), and (ii) **transmission dynamics**, where we can draw upon characterization of the epiphytic communities and the microbiome of neighboring plants in Aim 1 to define a pool of potential colonists. If **HGT** of **cooperation** genes is detected in our experiments in Aim 2, we will extend our model to consider **eco-evolutionary dynamics**. In addition, ongoing work by Allesina is extending the simple models from ref. 71 to include resource competition and higher-order interactions, as well as relaxing assumptions to allow the use of relative abundances and underdetermined (sparse) interactions. We will, of course, consider these newer models.

Comparing model output and real-world data: For each parameterized and well-formulated model that we develop, we will compare its predictions to patterns that we have measured in the real world (Aim 1). Patterns can be investigated at the level of isolates (with PEN-seq data), genes (with PEN-seq and metagenome data), species or communities (from amplicon data). For example, we have previously observed a higher **richness** in the **pathobiota** when the microbial community contains a greater **fraction** of **pathogenic taxa**[25]. By generating the **universe** of **persistent communities** under a particular model, and generating the relationship between the **fraction** of **pathogens** and the **diversity** of the **pathobiota**, we can test the ability of our model to capture this **qualitative pattern** as revealed in the data. Similarly, if we identify genes in the microbes that promote **cooperation**, or genes in the plants that promote cooperation, we will ask how the **prevalence** of those genes relates to **pathobiota diversity** both in the data and in the model predictions. And if we continue to see that *Sphingomonas* lessens the pressure on competitively inferior isolates (see *Section a*), we can examine the relationship between ***Sphingomonas* abundance** and **pathobiota diversity** in both model and data. There are many additional questions we can ask, and the ones enumerated here only serve as examples, to illustrate our general approach.

Anticipated knowledge gained: This last aim is open-ended, but we expect to **build** a **(i) general understanding** of **ecological** and **genetic** drivers of microbial community structure; **(ii)** their **relative importance**, and **(iii)** the **utility** of an **ecological versus genetic characterization** of microbial communities.

Contingency planning: The greatest risk is that the laboratory data to parameterize our models do not allow accurate prediction at one or several scales (broad-scale patterns vs. effects of specific environmental parameters, genetic variants etc.). An alternative approach that we can adopt involves fitting our models using a subset of field data and then testing our ability to predict patterns in the remaining field data. Note that this does not obviate Aims 1-4: all knowledge gained in the previous Aims will still be valid, as are the identification and characterization of ecologically important genes that will allow us to ask genetically driven questions in Aim 5.

Ensuring team integration

The three Principal Investigators and their co-investigators already have a **proven record of collaboration**. To continue to ensure seamless integration of team members at the three sites, we will have (i) **monthly "all hands" meetings** by video conferencing at 9 am Chicago/4 pm European time, (ii) **weekly project meetings** by video conferencing, (iii) **joint field sampling efforts** (of note, Bergelson has carried out joint field experiments in Sweden for several years now with collaborators from Sweden, Austria and the UK), (iv) **mandatory secondments** of postdocs and PhD students, and (v) **annual multi-day retreats** rotating among the three sites. We will use two major tools for project management, information exchange and internal reporting. We will use **Slack** for rapid, informal communication between team members, and Atlassian's **Confluence** platform for long-term coordination. Confluence provides a powerful wiki-like platform for team collaboration, and it allows for secure sharing of data and information, reporting as well as assignment and tracking of experiments.

