Heredity (2015) 114, 494–501
© 2015 Macmillan Publishers Limited All rights reserved 0018-067X/15
www.nature.com/hdy

ORIGINAL ARTICLE

Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads

A Gouin[1,2], F Legeai[1,2], P Nouhaud[1], A Whibley[3], J-C Simon[1] and C Lemaitre[2]

Unmapped reads are often discarded from the analysis of whole-genome re-sequencing, but new biological information and insights can be uncovered through their analysis. In this paper, we investigate unmapped reads from the re-sequencing data of 33 pea aphid genomes from individuals specialized on different host plants. The unmapped reads for each individual were retrieved following mapping to the *Acyrthosiphon pisum* reference genome and its mitochondrial and symbiont genomes. These sets of unmapped reads were then cross-compared, revealing that a significant number of these unmapped sequences were conserved across individuals. Interestingly, sequences were most commonly shared between individuals adapted to the same host plant, suggesting that these sequences may contribute to the divergence between host plant specialized biotypes. Analysis of the contigs obtained from assembling the unmapped reads pooled by biotype allowed us to recover some divergent genomic regions previously excluded from analysis and to discover putative novel sequences of *A. pisum* and its symbionts. In conclusion, this study emphasizes the interest of the unmapped component of re-sequencing data sets and the potential loss of important information. We here propose strategies to aid the capture and interpretation of this information.

Heredity (2015) **114,** 494–501; doi:10.1038/hdy.2014.85; published online 1 October 2014

INTRODUCTION

Next-generation sequencing and whole-genome re-sequencing is nowadays commonly used to identify genomic variants that underlie phenotypic variations, genetic diseases, adaptation or speciation in natural populations. Typically, the reads are mapped against a reference genome, and the genotypes (that is, single-nucleotide polymorphism (SNP) and structural variant calls) are based on these mapped reads (Altshuler *et al*., 2010; Nielsen *et al*., 2011). In addition to universal caveats regarding unknown insertions and/or genomic contamination, which can be overlooked in pure mapping approaches, non-model organisms may suffer from the poor quality of the nuclear reference genome and incomplete symbiont or organellar genomes. Moreover, mapping is constrained by the level of divergence between the reads and the available reference sequence (Sousa and Hey, 2013). The resulting ascertainment bias could be problematic, especially when studying adaptation or speciation processes, as genomic regions of interest are expected to display important levels of divergence. These different issues produce a non-negligible fraction of unmapped reads, whose sequences are generally disregarded in favor of the mapped reads in the subsequent steps of the analysis, despite potentially containing useful information. This study offers one strategy for mining the unmapped reads in order to extract the relevant biological knowledge, leading to advice and recommendations for other re-sequencing projects.

We investigated this question in the context of a large-scale re-sequencing project on the pea aphid species complex. The pea aphid *Acyrthosiphon pisum* is a phytophagous insect that feeds on host plants of >20 Fabaceae genera. This species forms a complex of sympatric populations, or biotypes, each specialized on one or a few legume species (Simon *et al*., 2003; Via, 1991). Peccoud *et al*. (2009a) showed that these biotypes include at least eight partially reproductively isolated host races and three cryptic species, forming a gradient of specialization and differentiation potentially through ecological speciation. This complex of biotypes started to diverge between 8000 and 16 000 years ago, with a burst of diversification at an estimated 3600–9500 years (Peccoud *et al*., 2009b). In addition, the pea aphid is associated with an obligatory endosymbiont, *Buchnera aphidicola*, which is found in specialized cells called bacteriocytes and provides its host with essential amino acids. The pea aphid also harbors several facultative symbionts whose distribution is strongly correlated with plant specialization of their hosts (Simon *et al*., 2003; Ferrari *et al*., 2012; Henry *et al*., 2013), and it has been posited that some of these symbionts could have a role in plant adaptation, although clear evidence is still lacking (Tsuchida *et al*., 2004; McLean *et al*., 2011).

This study was carried out on 33 aphid re-sequenced genomes from 11 different plant-adapted biotypes. The reads were mapped against the *A. pisum* reference genome, its mitochondrial genome and its known obligate (*B. aphidicola*) and facultative symbiont genomes. The *A. pisum* genome (530 Mb) was assembled using a combination of sequencing technologies (International Aphid Genomics Consortium, 2010; www.aphidbase.com). Although a second version of the *A. pisum* reference genome has since been released (International

[1]INRA, UMR 1349 INRA/Agrocampus Ouest/Université Rennes 1, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Le Rheu, France; [2]INRIA/IRISA/GenScale, Campus de Beaulieu, Rennes, France and [3]Department of Cell and Developmental Biology, John Innes Centre, Norwich Research Park, Norwich, UK
Correspondence: Dr J-C Simon, UMR 1349 IGEPP, Domaine de la Motte, INRA, INRA/Agrocampus Ouest/Université Rennes 1, Le Rheu, Rennes 35653, France.
E-mail: claire.lemaitre@inria.fr
or Dr C Lemaitre, INRIA/IRISA/GenScale, Campus de Beaulieu, Rennes Cedex 35042, France.
E-mail: jean-christophe.simon@rennes.inra.fr

Received 2 April 2014; revised 29 July 2014; accepted 4 August 2014; published online 1 October 2014

Aphid Genomics Consortium, 2010), the genome assembly remains highly fragmented (23 924 scaffolds), and it has not been subjected to the same level of scrutiny and finishing as the genomes of model organisms, such as *Drosophila*. Moreover, symbiont genome sequences may not be well characterized for this species, and genomic divergence is expected to be important within the whole complex. As a result, a sizeable portion of the reads was not mapped.

In this paper, we scrutinized these unmapped reads by performing cross-comparisons between the sets, assembling the reads by biotype and analyzing the resulting contigs. We used tools developed for next-generation sequencing, such as *ABySS* (Simpson *et al.*, 2009) and *Compareads* (Maillet *et al.*, 2012), and more classical ones, such as the BLAST suite of tools (Altschul *et al.*, 1990). This analysis revealed that meaningful biological information is contained in the unmapped reads and could help to recover some divergent genomic regions previously excluded from analyses and to discover putative novel sequences of *A. pisum* and its symbionts.

