Genomics of host specificity in the *Rhizobium*-legume symbiosis.

Imperial, J.[1, 2*], Laguerre, G.[3†], Brito, B.[1], Jorrín, B.[1]
[1] Centro de Biotecnología y Genómica de Plantas (CBGP) Universidad Politécnica de Madrid. Campus de Montegancedo, Pozuelo de Alarcón, 28223 Madrid, España. [2] Consejo Superior de Investigaciones Científicas (CSIC). Madrid, España. [3] LSTM, CIRAD-IRD, Université de Montpellier 2, Supagro, USC INRA, Montpellier. † Deceased, 17/1/2013.
[*] juan.imperial@upm.es

ABSTRACT
Most *Rhizobium leguminosarum* bv. *viciae* isolates are able to specifically nodulate plants of any of four different legume genera: *Pisum, Lens, Vicia,* and *Lathyrus.* However, previous evidence suggests that some genotypes are more adapted to a given plant host than others, and that the plant host can select specific genotypes among those present in a given soil population. We have used a population genomics approach to confirm that this is indeed the case, and to analyze the specific genotypic characteristics that each plant host selects.

One of the key aspects of the *Rhizobium*-legume symbiosis is its well-known specificity: specific rhizobia nodulate and fix nitrogen in specific legume hosts. However, this specificity is not absolute. Some tropical legumes (such as *Phaseolus* or siratro) are quite broad in their specificity requirements and are promiscuously nodulated by a large number of different rhizobial species and genera, whereas some rhizobia (such as *Sinorhizobium* sp. NGR234) are able to establish symbioses with very different plants. Some rhizobia, such as *Sinorhizobium meliloti* (Ballard *et al.*, 2005) or *Bradyrhizobium japonicum* (Koch *et al.*, 2010) can nodulate different hosts depending on a specific genetic complement, often uncovered after mutant screening or isolation of specific strains that are symbiotically active with just some of the hosts.
However, in some cases very subtle mechanisms of adaptation to a specific plant host might be in operation. This, for example, seems to be the case of rhizobial species were all (or most) of the isolates can effectively establish a diazotrophic symbiosis with plants of several different genera, often from different habitats and with different lifestyles, and is well exemplified by *Rhizobium leguminosarum* bv. viciae. Isolates belonging to this biovar establish effective symbioses with legumes belonging to four genera: *Pisum*, *Lens*, *Lathyrus* and *Vicia*. The last genus, in particular, includes species as diverse as vetch (*V. sativa*) and broad bean (*V. faba*). One set of nodulation and nitrogen fixation genes, harbored on a symbiotic plasmid, allows successful establishment and development of symbiosis with the different hosts (Surin and Downie, 1989) and, in cross-inoculation experiments, when challenged with any one of the above legume hosts, any *R. leguminosarum* bv. viciae strain is able to establish an efficient symbiosis. However, it has long been hypothesized that different rhizobial strains may be more adapted to a specific plant host than others, which may result in selection and enrichment of a specific strain or set of strains by the legume host from those present in a particular soil. Molecular evidence for plant-mediated selection of specific rhizobial genotypes from soil populations has been obtained by the research groups of Gisèle Laguerre (Depret *et al.*, 2004; Laguerre *et al.*, 2003; Louvrier *et al.*, 1995) and J. Peter W. Young (Mutch and Young, 2004; Palmer and Young, 2000). They used molecular markers and specific PCR amplification to obtain evidence that different plant hosts enrich specific genotypic marker variants of *R. leguminosarum*bv. viciae from those available in the soil.

These molecular studies were limited by the number and nature of the markers selected, and did not clarify the bases for enrichment of a given specific genotype. With the development and availability of new generation sequencing technologies this problem has been reappraised using genomic and population genomic approaches.

Genomics of Rhizobium.

After the original reports on genome sequencing of model rhizobia (*Mesorhizobium loti* 2000, *Sinorhizobium meliloti* 2001, *Bradyrhizobium japonicum* 2002, *R. etli* 2006 and *R. leguminosarum* 2006, a very large number of rhizobial genomes have been sequenced or are in the process of being sequenced, and among them about 100 *Rhizobium* isolates (as many as 96 complete or ongoing genome sequence projects in GOLD, the Genomes OnLine Database –http://www.genomesonline.org, as of June 30, 2013). Although data are still quite recent, several general conclusions on the genomic structure and organization of *Rhizobium* rhizobia emerge. In general, the *Rhizobium* contain large genomes, of *ca.* 7 Mb, and even larger in the case of members of the genus *Bradyrhizobium* (*ca.* 9 Mb).The occurrence of very large genomes in soil bacteria has been interpreted as an adaptation to this habitat, a complex, hostile and changing environment that demands the large metabolic and behavioral plasticity that can be provided by a large gene-encoding capacity. Contrary to members of the genus *Bradyrhizobium*, members of the genus *Rhizobium* (and both *Sinorhizobium*) present a multi-partite genome, harboring several large plasmids, some of which resemble chromosomes ("chromids", Harrison *et al*., 2010). On average, 30-40% of the genome in these bacteria is present in the form of plasmids (Galardini *et al*., 2013; Harrison *et al*., 2010; Mazur *et al*., 2011). This characteristic is shared with the Roseobacter clade (Petersen *et al*., 2013), and affords a large genomic plasticity, especially since many of these plasmids incorporate conjugative systems (Crossman *et al*., 2008). This plasticity liberates these bacteria from the constraints of long replication times associated to a single, very large chromosome, the situation found with the bradyrhizobia.