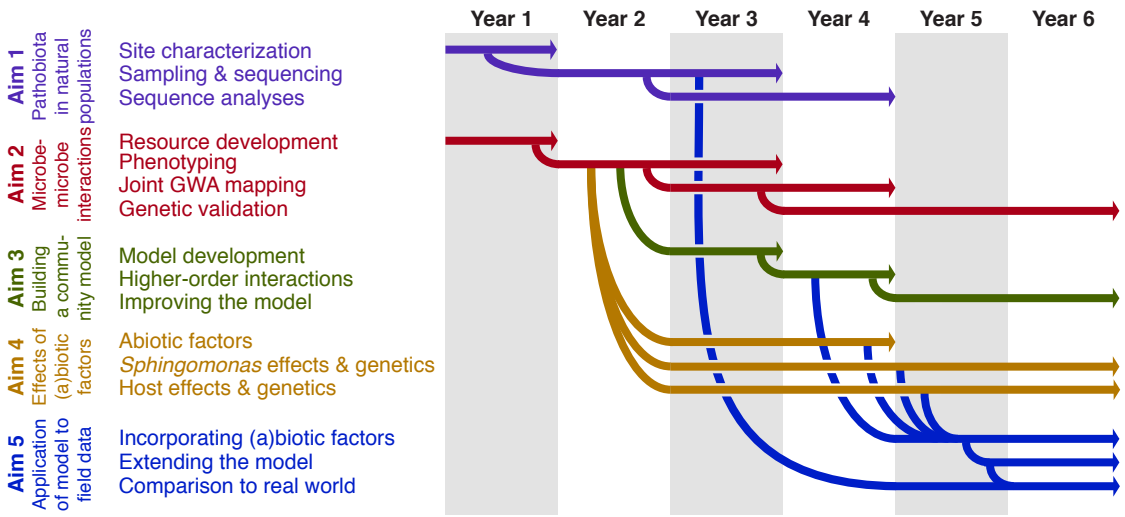
Risk statement and outlook

How does PATHOCOM meet the high-risk/high-gain profile expected of ERC projects? PATHOCOM goes **beyond** the **current state of the art** by taking **full advantage** of very rapid developments in the **genomics** of **wild communities, ultra-high-throughput methods** for studying **microbe-microbe interactions** ***in planta*** and our ability to **model very large data sets**. With the exception of human gut microbiome research, we are not aware of other programs with similar ambitions to understand **drivers** of **microbe-microbe interactions** in the **real world**, at the **level** of **community, environment** and **genetics**. While lab studies have been very successful in informing us about general principles and mechanisms of pathogen recognition by the host, and evasion of recognition by pathogens, we are still largely in the dark when it comes to understanding how microbes interact in the **context** of **natural infections** to **overcome plant defense**s. We will redress this situation in PATHOCOM.

Our aims span a **range** of **approaches** with **increasing risk**. Aim 1 will deliver **rich knowledge** about **microbe diversity** in an **exceptionally large sample** of **wild plants**, across a sufficient number of seasons and geographic regions that general patterns can be distinguished from local idiosyncrasies. Importantly, because we will capture genome-wide genetic variation, this Aim will set the stage for **subsequent mechanistic studies**. Aims 2 and 4 will produce an orthogonal data set of **pathogen-pathogen interactions** at a **similarly ambitious scale**, backed up by complete genome information. Note that the **data sets** that are **generated** in **Aims 1, 2** and **4** can be mined by others for **discovery** of **genetic** and **ecological mechanisms** with alternative analytical approaches. Aims 3 and 5 are the aims with the most uncertain outcomes. In these Aims, we will **synthesize** the **observational** and **empirical sides** of the **project**, by developing cutting-edge models to explain patterns of microbe-microbe interactions observed in simplified lab settings, in controlled field experiments, and directly in data from wild plants. These models will predict **broad-scale patterns** from **small-scale** characterization of species **interactions** and thus build a general understanding of how microbial community structure is shaped, with the less certain part being how far they will go in predicting the effects of individual isolates, genes or environmental factors.

If successful, a systematic understanding of forces that **shape** the **success** of **pathogenic microbes** in **wild plants** will have important implications not only for anticipating disease development in agroecosystems, but also for the design of **new intervention strategies** based on interfering with synergistic positive interactions among co-infecting pathogens.