MATERIALS AND METHODS

Next-generation sequencing data

Thirty-three pea aphid genomes were paired-end re-sequenced using the Illumina HiSeq 2000 instrument (Illumina inc., San Diego, CA, USA) with around 15× coverage for each genome. The individuals belonged to different populations each referred to as a biotype due to their adaptation to a specific host plant. In this study, 11 biotypes were each represented by 3 individuals (Supplementary Table S1 in Supplementary Material). Reads were 100 bp long, sequenced in pairs with a mean insert size of 250 bp and between 32.5 and 59.2 million read pairs (42.5 million on average) were obtained for each individual (see Supplementary Material). The fastq files of the paired reads from the 33 genomes were stored at the Sequence Read Archive of the National Center for Biotechnology Information database, of the BioProject ID PRJNA255937.

Reads were mapped using *Bowtie2* (Langmead and Salzberg, 2012) with default parameters (up to 10 mismatches per read, or fewer if indels are present—command-line in Supplementary Material) to a set of reference genomes. We also tested another popular mapper, BWA (Li and Durbin, 2009), but the percentage of unmapped reads was higher than for *Bowtie2* (on average over the 33 individuals, 6.1% vs 3.7% for BWA and *Bowtie2*, respectively). The reference set comprised the published pea aphid *A. pisum* reference genome (International Aphid Genomics Consortium, 2010) and its mitochondrial genome along with the genome of its primary bacterial symbiont and several secondary symbiont genomes reported for the pea aphid (*Hamiltonella defensa*, PAXS or X-type, *Regiella insecticola*, *Rickettsia* sp., *Rickettsiella* sp., *Serratia symbiotica*, *Spiroplasma* sp., *Wolbachia* sp., Oliver *et al.*, 2010; Russell *et al.*, 2013). When available, we took the reference genome sequence of the symbiont associated with the pea aphid (that is, *Hamiltonella defensa 5AT* (CP001277.1), *Regiella insecticola R5.15* (AGCA00000000.1), *Serratia symbiotica* str. Tucson (AENX00000000.1)), otherwise genomes of the closest symbionts were used as reference (that is, *Rickettsia* sp. endosymbiont of *Ixodes scapularis* (NZ_CM000770.1), *Rickettsiella grylli* (AAQJ00000000.2), *Spiroplasma melliferum* KC3 (AGBZ00000000.1) and *Wolbachia* sp. strain wRi (CP001391.1)). Note that we could not map reads to PAXS sequences, because no genome is currently available for this symbiont either for *A. pisum* or other host organisms. Various statistics about the quality of the mapping were recorded, and we calculated for each individual the average coverage for each reference genome used.

Extraction of unmapped reads

Fragments for which both reads of the pair did not map to the reference genomes were extracted from the BAM file (mapping result file) using *Samtools* features (Handsaker *et al.*, 2011). In order to check the quality of the unmapped reads, *Prinseq* (Schmieder and Edwards, 2011) was used. Sequences were trimmed if, working from the 3′ end of the read, base quality dropped below 20 within a window of 10 nucleotides. Read pair information was not preserved, and only sequences of at least 66 nucleotides in length were retained for the analysis. Quality-trimmed single-end unmapped read sets were used as the input to the pipeline.

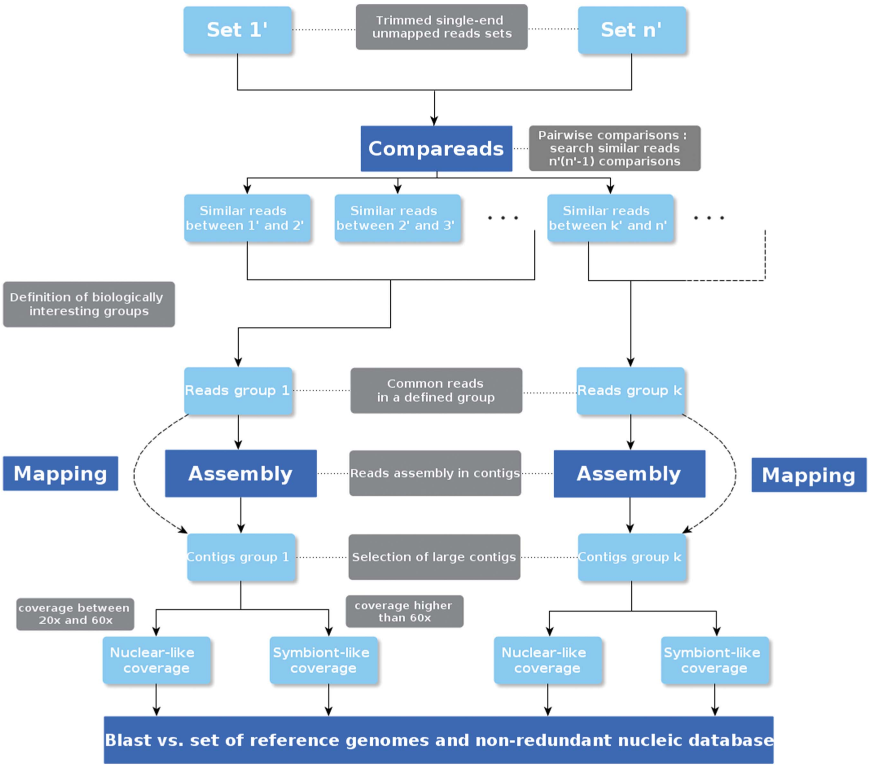
Set 1'
Trimmed single-end unmapped reads sets
Set n'
Compareads
Pairwise comparisons : search similar reads n'(n'-1) comparisons
Similar reads between 1' and 2'
Similar reads between 2' and 3'
. . .
Similar reads between k' and n'
. . .
Definition of biologically interesting groups
Reads group 1
Common reads in a defined group
Reads group k
Mapping
Assembly
Reads assembly in contigs
Assembly
Mapping
Contigs group 1
Selection of large contigs
Contigs group k
coverage between 20x and 60x
coverage higher than 60x
Nuclear-like coverage
Symbiont-like coverage
Nuclear-like coverage
Symbiont-like coverage
Blast vs. set of reference genomes and non-redundant nucleic database

Figure 1 Global overview of the pipeline followed for the analysis of unmapped reads.