Genomics and the rhizobia in the soil.

Soil microbial communities have the highest level of prokaryotic diversity (up to 10^9 microorganisms per gram, Knietsch *et al*., 2003). Metagenomic approaches would appear to be the ideal approximation to such a complex system, allowing the study of the nature, composition and function of microbial communities in soil. However, even metagenomics is limitedby this complexity, in view of: a) the very large size of these datasets limits our ability to analyze them; b) soil changes rapidly not only temporally but also spatially, even at the micro level, and its physicochemical properties affect microbial distribution within the soil matrix, imposing important technical and methodological problems. Despite these caveats metagenomics constitutes a powerful approach to obtain information about the nature, composition and function of microbial communities in soil.

The specificity of the Rhizobium-legume symbiosis has classically allowed the use of most probable number (MPN) techniques to enumerate soil Rhizobia that are able to nodulate trap plants. For *R. leguminosarum* bv. viciae, representative abundances are on the order of 10^4-10^5 viable cells per gram of soil). Louvrier and collaborators developed a semi-selective medium to isolate *R. leguminosarum* directly from soils (Louvrier *et al*., 1995). The numbers they obtained in soils from Eastern France were *ca.* 10^4 per gram of soil. Overall, it can be concluded that, although cultivation of the plant host results in an increase in rhizobial soil counts (modest in the case of *R. leguminosarum*), established soil populations of *R. leguminosarum* are, at most, on the order of 10^4 to 10^5 per gram of soil. If typical soils contain *ca.* 10^9 bacteria per gram of soil, *R.*

leguminosarum would amount to less than 0.01% of the total soil microbiota. In practical terms, this implies that even in one of the largest metagenomic datasets (*ca.* 1 Tb), at most 100 Mb would be *R. leguminosarum* DNA. This, barring the crucial problem of how to specifically identify these sequences, would represent at most a 15x coverage of a single *R. leguminosarum* genome and would barely be representative of the population diversity. Thus, the low natural abundance of rhizobia in soils precludes the use of purely metagenomic methods to study their diversity and demands an alternative approach.

In view of these difficulties, we decided to adopt a Pool-Seq approach to the study of genotype selection by the host plant. Kofler and collaborators, working with *Drosophila melanogaster* populations, proposed for the first time the Pool-Seq term in 2011 for the next-generation sequencing and analysis of pooled DNA samples from natural populations. It constitutes a feasible (and affordable) genome-wide approach for comparison of population samples, thus allowing an easy scaling from the limitations of single markers to population genomics (Kofler *et al.*, 2011 and references therein).
We reasoned that sequencing pooled DNA samples from *R.leguminosarum* bv. viciae nodule isolates obtained from different legume plant hosts used as rhizobial traps would allow an experimental test of the hypothesis that different plant hosts select specific subpopulations of rhizobia from the available population present in a given soil. We compared four populations (*P. sativum, L. culinaris, V, sativa* and *V. faba*) originating from the same agricultural soil and consisting each of one hundred nodulae isolates that were grown independently and then pooled; the genomic DNA of the pools was extracted and sequenced (Illumina Hi-Seq 2000, 180 bp PE libraries, 100 bp reads, 12 Mreads) at BGI (Hong Kong and Shenzhen, China).
For analysis of the rhizobial Pool-Seq data, two specific considerations were taken into account.First, plant-specific subpopulations derive from the same unselected, resident soil population, whose genomic composition is, by definition, unknown, since their low numbers preclude any unselected genomic analysis. It is likely that this resident population contains both major and minor genomic types, resulting both from the soil's edapho-climatic properties and from its agricultural history. Thus, specific selection by the legume host will operate –if it does– on this original composition which, although distorted by the plant effect, will still be present in the plant specific isolates. Second, the large size and the multipartite composition of the *R. leguminosarum* genome favor both an open pan-genome and a large non-conserved genome. With *R. leguminosarum* bv. viciae we have estimated that 20-30% of the genes are strain specific. This suggests that plant host selection of specific rhizobial genotypes may implicate specific genes or groups of genes (eg. transport and metabolism of substrates). However, identification of these genes from Pool-Seq data is technically complex, since any DNA assembly will result, necessarily, in the formation of chimaeras with no biological meaning.
With these limitations in mind, we decided to restrict the Pool-Seq comparative data analysis to conserved genes, and reads for those genes were identified following recruitment by a reference genome, which in our case was that of *R. leguminosarum* bv. viciae 3841. A data analysis pipeline was designed and implemented, where reads are aligned to the reference genome, and both coverage and single nucleotide polymorphism (SNP) analysis are performed and compared between subpopulations. These analyses were carried out both for the complete genome and for relevant markers (16S rDNA, *nod* genes, *nif* genes, *recA* and *glnII* housekeeping markers, etc.). The data clearly show, both at the genome-wide and at the specific marker levels, that specific genotypes are indeed selected by the plant host, thus confirming previous indications.