Timeline



Section c. Resources follows Bibliography on page 21

Bibliography

1. Alexander, H. M. Disease in Natural Plant Populations, Communities, and Ecosystems: Insights into Ecological and Evolutionary Processes. *Plant Dis.* **94**, 492–503 (2010).
2. Stukenbrock, E. H. & McDonald, B. A. The origins of plant pathogens in agro-ecosystems. *Annu. Rev. Phytopathol.* **46**, 75–100 (2008).
3. McDonald, B. A. & Stukenbrock, E. H. Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, (2016).
4. Kemen, E. Microbe-microbe interactions determine oomycete and fungal host colonization. *Curr. Opin. Plant Biol.* **20**, 75–81 (2014).
5. Agler, M. T. *et al.* Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLoS Biol.* **14**, e1002352 (2016).
6. Griffiths, E. C., Pedersen, A. B., Fenton, A. & Petchey, O. L. Analysis of a summary network of co-infection in humans reveals that parasites interact most via shared resources. *Proc. Biol. Sci.* **281**, 20132286 (2014).
7. Tollenaere, C., Susi, H. & Laine, A.-L. Evolutionary and Epidemiological Implications of Multiple Infection in Plants. *Trends Plant Sci.* **21**, 80–90 (2016).
8. Belhaj, K. *et al.* Arabidopsis late blight: infection of a nonhost plant by Albugo laibachii enables full colonization by Phytophthora infestans. *Cell. Microbiol.* **19**, (2017).
9. Abdullah, A. S. *et al.* Host-Multi-Pathogen Warfare: Pathogen Interactions in Co-infected Plants. *Front. Plant Sci.* **8**, 1806 (2017).
10. Gorsich, E. E. *et al.* Opposite outcomes of coinfection at individual and population scales. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7545–7550 (2018).
11. Jouet, A. *et al.* Albugo candida race diversity, ploidy and host-associated microbes revealed using DNA sequence capture on diseased plants in the field. *New Phytol.* **221**, 1529–1543 (2019).
12. Sibley, C. D. & Surette, M. G. The polymicrobial nature of airway infections in cystic fibrosis: Cangene Gold Medal Lecture. *Can. J. Microbiol.* **57**, 69–77 (2011).
13. McGuigan, L. & Callaghan, M. The evolving dynamics of the microbial community in the cystic fibrosis lung. *Environ. Microbiol.* **17**, 16–28 (2015).
14. Griffiths, E. C., Pedersen, A. B., Fenton, A. & Petchey, O. L. The nature and consequences of coinfection in humans. *J. Infect.* **63**, 200–206 (2011).
15. Grünwald, N. J., Garbelotto, M., Goss, E. M., Heungens, K. & Prospero, S. Emergence of the sudden oak death pathogen Phytophthora ramorum. *Trends Microbiol.* **20**, 131–138 (2012).
16. Gladieux, P. *et al.* The population biology of fungal invasions. *Mol. Ecol.* **24**, 1969–1986 (2015).
17. McMullan, M. *et al.* The ash dieback invasion of Europe was founded by two genetically divergent individuals. *Nat Ecol Evol* **2**, 1000–1008 (2018).
18. Bevan, J. R., Crute, I. R. & Clarke, D. D. Diversity and variation in expression of resistance to Erysiphe fischeri in Senecio vulgaris. *Plant Pathol.* **42**, 647–653 (1993).
19. Laine, A.-L. Resistance variation within and among host populations in a plant-pathogen metapopulation: implications for regional pathogen dynamics. *J. Ecol.* **92**, 990–1000 (2004).
20. Burdon, J. J. Phenotypic and genetic patterns of resistance to the pathogen Phakopsora pachyrhizi in populations of Glycine canescens. *Oecologia* **73**, 257–267 (1987).
21. Meyer, S. e. *et al.* Genetic Variation in Ustilago bullata: Molecular Genetic Markers and Virulence on Bromus tectorum Host Lines. *Int. J. Plant Sci.* **166**, 105–115 (2005).
22. Barrett, L. G., Thrall, P. H. & Burdon, J. J. Evolutionary diversification through hybridization in a wild host-pathogen interaction. *Evolution* **61**, 1613–1621 (2007).
23. Susi, H., Barrès, B., Vale, P. F. & Laine, A.-L. Co-infection alters population dynamics of infectious disease. *Nat. Commun.* **6**, 5975 (2015).
24. Karasov, T. L. *et al.* Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* **24**, 168–179 (2018).
25. Bartoli, C. *et al.* In situ relationships between microbiota and potential pathobiota in Arabidopsis thaliana. *ISME J.* **12**, 2024–2038 (2018).
26. Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006).
27. Lamichhane, J. R. & Venturi, V. Synergisms between microbial pathogens in plant disease complexes: a growing trend. *Front. Plant Sci.* **6**, 385 (2015).
28. Barrett, L. G., Kniskern, J. M., Bodenhausen, N., Zhang, W. & Bergelson, J. Continua of specificity and