Pipeline for the analysis of unmapped reads

The analysis pipeline, shown in Figure 1, was composed of three major stages: (i) pairwise comparisons between unmapped read sets, (ii) *de novo* assembly of pooled sets of reads, and (iii) analysis of the assembled contigs. Pairwise comparisons between the read sets were computed in order to identify biologically relevant signals and to define groups of individuals based on the quantity of similar reads. The second part consisted of the assembly of the common reads within previously defined groups. The contigs >1 kb were then analyzed in terms of size, read coverage and similarity to reference genomes.

Comparison of unmapped reads. *Compareads* (Maillet *et al.*, 2012) was used to compare the read content of the trimmed unmapped read set in each individual in a pairwise manner: this software can find similar reads between two sets of reads in an assembly-free manner. To be considered a match, a read of set A needs to share at least two non-overlapping k-mers of size 33 with at least one read of set B. This comparison thus gives two percentages of similarity between sets A and B: the percentage of reads of A similar to reads of B and *vice versa*. For all pairwise comparisons, a symmetric similarity score was also provided, computed as follows: $\frac{AinterB+BinterA}{N_A+N_B}$, with *AinterB* the number of reads in set A similar to reads in set B, *BinterA* the number of reads in set B similar to reads in set A, and N_A and N_B the total number of reads in sets A and B, respectively.

The 33 samples were classified based on this similarity measure, using the *R* (version 2.15) software with the maximum distance for the distance matrix and the complete linkage method for hierarchical clustering (function heatmap.2 from gplots package).

Assembly. We pooled common unmapped reads from the three individuals that belonged to the same biotype, that is, reads present in at least one pairwise comparison between individuals of a biotype were all concatenated in one fastq file. The *de novo* assembler *ABySS* (Simpson *et al.*, 2009) was used to assemble these common unmapped reads for each biotype. The size of the k-mer for the De Bruijn graph was set to 31.

To calculate the contig coverage statistics, the sets of unmapped reads were re-mapped to the obtained contig sequences using *Bowtie2* (default parameters), and the number of mapped reads was obtained using *Samtools*. In accordance with the mean coverage observed in the main data set (that is, where the pea aphid nuclear genome had an average coverage of 15× in each individual), and as reads from two to three individuals were pooled at this step, we considered those contigs with coverage ranging from 20× to 60× likely issued from the pea aphid (nuclear-like coverage), whereas contigs with higher coverage were considered more likely to derive from the symbionts (symbiont-like coverage) or repetitive sequences.

Comparison and analyses of contigs. *BLASTClust* was used to assess whether large homologous contigs (longer that 1 kb) could be found in different biotypes. A match was retained between two sequences if they were 80% identical over at least 90% of each sequence length. The contigs were then assayed by BLASTn search against the pea aphid reference genomes (nuclear, mitochondrion and symbionts) in order to ascertain their origin. Contigs with hits with an *e*-value $<$1e-50 were considered to represent highly divergent region of the *A. pisum* or its known symbiont genomes, that is, assumed to contain reads that could not be mapped during the first mapping step.

De novo assembly and characterization of an aphid symbiont genome

We performed the following analyses to assemble the genome of a bacterial symbiont detected in the unmapped read set of the individual Vc3. First, starting from the full read set of Vc3 (40.3 million read pairs), reads were filtered according to their k-mer coverage to obtain only the reads originating from the targeted genome and thus avoid simultaneously assembling the whole nuclear pea aphid genome. Given that the targeted genome had an average read depth in Vc3 of around 600×, only reads for which 68% of the length was covered by 31-mers present at least 100 times in the data set were retained, using *readFilter* (P Peterlongo *et al.*, unpublished) a custom software based on k-mer counts performed by the *DSK* software (Rizk *et al.*, 2013). Reads that could be mapped to the *B. aphidicola* or mitochondrial genomes were removed as their coverage levels would otherwise lead them to be retained in this read set. Only read pairs that remained intact following these filtering steps were included, and these pairs (which totaled 8.8 million read pairs) were assembled using *SPAdes* (Bankevich *et al.*, 2012), which has been reported to perform well with bacterial genomes (Magoc *et al.*, 2013). Several k-mer sizes were combined in *SPAdes* (31, 41, 63, 81, 89), with default values employed for the other parameters. We kept contigs >500 bp and removed those aligning with the non-*Spiroplasma* reference genomes. Alignments were performed with the global aligner *Mummer* (Kurtz *et al.*, 2004). We used *GeneMarkS+* (Besemer *et al.*, 2001) to predict proteins in the remaining contigs. These proteins were then compared with the NR database (version 22/01/2014) using BLASTp.

Identification and analysis of potentially divergent regions of the reference genome

To delineate potentially divergent regions of the reference genome that were present in the most divergent biotype (*Lathyrus pratensis*), contigs obtained from the unmapped reads of this biotype were aligned against the nuclear pea aphid genome with *Mummer*. The regions matching the reference genome with $>80\%$ identity and >500 bp were retained for further analyses.

In these regions, several metrics were computed, including the read depth at multiple mapping stringencies and SNP calling statistics. Read depth was computed first from the initial mapping obtained with Bowtie2 and also following mapping with *Stampy* (Lunter and Goodson, 2011), an aligner which is reported to perform well when mapping to a divergent reference. SNP calling statistics were collated from the results of the *GATK* (DePristo *et al.*, 2011) pipeline applied to the complete data set of 33 genomes. This pipeline consisted of PCR duplicate removal, indel realignment, base quality recalibration and genotyping with the UnifiedGenotyper. We used the number of 'undefined' calls, that is, polymorphic positions in the genome for which the genotype could not be determined by UnifiedGenotyper, as a proxy for alignment success. Finally, the gene content of these regions has been established using the version 2.1 of the official gene set of the pea aphid provided by AphidBase (Legeai *et al.*, 2010).