A major outcome of this study is one of methodology for the study of natural rhizobial populations in the soil. Given unlimited resources for sequencing and data analysis, it is clear that individual genome sequencing of isolates, followed by assembly and multiple genome comparisons represents a more powerful tool than the Pool-Seq approach. However, such a situation is unlikely to occur, and the advantages and disadvantages of Pool-Seq must be evaluated for each project. When this project was designed (2010), it was not feasible to individually sequence and assemble 200 rhizobial strains. Even with today's higher capacity and lower costs, the pooled DNA approach allows for higher sequencing depths, and thus for potentially better descriptions of the populations, and for more facile analysis with our optimized pipeline. However, the Pool-Seq approach suffers from important drawbacks for this type of analysis in rhizobia. First, since the plant specific genotype enrichment will necessarily reflect the original structure of the rhizobial population in the soil, and this can vary from soil to soil, the analysis should be repeated with different types of soil. More importantly, the impossibility to assemble reads without the generation of chimeras makes it very difficult to identify specific genes that are not present in the reference genome but that may be specifically enriched in plant-selected subpopulations. These genes are important because they can provide not only specific host-linked markers but also evidence for the structural or functional nature of the phenotypes selected by the plant. We are addressing these limitations by means of two complementary approaches. First, the complexity of the plant-enriched subpopulations can be reduced by any number of typing methods, for instance RAPD analysis, making this reduced number of isolates more amenable to direct genome sequencing and assembly. Second, the Pool-Seq pipeline for coverage and SNP data analysis can be repeated with different *R. leguminosarum* reference genomes in order to incorporate coverage and SNP analyses for genes that were absent from the original reference genome. Results from both these strategies strengthen the power of the Pool-Seq approach with a mínimum investment in sequencing and data analysis.

ACKNOWLEDGMENTS

This work was supported by the Spanish Consolider-Ingenio Program (Microgen Project, CSD2009-00006) to J.I. We thank Rosabel Prieto for help with isolation of strains from root nodules and DNA preparations. We also thank Gonzalo Martín for IT.

REFERENCES

Ballard, R.A., *et al.* (2005). Aust. J. Exp. Agr. 45: 209-216.
Crossman, L.C., *et al*. (2008). PloS ONE 3: e2567.
Depret, G., *et al*. (2004). FEMS Microbiol. Ecol. 51: 87-97.
Galardini, M., *et al*. (2013). Genome Biol. Evol. 5: 542-558.
Harrison, P.W., *et al.* (2010). Trends Microbiol. 18: 141-148.
Knietsch, A., *et al*. (2003). J. Mol. Microbiol. Biotechnol. 5: 46-56.
Koch, M., *et al*. (2010). Mol. Plant-Microbe Interact. 23: 784-790.
Kofler, R., *et al*. (2011). Bioinformatics 27: 3435-3436.
Laguerre, G., *et al*. (2003). Appl. Environ. Microbiol. 69: 2276-2283.
Louvrier, P., *et al*. (1995). Soil Biol. Biochem. 27: 919-924.
Mazur, A., *et al*. (2011). BMC Microbiol. 11: 123.
Mutch, L.A., and Young, J.P.W. (2004). Mol. Ecol. 13: 2435-2444.
Palmer, K.M., and Young, J.P.W. (2000). Appl. Environ. Microbiol. 66: 2445-2450.
Petersen, J., *et al*. (2013). Appl. Microbiol. Biotechnol. 97: 2805-2815.
Surin, B.P., and Downie, J.A. (1989). Plant Mol. Biol. 12: 19-29.