virulence in plant host-pathogen interactions: causes and consequences. *New Phytol.* **183**, 513–529 (2009).
29. Vorholt, J. A., Vogel, C., Carlström, C. I. & Müller, D. B. Establishing Causality: Opportunities of Synthetic Communities for Plant Microbiome Research. *Cell Host Microbe* **22**, 142–155 (2017).
30. Paredes, S. H. *et al.* Design of synthetic bacterial communities for predictable plant phenotypes. *PLoS Biol.* **16**, e2003962 (2018).
31. Nordborg, M. *et al.* The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biol.* **3**, e196 (2005).
32. Horton, M. *et al.* Genome-wide pattern of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat. Genet.* **44**, 212–216 (2012).
33. 1001 Genomes Consortium. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**, 481–491 (2016).
34. Fulgione, A. & Hancock, A. M. Archaic lineages broaden our view on the history of Arabidopsis thaliana. *New Phytol.* **219**, 1194–1198 (2019).
35. Exposito-Alonso, M. *et al.* The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.* **14**, e1007155 (2018).
36. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* **317**, 338–342 (2007).
37. Aranzana, M. J. *et al.* Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet.* **1**, e60 (2005).
38. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–631 (2010).
39. Brachi, B. *et al.* Linkage and association mapping of Arabidopsis thaliana flowering time in nature. *PLoS Genet.* **6**, e1000940 (2010).
40. Todesco, M. *et al.* Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* **465**, 632–636 (2010).
41. Huard-Chauveau, C. *et al.* An atypical kinase under balancing selection confers broad-spectrum disease resistance in Arabidopsis. *PLoS Genet.* **9**, e1003766 (2013).
42. Karasov, T. L. *et al.* The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature* **512**, 436–440 (2014).
43. Horton, M. W. *et al.* Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).
44. Brachi, B. *et al.* Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4032–4037 (2015).
45. Frachon, L. *et al.* Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nat Ecol Evol* **1**, 1551–1561 (2017).
46. Rubio, B. *et al.* Genome-wide association study reveals new loci involved in Arabidopsis thaliana and Turnip mosaic virus (TuMV) interactions in the field. *New Phytol.* **221**, 2026–2038 (2019).
47. Voichek, Y. & Weigel, D. Finding genetic variants in plants without complete genomes. *bioRxiv* (2019).
48. Jakob, K. *et al.* Pseudomonas viridiflava and P. syringae--natural pathogens of Arabidopsis thaliana. *Mol. Plant. Microbe. Interact.* **15**, 1195–1203 (2002).
49. Goss, E. M. & Bergelson, J. Fitness consequences of infection of Arabidopsis thaliana with its natural bacterial pathogen Pseudomonas viridiflava. *Oecologia* **152**, 71–81 (2007).
50. Jakob, K., Kniskern, J. M. & Bergelson, J. The role of pectate lyase and the jasmonic acid defense response in Pseudomonas viridiflava virulence. *Mol. Plant. Microbe. Interact.* **20**, 146–158 (2007).
51. Bodenhausen, N., Horton, M. W. & Bergelson, J. Bacterial communities associated with the leaves and the roots of Arabidopsis thaliana. *PLoS One* **8**, e56329 (2013).
52. Karasov, T. L., Barrett, L., Hershberg, R. & Bergelson, J. Similar levels of gene content variation observed for Pseudomonas syringae populations extracted from single and multiple host species. *PLoS One* **12**, e0184195 (2017).
53. Wang, M. *et al.* Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E5440–E5449 (2018).
54. Durán, P. *et al.* Microbial Interkingdom Interactions in Roots Promote Arabidopsis Survival. *Cell* **175**, 973–983.e14 (2018).
55. Regalado, J. *et al.* Combining whole genome shotgun sequencing and rDNA amplicon analyses to improve detection of microbe-microbe interaction networks in plant leaves. *bioRxiv* 823492 (2019). doi:10.1101/823492
56. Karasov, T. L. *et al.* The relationship between microbial biomass and disease in the Arabidopsis thaliana phyllosphere. *bioRxiv* 828814 (2019). doi:10.1101/828814
57. Müller, D. B., Vogel, C., Bai, Y. & Vorholt, J. A. The Plant Microbiota: Systems-Level Insights and