RESULTS

Mapping to reference genomes confirms variation in symbiotic composition between individual host genomes

The coverage of the *A. pisum* nuclear genome was 14.3× on average (min = 10.6× and max = 19.96×), whereas its mitochondrial genome was covered 946.0× on average (min = 257.09× and max = 3245.60×) and its obligate symbiont genome, 748.8× on average (min = 138.08× and max = 1509.03×). The coverage of the facultative symbiont genomes depended strongly on the individual host and varied from 0× to 117.7×. Observed variation for symbiont genome coverage among pea aphids was strongly linked to the infection status of the hosts. Indeed, when we compared the expected symbiotic composition based on PCR detection tests and results of mapping, we obtained a good match in most cases: the presence of a given symbiont as detected by a diagnostic PCR was confirmed by $>2\times$ coverage of reads that mapped against the reference genome (Supplementary Table S2 in Supplementary Material). There were, however, several exceptions to this pattern, namely *Ricketssia*, *Rickettsiella* and *Spiroplasma* symbionts for which genomes from a pea aphid host are currently not available. It should be noted that positive individuals for each of these three symbionts showed a weak but detectable number of reads that mapped against the closest reference genome found in databases (that is, Ps1, Ml1 and Ml3 individuals infected by *Rickettsia*, Tp3, Vc1 and Vc3 individuals infected by *Spiroplasma* and Ms1 individual infected by *Rickettsiella* in Supplementary Table S2). Note also that no reads mapped to the *Wolbachia* genome, confirming the absence of this symbiont in our selection of *A. pisum* genotypes.

A non-negligible fraction of reads does not map

For a given individual, there were between 0.6 and 7 million pairs of reads (mean = 1.3 million) where both reads did not map to any of the reference genomes (nuclear genome, mitochondrion or known symbionts). This constituted an average of 3.7% of the initial read sets. Moreover, most of these reads were of good quality, as shown in Figure 2, as few reads were removed (about 17%) by quality trimming (see Methods).

A direct analysis of these read data sets did not allow to characterize the unmapped reads in comparison to the mapped ones, in terms of sequence complexity (Shannon entropy) or signal for repeats (no enrichment of sequence matches with small RNAs targeting pea aphid transposable elements). However, the small size of the reads makes such direct analyses difficult and limits the sensitivity of such a characterization.

We can also see in Figure 2 that the fraction of unmapped reads varied between individuals. In particular, the individual Vc3 showed an atypically large amount of unmapped reads with >14 million reads representing 18.5% of the initial read set for this individual. For some biotypes, the fraction of unmapped reads was very similar across all individuals, perhaps implying a common cause of mapping failure. However, the fraction of reads did not seem to be correlated with the divergence of the individuals (or biotypes) with respect to the reference genome. The absence of this correlation suggests that the failure to map is not a simple consequence of inappropriate mapping parameters, as if mapping were too stringent we would expect to obtain a correlation between the unmapped fraction and biotype divergence from the reference.

Unmapped reads contain biologically meaningful information

Each set of unmapped reads was compared with all other sets using *Compareads*. Across the 1056 (33 × 32) pairwise comparisons, the percentage of common reads between two individuals varied greatly, from 6% to 95% with an average value of 50% (see Supplementary Figure S3 of Supplementary Material). For all but one individual, there was at least one other individual with which it shared 50% of its reads.

Interestingly, there was a significant difference when comparing individuals of the same biotype, where on average 70% of reads were shared between individuals, versus comparisons between individuals of different biotypes, which on average shared 48% of reads (P-value $<10^{-16}$ for the Welch two-sample test). This trend was confirmed by the hierarchical classification of individuals based on the pairwise similarity scores computed from the read set intersections (see Methods). Indeed, we can see in Figure 3 that individuals belonging to the same biotype were largely clustered together.

One extreme case is the *L. pratensis* biotype, which is known to be the most divergent biotype and is considered a cryptic species (Peccoud *et al.*, 2009a). It showed a very specific profile on the heatmap with strong similarity within this biotype (yellow group on Figure 3): a *L. pratensis* individual shared on average 72% of its unmapped reads with another *L. pratensis* individual, whereas only 23% were shared with an individual of another biotype.

These results show that the sets of unmapped reads contain sequence information specific to biotype or group of individuals and therefore may contain valuable sequences for biological analyses.

Where do these sequences come from?

In order to get longer and more readily interpretable sequences, we assembled them conjointly by biotype, using the assembler *ABySS*. Pools of unmapped read sets were used as inputs to obtain sufficient coverage for good quality assemblies. As the individual classification accorded with the biotype composition and because individuals from the same biotype are genetically closer than those from other biotypes (Peccoud *et al.*, 2009a), we pooled unmapped reads that were shared between at least two of the three individuals that belonged to the same biotype. By removing reads uniquely present in a single individual, putatively low coverage sequences were excluded, limiting one potential source of noise to the assembly process. Overall, 94 Mb of contig sequences, each ranging from 100 bp (shorter contigs were filtered) to 35.6 kb, were assembled. On average, 45% of the unmapped reads could be remapped to the assembled contigs. The average N50 was low (around 428 bp), but we obtained >11 800 contigs >1 kb (see Table 1).

The subsequent analysis considered contigs >1 kb in more detail. Coverage of the contigs varied considerably, with 57% of them having a nuclear-like coverage, that is, between 20× and 60× (see Material and Methods), consistent with an origin from the pea aphid nuclear genome. On the other hand, 14% of contigs had coverage >60×,