Perspectives. *Annual Review of Genetics* **50**, 211–234 (2016).
58. Innerebner, G., Knief, C. & Vorholt, J. A. Protection of Arabidopsis thaliana against leaf-pathogenic Pseudomonas syringae by Sphingomonas strains in a controlled model system. *Appl. Environ. Microbiol.* **77**, 3202–3210 (2011).
59. Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E. & Schulze-Lefert, P. Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* **64**, 807–838 (2013).
60. Chowdhury, S. P., Hartmann, A., Gao, X. & Borriss, R. Biocontrol mechanism by root-associated Bacillus amyloliquefaciens FZB42 - a review. *Front. Microbiol.* **6**, 780 (2015).
61. Gómez Expósito, R., de Bruijn, I., Postma, J. & Raaijmakers, J. M. Current Insights into the Role of Rhizosphere Bacteria in Disease Suppressive Soils. *Front. Microbiol.* **8**, 2529 (2017).
62. Levy, A., Conway, J. M., Dangl, J. L. & Woyke, T. Elucidating Bacterial Gene Functions in the Plant Microbiome. *Cell Host Microbe* **24**, 475–485 (2018).
63. Bass, D., Stentiford, G. D., Wang, H.-C., Koskella, B. & Tyler, C. R. The Pathobiome in Animal and Plant Diseases. *Trends Ecol. Evol.* (2019). doi:10.1016/j.tree.2019.07.012
64. Cordovez, V., Dini-Andreote, F., Carrión, V. J. & Raaijmakers, J. M. Ecology and Evolution of Plant Microbiomes. *Annu. Rev. Microbiol.* (2019). doi:10.1146/annurev-micro-090817-062524
65. Moura, M. L., Brito, L. M., Mourao, I. M., Jacques, M. A. & Duclos, J. Tomato pith necrosis (TPN) caused by P-corrugata and P-mediterranea: Severity of damages and crop loss assessment. in *Proceedings of the 1st International Symposium on Tomato Diseases* (eds. Momol, M. T., Ji, P. & Jones, J. B.) 365–371 (International Society Horticultural Science, 2005).
66. Grube, M., Fürnkranz, M., Zitzenbacher, S., Huss, H. & Berg, G. Emerging multi-pathogen disease caused by Didymella bryoniae and pathogenic bacteria on Styrian oil pumpkin. *Eur. J. Plant Pathol.* **131**, 539 (2011).
67. Ellner, S. P., Seifu, Y. & Smith, R. H. Fitting population dynamic models to time-series data by gradient matching. *Ecology* **83**, 2256–2270 (2002).
68. Ionides, E. L., Bretó, C. & King, A. A. Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 18438–18443 (2006).
69. Bretó, C., He, D., Ionides, E. L. & King, A. A. Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3**, 319–348 (2009).
70. Xiao, Y. *et al.* Mapping the ecological networks of microbial communities. *Nat. Commun.* **8**, 2042 (2017).
71. Maynard, D. S., Miller, Z. R. & Allesina, S. Predicting coexistence in experimental ecological communities. *Nat. Ecol. Evol.* accepted in principle (bioRxiv 598326) (2019).
72. Abreu, C. I., Friedman, J., Andersen Woltz, V. L. & Gore, J. Mortality causes universal changes in microbial community composition. *Nat. Commun.* **10**, 2120 (2019).
73. Thiergart, T. *et al.* Root microbiota assembly and adaptive differentiation among European Arabidopsis populations. *bioRxiv* 640623 (2019). doi:10.1101/640623
74. Roux, F. & Bergelson, J. The Genetics Underlying Natural Variation in the Biotic Interactions of Arabidopsis thaliana: The Challenges of Linking Evolutionary Genetics and Community Ecology. *Curr. Top. Dev. Biol.* **119**, 111–156 (2016).
75. Bulgarelli, D. *et al.* Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
76. Lundberg, D. S. *et al.* Defining the core Arabidopsis thaliana root microbiome. *Nature* **488**, 86–90 (2012).
77. Brachi, B. *et al.* Plant genes influence microbial hubs that shape beneficial leaf communities. *bioRxiv* 181198 (2017). doi:10.1101/181198
78. Bergelson, J., Mittelstrass, J. & Horton, M. W. Characterizing both bacteria and fungi improves understanding of the Arabidopsis root microbiome. *Sci. Rep.* **9**, 24 (2019).
79. Ning, Y., Liu, W. & Wang, G.-L. Balancing Immunity and Yield in Crop Plants. *Trends Plant Sci.* **22**, 1069–1079 (2017).
80. Krattinger, S. G. & Keller, B. Molecular genetics and evolution of disease resistance in cereals. *New Phytol.* **212**, 320–332 (2016).
81. Bartoli, C. & Roux, F. Genome-Wide Association Studies In Plant Pathosystems: Toward an Ecological Genomics Approach. *Front. Plant Sci.* **8**, 763 (2017).
82. Debieu, M., Huard-Chauveau, C., Genissel, A., Roux, F. & Roby, D. Quantitative disease resistance to the bacterial pathogen Xanthomonas campestris involves an Arabidopsis immune receptor pair and a gene of unknown function. *Mol. Plant Pathol.* **17**, 510–520 (2016).
83. Aoun, N. *et al.* Quantitative Disease Resistance under Elevated Temperature: Genetic Basis of New Resistance Mechanisms to Ralstonia solanacearum. *Front. Plant Sci.* **8**, 1387 (2017).
84. Vetter, M. M. *et al.* Flagellin perception varies quantitatively in Arabidopsis thaliana and its relatives. *Mol. Biol. Evol.* **29**, 1655–1667 (2012).

85. Roux, F., Noël, L., Rivas, S. & Roby, D. ZRK atypical kinases: emerging signaling components of plant immunity. *New Phytol.* **203**, 713–716 (2014).
86. Wagner, M. R. *et al.* Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nat. Commun.* **7**, 12151 (2016).
87. Walters, W. A. *et al.* Large-scale replicated field study of maize rhizosphere identifies heritable microbes. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7368–7373 (2018).
88. Frachon, L. *et al.* A Genomic Map of Climate Adaptation in Arabidopsis thaliana at a Micro-Geographic Scale. *Front. Plant Sci.* **9**, 967 (2018).
89. Palmer, M. *et al.* Phylogenomic resolution of the bacterial genus Pantoea and its relationship with Erwinia and Tatumella. *Antonie Van Leeuwenhoek* **110**, 1287–1309 (2017).
90. Bomblies, K. *et al.* Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genet.* **6**, e1000890 (2010).
91. Kniskern, J. M., Traw, M. B. & Bergelson, J. Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on Arabidopsis thaliana. *Mol. Plant. Microbe. Interact.* **20**, 1512–1522 (2007).
92. Brachi, B. *et al.* Investigation of the geographical scale of adaptive phenological variation and its underlying genetics in Arabidopsis thaliana. *Mol. Ecol.* **22**, 4222–4240 (2013).
93. Sickel, W. *et al.* Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. *BMC Ecol.* **15**, 20 (2015).
94. Cuénoud, P. *et al.* Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. *Am. J. Bot.* **89**, 132–144 (2002).
95. Barthet, M. M. & Hilu, K. W. Expression of matK: functional and evolutionary implications. *Am. J. Bot.* **94**, 1402–1412 (2007).
96. Białas, A. *et al.* Lessons in Effector and NLR Biology of Plant-Microbe Systems. *Mol. Plant. Microbe. Interact.* **31**, 34–45 (2018).
97. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
98. Van de Weyer, A.-L. *et al.* A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell* **178**, 1260–1272.e14 (2019).
99. Jones, J. D. G., Vance, R. E. & Dangl, J. L. Intracellular innate immune surveillance devices in plants and animals. *Science* **354**, (2016).
100. Thilliez, G. J. A. *et al.* Pathogen enrichment sequencing (PenSeq) enables population genomic studies in oomycetes. *New Phytol.* **221**, 1634–1648 (2019).
101. Bai, Y. *et al.* Functional overlap of the Arabidopsis leaf and root microbiota. *Nature* **528**, 364–369 (2015).
102. Boulanger, A. & Noël, L. D. Xanthomonas Whole Genome Sequencing: Phylogenetics, Host Specificity and Beyond. *Frontiers in microbiology* **7**, 1100 (2016).
103. Levy, A. *et al.* Genomic features of bacterial adaptation to plants. *Nat. Genet.* **50**, 138–150 (2018).
104. Garud, N. R., Good, B. H., Hallatschek, O. & Pollard, K. S. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biol.* **17**, e3000102 (2019).
105. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
106. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
107. Jiang, L. *et al.* A mapping framework of competition-cooperation QTLs that drive community dynamics. *Nat. Commun.* **9**, 3010 (2018).
108. Barret, M. *et al.* Emergence shapes the structure of the seed microbiota. *Appl. Environ. Microbiol.* **81**, 1257–1266 (2015).
109. Berge, O. *et al.* A user's guide to a data base of the diversity of Pseudomonas syringae and its application to classifying strains in this phylogenetic complex. *PLoS One* **9**, e105547 (2014).
110. Choi, K.-H. *et al.* A Tn7-based broad-range bacterial cloning and expression system. *Nat. Methods* **2**, 443–448 (2005).
111. Wei, H.-L., Zhang, W. & Collmer, A. Modular Study of the Type III Effector Repertoire in Pseudomonas syringae pv. tomato DC3000 Reveals a Matrix of Effector Interplay in Pathogenesis. *Cell Rep.* **23**, 1630–1638 (2018).
112. Choi, K.-H. & Schweizer, H. P. mini-Tn7 insertion in bacteria with single attTn7 sites: example Pseudomonas aeruginosa. *Nat. Protoc.* **1**, 153–161 (2006).
113. Jittawuttipoka, T. *et al.* Mini-Tn7 vectors as genetic tools for gene cloning at a single copy number in an industrially important and phytopathogenic bacteria, Xanthomonas spp. *FEMS Microbiol. Lett.* **298**, 111–117 (2009).