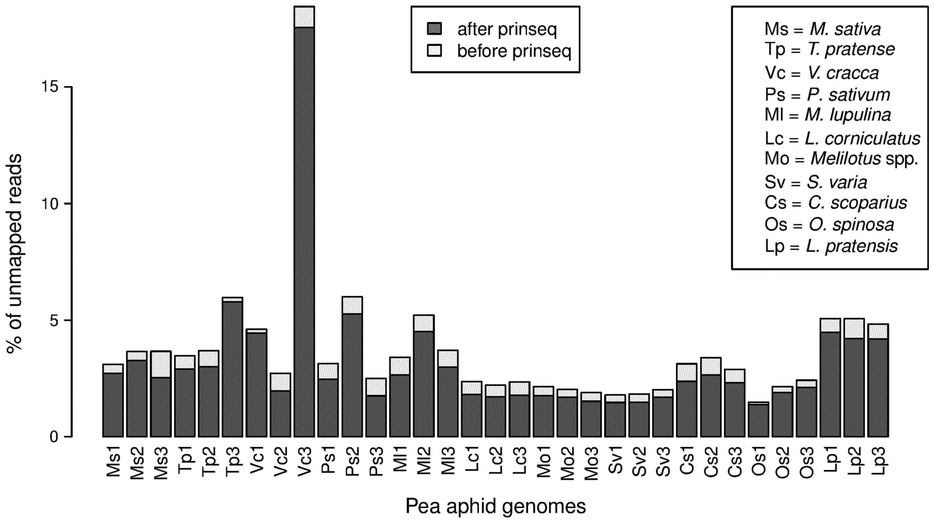

Figure 2 Percentage of unmapped reads (unmapped by pair) for each individual, after and before cleaning for quality. Individuals are grouped by biotype and sorted according to their known divergence with respect to the reference genome, the most divergent ones being at the right side of the figure.

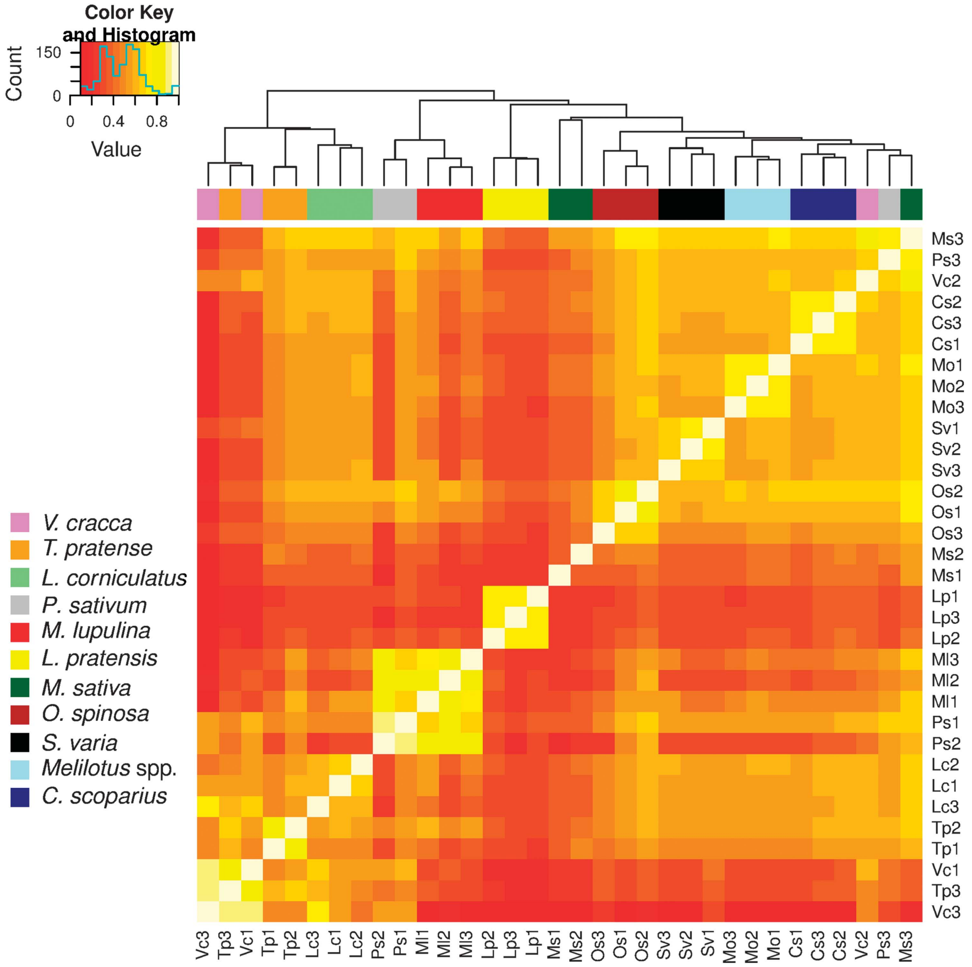

Figure 3 Hierarchical classification of the sets of unmapped reads. Each color below the tree corresponds to a biotype. Colors in the heatmap are function of the similarity score between two samples, from low similarity in red to high similarity in yellow.

which would be consistent with an origin from bacterial symbionts, the mitochondrion or repeated sequences. Contigs with coverage <20× (29%) could correspond to sequences from other microbes (including unreported symbionts) that are in low abundance in the aphid host.

Alignment of the contigs to the set of reference sequences can also suggest a potential genomic source. Overall, 63% of the contigs had a significant blast hit to one of the reference genomes, with the large majority matching with the nuclear pea aphid genome (89%).

As was found in the coverage analysis, the BLAST analysis revealed biotype-specific trends (Figure 4). Both approaches can thus be applied to classify the contigs into one of the two main origins: either symbiotic or nuclear. Moreover, the attributions by coverage and BLAST are largely consistent, with a concordant origin for 93% of the contigs with an origin assigned by both methods.

Sequences of symbiotic origin. Three biotypes contained a sizeable proportion of sequences with a putative symbiotic origin: *Pisum sativum*, *Vicia cracca* and *Medicago lupulina*. Both the *P. sativum* and *M. lupulina* biotype contig sequences predominantly showed significant similarity to reference symbiont genomes (see Figure 4). In line with this symbiont status, these contigs had a high coverage (>140× on average). The detected similarity was due to one particular symbiont genome: *Rickettsia* sp. endosymbiont of *Ixodes scapularis* (and included in the reference genome set). However, very few reads had mapped initially to this genome, which had an overall coverage of only 3× for the *P. sativum* and *M. lupulina* biotypes, and was absent from all other biotypes. This suggested that the chosen reference genome for *Rickettsia* was too distant from the actual pea aphid symbiont. By comparing these contigs to other *Rickettsia* species, we identified *R. bellii* as a more closely related species. This closer relationship was also confirmed by a phylogenetic analysis of 16S ribosomal RNA genes from all available *Rickettsia* species having their complete genome sequenced (Supplementary Figure S1 in Supplementary Material). Substituting this genome as a reference resulted in an improved coverage for both *P. sativum* and *M. lupulina* biotypes and confirmed the presence of this facultative symbiont in individuals that showed negligible coverage when their reads were mapped to *R. ixodes* (Supplementary Table S2 in Supplementary Material). However, the discrepancy between this coverage level and that observed in the assembled contigs, which was over twofold higher, suggests that *Rickettsia* from *A. pisum* may diverge significantly from *R. bellii*, which requires to characterize its genome.

Unlike the *P. sativum* and *M. lupulina* biotypes, the *V. cracca* biotype contigs showed almost no similarity to the reference symbionts despite a coverage signal that averaged 602× and was thus consistent with symbiont origin. When aligning these contigs to the NR nucleic acid database, we found few matches and typically only low similarity scores but noted that these hits were enriched in

Table 1 Contig statistics

Biotype	n *reads (M)*	*Contigs > 100 bp*				*Contigs > 1 kb*		
		nb	*assbl. Mb*	*% reads*	*N50*	*nb*	*assbl. Mb*	*% reads*
M. sativa	3.68	21 110	5.98	41.29	380	669	1.09	18.61
T. pratense	7.07	29 298	8.75	40.25	415	1107	2.09	19.04
V. cracca	18.29	21 907	7.41	39.00	520	1135	2.25	26.00
P. sativum	6.26	21 123	7.13	49.34	510	1055	2.12	39.1
M. lupulina	7.56	20 932	7.01	48.66	508	1075	2.11	37.14
L. corniculatus	3.34	25 772	7.43	47.95	403	869	1.43	21.99
Melilotus spp.	3.68	23 792	6.9	44.21	408	879	1.41	18.83
S. varia	2.96	23 340	6.75	50.18	402	839	1.33	24.35
C. scoparius	5.01	27 081	7.84	33.55	410	1026	1.65	13.55
O. spinosa	3.67	25 170	7.4	45.84	418	977	1.56	20.46
L. pratensis	8.98	83 344	21.41	53.2	331	2211	3.42	22.67

For each biotype, the number of unmapped reads in million (*n* reads) used for the assembly is indicated along with several statistics describing the properties for two contig length cutoffs (100 bp and 1 kb), namely, the number of obtained contigs (nb), their cumulative length (assbl. Mb), the percentage of reads (% reads) that could be mapped to the contigs and the N50 value.

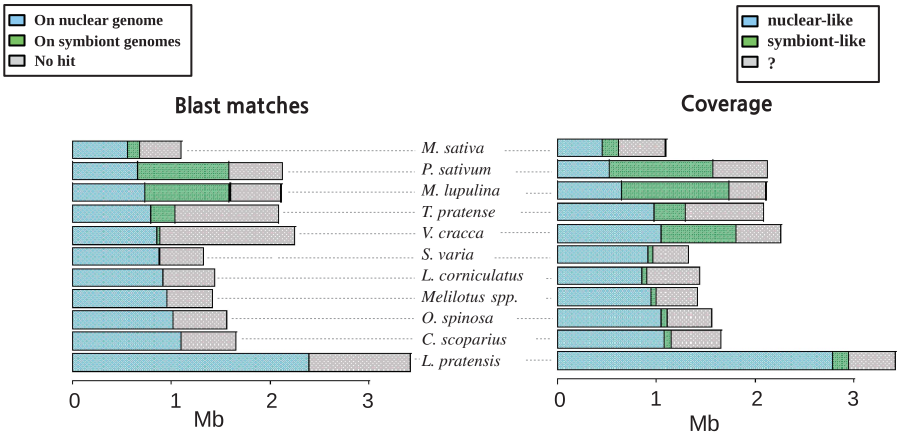

Figure 4 Analysis of contigs >1 kb in terms of blast matches and read coverage.

sequences from the Mollicute group and, more precisely, the *Spiroplasma* genus. This implied that some individual genomes of the pea aphid contained a *Spiroplasma* symbiont whose genome is distant from available *Spiroplasma* genomes in the public databases. This hypothesis was confirmed by the fact that three genotypes (of which two *V. cracca*) of the pea aphid were positive for *Spiroplasma* infection based on PCR-specific detection (Supplementary Table S2 in supplementary material). Therefore *Spiroplasma* was likely present in at least three individuals but in high abundance in only one: Vc3. Indeed, most of the *V. cracca* unmapped reads came from this single *V. cracca* individual, which had five times more unmapped reads than the average (>14 million reads, see Figure 2). Based on the hierarchical classification in Figure 3, Vc3 was grouped with the individuals Vc1 and Tp3 (in agreement with PCR results), with 90.5 and 95% of its unmapped reads being similar to these two individuals, respectively. The high abundance of reads from this uncharacterized source in Vc3 led us to attempt the *de novo* assembly of this *Spiroplasma* genome. The assembly was performed with *SPAdes*, after having extracted only putative '*Spiroplasma*' reads from the full Vc3 read set (see Material and Methods). The final assembly contained 509 contigs >500 bp (2442 bp on average), totaling 1.2 Mb of sequence. Although at the nucleotide level, these contigs showed weak similarity to available *Spiroplasma* genomes, at the protein level their relationship with Mollicute (and mainly *Spiroplasma*) proteins was confirmed for 546 annotated genes (on contigs summing to 780 kb). Moreover, the assembly size, low GC content (24%) and the fragmented assembly were consistent with known *Spiroplasma* genome features: the genome size varies from 1.4 to 1.9 Mb, GC content is around 26%, and the genomes contain lots of repeated sequences and viral elements that make the assembly task harder (Carle *et al*., 2010; Lo *et al*., 2013). Additionally, a phylogenetic analysis of its 16S RNA gene confirmed its membership to the *Spiroplasma* genus and the absence of any close relative with a complete genome available in the databases (Supplementary Figure S2 in Supplementary Material).