114. Romero-Jiménez, L., Rodríguez-Carbonell, D., Gallegos, M. T., Sanjuán, J. & Pérez-Mendoza, D. Mini-Tn7 vectors for stable expression of diguanylate cyclase PleD* in Gram-negative bacteria. *BMC Microbiol.* **15**, 190 (2015).
115. Schlechter, R. O. *et al.* Chromatic Bacteria - A Broad Host-Range Plasmid and Chromosomal Insertion Toolbox for Fluorescent Protein Expression in Bacteria. *Front. Microbiol.* **9**, 3052 (2018).
116. Levy, S. F. *et al.* Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).
117. Miki, D., Zhang, W., Zeng, W., Feng, Z. & Zhu, J.-K. CRISPR/Cas9-mediated gene targeting in Arabidopsis using sequential transformation. *Nat. Commun.* **9**, 1967 (2018).
118. Vernikos, G., Medini, D., Riley, D. R. & Tettelin, H. Ten years of pan-genome analyses. *Curr. Opin. Microbiol.* **23**, 148–154 (2015).
119. Chen, P. E. & Shapiro, B. J. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.* **25**, 17–24 (2015).
120. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
121. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
122. Nogueira, T. *et al.* Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr. Biol.* **19**, 1683–1691 (2009).
123. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic elements, and why? *Heredity* **106**, 1–10 (2011).
124. Dimitriu, T., Misevic, D., Lindner, A. B. & Taddei, F. Mobile genetic elements are involved in bacterial sociality. *Mob. Genet. Elements* **5**, 7–11 (2015).
125. Hall, J. P. J., Brockhurst, M. A. & Harrison, E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, (2017).
126. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
127. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *ISME J.* **13**, 2437–2446 (2019).
128. Nowak, M. A., Tarnita, C. E. & Wilson, E. O. The evolution of eusociality. *Nature* **466**, 1057–1062 (2010).
129. Subrahmaniam, H. J. *et al.* The genetics underlying natural variation of plant-plant interactions, a beloved but forgotten member of the family of biotic interactions. *Plant J.* **93**, 747–770 (2018).
130. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
131. Bergelson, J. & Roux, F. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat. Rev. Genet.* **11**, 867–879 (2010).
132. Weissbrod, O., Rothschild, D., Barkan, E. & Segal, E. Host genetics and microbiome associations through the lens of genome wide association studies. *Curr. Opin. Microbiol.* **44**, 9–19 (2018).
133. Grilli, J., Barabás, G., Michalska-Smith, M. J. & Allesina, S. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548**, 210–213 (2017).
134. Engering, A., Hogerwerf, L. & Slingenbergh, J. Pathogen-host-environment interplay and disease emergence. *Emerg. Microbes Infect.* **2**, e5 (2013).
135. Brader, G. *et al.* Ecology and Genomic Insights into Plant-Pathogenic and Plant-Nonpathogenic Endophytes. *Annu. Rev. Phytopathol.* **55**, 61–83 (2017).
136. Ross, I. L., Alami, Y., Harvey, P. R., Achouak, W. & Ryder, M. H. Genetic diversity and biological control activity of novel species of closely related pseudomonads isolated from wheat field soils in South Australia. *Appl. Environ. Microbiol.* **66**, 1609–1616 (2000).
137. Nemri, A. *et al.* Genome-wide survey of Arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 10302–10307 (2010).
138. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
139. Gautier, M. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics* **201**, 1555–1579 (2015).
140. Tian, D., Traw, M. B., Chen, J. Q., Kreitman, M. & Bergelson, J. Fitness costs of R-gene-mediated resistance in Arabidopsis thaliana. *Nature* **423**, 74–77 (2003).