Finally, when unmapped reads were re-mapped to the partially assembled genome of *Spiroplasma* isolated from *A. pisum*, individuals which had been found to be positive for *Spiroplasma* by a PCR-based diagnostic assay but which had a negligible coverage when their reads were mapped to *S. melliferum* registered a high coverage on contigs from the *A. pisum*-derived *Spiroplasma* (up to 1185× for some individuals, Supplementary Table S2,Supplementary Material). In one case, the sensitivity of the sequence analysis may have exceeded that of the PCR test as individual Lc3 was PCR-negative for *Spiroplasma* but recorded on re-mapping, suggesting a possible infection under the threshold of PCR detection.

Sequences of nuclear origin. All biotypes possessed contigs with a putative nuclear origin, as shown on Figure 4. Some of these contigs were similar between several biotypes or even between all biotypes. We clustered the contigs together using BlastClust and obtained overall 10.1 Mb of distinct sequences having a nuclear-like coverage, of which 4.2 Mb had no similarity to the reference genome of *A. pisum*. Some of these are likely to be insertion polymorphisms, whereas the 8.6 kb that are shared in at least eight biotypes could represent pea aphid sequences missing from the current reference assembly either due to error or to deletions in the individual genome that was used to build the reference genome.

Aside from these common sequences, the *L. pratensis* biotype was particularly enriched in sequences with a putative nuclear origin (Figure 4). Most of its contig sequences had a significant blast hit to the nuclear reference genome (2.4 Mb (69.8%) of total contig length) and a nuclear-like coverage (86% of total length), suggesting that these contigs were assembled from reads that were too divergent to map in the first place.

One thousand one hudred and thirty-seven regions (covering 1001 kb) that exhibit similarity to a *L. pratensis* contig over at least 500 bp were then delimited on the reference genome, using the global aligner *Mummer*. The analysis of read coverage in these regions uncovered two types of region: 'low-coverage' regions in which very few reads had mapped (coverage $<30\times$ for the three *L. pratensis* individuals combined, 377 regions summing to 337 kb), and 'normal-to-high-coverage' regions (760 regions, 663 kb). Although the latter could be explained by one or several divergent copies not present in the reference genome, the former are likely to be regions that are too divergent in all the *L. pratensis* genomes, in which we may miss important biological information. Indeed, we observed a high proportion of undefined SNP calls for *L. pratensis* samples in these 'low-coverage' regions. On average, each *L. pratensis* individual had 61% of undefined calls, whereas this percentage never exceeded 20% in these same regions for other biotypes (with an average of 13%). These are high values compared with the proportion of undefined calls over the whole genome (on average, 9% for *L. pratensis* samples and 3.7% for other biotypes). Moreover, half of the regions showed $>50\%$ of undefined calls for all three *L. pratensis* individuals. This supports the assertion that SNP information is lost because of unmapped reads.

When using a more sensitive mapping approach with *Stampy*, some of these missed SNPs could be recovered. For the three *L. pratensis* genomes, overall 64% of the initially unmapped reads were re-mapped onto the set of reference genomes. Among these rescued reads, 0.66% mapped to the 'low-coverage' regions, which was more than expected knowing that these regions of interest represent only 0.06% of the whole genome. This sensitive mapping enabled recovery of, on average, 60% of undefined SNP calls, with 12.5% of regions completely resolved (that is, with no undefined SNPs). However, for 54% of the regions, the total coverage (*Bowtie2+Stampy*) still did not reach normal levels and remained $<30\times$.

DISCUSSION AND CONCLUSION

Although approaches for mining unmapped read sets for specific purposes have been described, for example, for pathogen discovery (for example, Kostic *et al*., 2011), this portion of reads is typically disregarded in re-sequencing projects. The sources of unmapped reads are various: they may derive from characterized or uncharacterized symbionts, bacterial, viral or eukaryotic pathogens, highly divergent genomic regions, genomic insertion sequences or library contaminants. The relative proportions of these contributions can vary, and factors such as reference genome quality and the genetic distance between reference and target can have a major role. Therefore, in non-model systems, both the contribution of unmapped reads to the data set and the likelihood that these reads are a reservoir of useful biological information are increased. However, identifying the unmapped reads and ascribing them to specific sources is not trivial. We have here proposed a novel approach to rescue some of this potentially lost information and have explored the unmapped read sets in the context of 33 re-sequenced genomes from biotypes of the pea aphid species complex.

The direct pairwise comparisons of read sets, before assembly, enabled the rapid identification of similar read sets and highlighted atypical samples and biotypes. Moreover, as the coverage of each individual alone was too low to expect a good quality assembly, merging samples in order to achieve sufficient coverage was necessary for *de novo* assembly quality. However, selecting and merging only reads common to a single biotype or population would preclude the identification of other interesting sequences specific to one genotype or to a combination of individuals of different biotypes. Therefore a more in-depth analysis of the pairwise comparisons followed by the assembly of particular combinations of read sets could be interesting to conduct and may help to uncover unexpected links between individuals.

The assembly phase generated longer sequences than the preprocessed read sets, and these can be more efficiently analyzed and compared with sequence databases. However, although bacterial sequences, such as the ones obtained from *Rickettsia* and *Spiroplasma*, could be relatively easily assembled and led to large contigs, we observed that the remaining contigs were usually very short (N50 around 400 bp), and probably one of the consequence is that a large fraction of the unmapped read sets could not be remapped on the shortest contigs. The recovery of short contigs may be influenced by our methods: extracting and assembling only unmapped read pairs would mean that we assemble only regions of high divergence, which may be interspersed in the genome with less divergent regions that are well served by the mapping. In this case, we would predict that samples from the more divergent populations would have, on average, larger contigs of nuclear origin. This is supported in our case by the most distant biotype from the reference, *L. pratensis*, which shows the greatest proportion of large contigs with similarity with the nuclear genome and *de facto* the highest number of remapped reads (53.2%).

The final step of our approach was to align the contigs against the reference genomes (nuclear and symbionts) with less stringent similarity criteria than those used during the first mapping step of our process. This, together with the average read coverage of contigs, allowed us to ascribe a putative origin (nuclear or symbiont) of most of the larger contigs. For contigs of symbiont origin, this revealed notably the mis-specification of a reference genome and identified a closer representative species. Without this analysis, we would have concluded from the first mapping that this symbiont was absent (or at very low abundance) from all individuals. Moreover, this revealed the presence in three individuals of a symbiotic bacterium of the genus *Spiroplasma*, which has been previously reported for the pea aphid (Fukatsu *et al*., 2001) but never sequenced and for which we produced a first draft assembly of its genome. Again, the presence of this symbiont would have been completely missed with the first mapping.

In addition, this analysis allowed us to highlight specific parts of the nuclear genome that are enriched in the unmapped read set. These are large regions which are either absent from the reference genome or show high divergence to the corresponding reference sequence such that each of the read pairs originating from it cannot be mapped. The latter explanation seems to be the most frequent in our data set. This highlights the major drawback of classical comparative genomics approaches relying on a reference genome. The regions of the reference genome with important genomic divergence for some individuals will contain fewer mapped reads from these individuals and ultimately little divergence will be detected, leading to an erroneous interpretation. We confirmed this consequence of unmapped reads by observing an increased level of unassigned genotypes in these particular parts of the reference genome for the most divergent biotype.

This mapping issue could lead to the loss of valuable biological information or biases in the analysis of genomic variation. Careful calibration of mapping parameters to better handle sequence mismatches between reads and the reference genome can reduce the fraction of reads that cannot be mapped. We explored this by using the Stampy aligner to reprocess the unmapped reads and could recover 64%. Although this offers an improvement on the original Bowtie mapping, most of the observed regions with missing genotype information remained unresolved, and it is important to note that relaxing these settings will increase false positive mapping and also increases the time and computing resources required to process the data sets.

Here, our approach helped to recover those divergent regions, and having applied this strategy, the biological signals and functions of these regions can then be interrogated. In the case of the pea aphid data set, the genic content of the regions will be investigated with a view to determining whether they are enriched in genes involved in host-plant adaptation (for example, receptors and enzymes). More generally, recovery of these regions enabled them to be subjected to further study, for example, to identify signatures of positive selection.

DATA ARCHIVING

Fastq files of the paired reads from the 33 genomes are available from the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) database: BioProject ID PRJNA255937.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by ANR Blanc SPECIAPHID (ANR 11 BSV7 005 01) and by the Région Bretagne (Adaptro). We warmly thank Jean Peccoud for aphid collection and fruitful discussions and Lucie Mieuzet for DNA extractions.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG *et al.* (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al.* (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Besemer J, Lomsadze A, Borodovsky M (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* **29**: 2607–2618.

Carle P, Saillard C, Carrere N, Carrere S, Duret S, Eveillard S *et al.* (2010). Partial chromosome sequence of *Spiroplasma citri* reveals extensive viral invasion and important gene decay. *Appl Environ Microbiol* **76**: 3420–3426.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Ferrari J, West JA, Via S, Godfray HCJ (2012). Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. *Evolution* **66**: 375–390.

Fukatsu T, Tsuchida T, Nikoh N, Koga R (2001). Spiroplasma symbiont of the pea aphid, *Acyrthosiphon pisum* (Insecta: Homoptera). *Appl Environ Microbiol* **67**: 1284–1291.

Handsaker RE, Korn JM, Nemesh J, McCarroll SA (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**: 269–276.

Henry LM, Peccoud J, Simon JC, Hadfield JD, Maiden MJC, Ferrari J *et al.* (2013). Horizontally transmitted symbionts and host colonization of ecological niches. *Curr Biol* **23**: 1713–1717.

International Aphid Genomics Consortium (2010). Genome sequence of the pea aphid *Acyrthosiphon pisum. PLoS Biol* **8**: e1000313.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G *et al.* (2011). PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29**: 4–7.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–U354.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* **25**: 1754–1760.

Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispe C, Collin O *et al.* (2010). AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol Biol* **19**: 5–12.

Lo W-S, Chen L-L, Chung W-C, Gasparich GE, Kuo C-H (2013). Comparative genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. *BMC Genomics* **14**: 22.

Lunter G, Goodson M (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* **21**: 936–939.

Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D *et al.* (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* **29**: 1718–1725.

Maillet N, Lemaitre C, Chikhi R, Lavenier D, Peterlongo P (2012). Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics* **13** (Suppl 19): S10.

McLean AHC, van Asch M, Ferrari J, Godfray HCJ (2011). Effects of bacterial secondary symbionts on host plant use in pea aphids. *Proc R Soc Biol Sci Ser B* **278**: 760–766.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Oliver KM, Degnan PH, Burke GR, Moran NA (2010). Facultative symbionts in aphids and the horizontal transfer of ecologically important traits. *Annu Rev Entomol* **55**: 247–266.

Peccoud J, Ollivier A, Plantegenest M, Simon J-C (2009a). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proc Natl Acad Sci USA* **106**: 7495–7500.

Peccoud J, Simon J-C, McLaughlin HJ, Moran NA (2009b). Post-Pleistocene radiation of the pea aphid complex revealed by rapidly evolving endosymbionts. *Proc Natl Acad Sci USA* **106**: 16315–16320.

Rizk G, Lavenier D, Chikhi R (2013). DSK: k-mer counting with very low memory usage. *Bioinformatics* **29**: 652–653.

Russell JA, Weldon S, Smith AH, Kim KL, Hu Y, Lukasik P *et al.* (2013). Uncovering symbiont-driven genetic diversity across North American pea aphids. *Mol Ecol* **22**: 2045–2059.

Schmieder R, Edwards R (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Simon J-C, Carré S, Boutin M, Prunier–Leterme N, Sabater–Muñoz B, Latorre A *et al.* (2003). Host-based divergence in populations of the pea aphid: insights from nuclear markers and the prevalence of facultative symbionts. *Proc R Soc Biol Sci Ser B* **270**: 1703–1712.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117–1123.

Sousa V, Hey J (2013). Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Rev Genet* **14**: 404–414.

Tsuchida T, Koga R, Fukatsu T (2004). Host plant specialization governed by facultative symbiont. *Science* **303**: 1989.

Via S (1991). Specialized host plant performance of pea aphid clones is not altered by experience. *Ecology* **72**: 1420–1427.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)Supplementary Information