Bioinformatics Journal Club

22.05.2015

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

A Gouin, F Legeai, P Nouhaud, A Whibley, J-C Simon and C Lemaitre

Heredity: Volume 114, Issue 5 (May 2015)
Special Issue on Environmental Genomics

This special issue reflects the recent advances in the field of environmental genomics and exposes the attractive prospects in the light of the new, rapidly-evolving tools that are next generation sequencing (NGS) approaches. Understanding the ecology, evolution, adaptation and biodiversity of organisms in their ecosystems is one of the most challenging scientific issues to which NGS and other "omics" may contribute. The papers deal with a broad range of organisms (eukaryotes and prokaryotes) and environments, including biotic interactions and methodological enhancements. **The special issue highlights the exciting paths opened by NGS in environmental genomics and the novel opportunities to go deeper in the understanding and characterization of complex biological systems.**

Lehetäide elust. Mõistvalt

Külli Hiiesaar

“Eesti Loodus”, Juuli 1997

*Lehetäide elu pole kerge**. Kui oled väike, õrn, kaitsetu ja peale selle veel kohmakas**, on maailm vaenlasi täis. Põgenemislootused on väikesed, sest hüpata nad ei oska, jalad kaugele ei kanna ja tiivad on ainult periooditi. Ähvarduspoosi pole mõtet võttagi, sellega ei hirmuta küll kedagi, keha mürki ei sisalda ning suised kõlbavad vaid taimemahla imemiseks.*

Varjevärvus võib ära petta ainult inimese ja sedagi ajutiselt. "Klassikaaslastega" see ei õnnestu. Neid ajendab nälg ja sigimisinstinkt ning nad on varustatud kõige mitmekülgsema relvaarsenaliga. Nende lõhnaretseptorid aitavad juba kaugelt ohvrit avastada, hästi liikuv pea ja suured silmad täpsustavad asukoha, kompides ja maitstes tehakse kindlaks saagi söögikõlblikkus. Saagimaiaid röövleid kannavad kohale tiivad või väledad jalad, ohvriga toimetulekuks on tugevad suised, mürk ja kavalus.

***Ainus võimalus ellu jääda on kiiresti sigida**. Lehetäisid on maailmas kirjeldatud 20 000 liiki, kuid tõenäoliselt on neid veelgi rohkem. Neid võib näha nii aias, põllul, metsas kui ka koduaknal lillepotis, nii taime maapealsel kui ka maa-alusel osal.*

***Päris ilma sõpradeta ei ole isegi lehetäi**. Elades sipelgatega sümbioosis, on kasu vastastikune: sipelgad pakuvad lehetäile kaitset vaenlaste rünnakute eest, vastutasuks saavad magusat nestet toiduks. Kuid ega sõpradegi peale saa alati kindel olla: kui sipelgatel tekib valgulise toidu nappus, süüakse lehetäi ära.*

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. **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.

Friends on board:

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, and it has been posited that some of these symbionts could have a role in plant adaptation, although clear evidence is still lacking.

Aim of the study:

- 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.**
- **Read mapping is constrained by the level of divergence** between the reads and the available reference sequence.
- **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.**

Datasets:

- 33 pea aphid genomes were paired-end re-sequenced using the Illumina HiSeq 2000 instrument 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.
- 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.

Toolbox:

- Read mapping - **Bowtie2** (Langmead and Salzberg, 2012), ~~**BWA**~~ (Li and Durbin, 2009), **Stampy** (Lunter and Goodson, 2011)
- Read quality filtering - **Prinseq** (Schmieder and Edwards, 2011)
- Assembly - **ABySS** (Simpson et al., 2009), – **SPAdes** (Bankevich et al., 2012)
- Read sorting - **Compareads** (Maillet et al., 2012)
- Alignment **- BLAST suite** of tools (Altschul et al., 1990), the global aligner **Mummer** (Kurtz et al., 2004)
- Bam file processing & coverage statistics - **samtools** features (Handsaker et al., 2011)

Toolbox: (2)

- **GeneMarkS+** (Besemer et al., 2001) to predict proteins in the remaining contigs.
- SNP calling statistics were collated from the results of the ***GATK*** (DePristo *et al.*, 2011). 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.
- The gene content of the regions was established using the version 2.1 of the official gene set of the pea aphid provided by **AphidBase** (Legeai et al., 2010).

Workflow:

Read mapping:
Bowtie2 *Ref: A. pisum* reference genomic + mitochondrion + primary bacterial and several secondary symbiont genomes

Extraction of unmapped reads:
samtools+Prinseq Retain: reads with both in the pair unmapped; > Q20 quality cutoff, >66 bp, converted to fragment read library

Pipeline for the analysis of unmapped reads

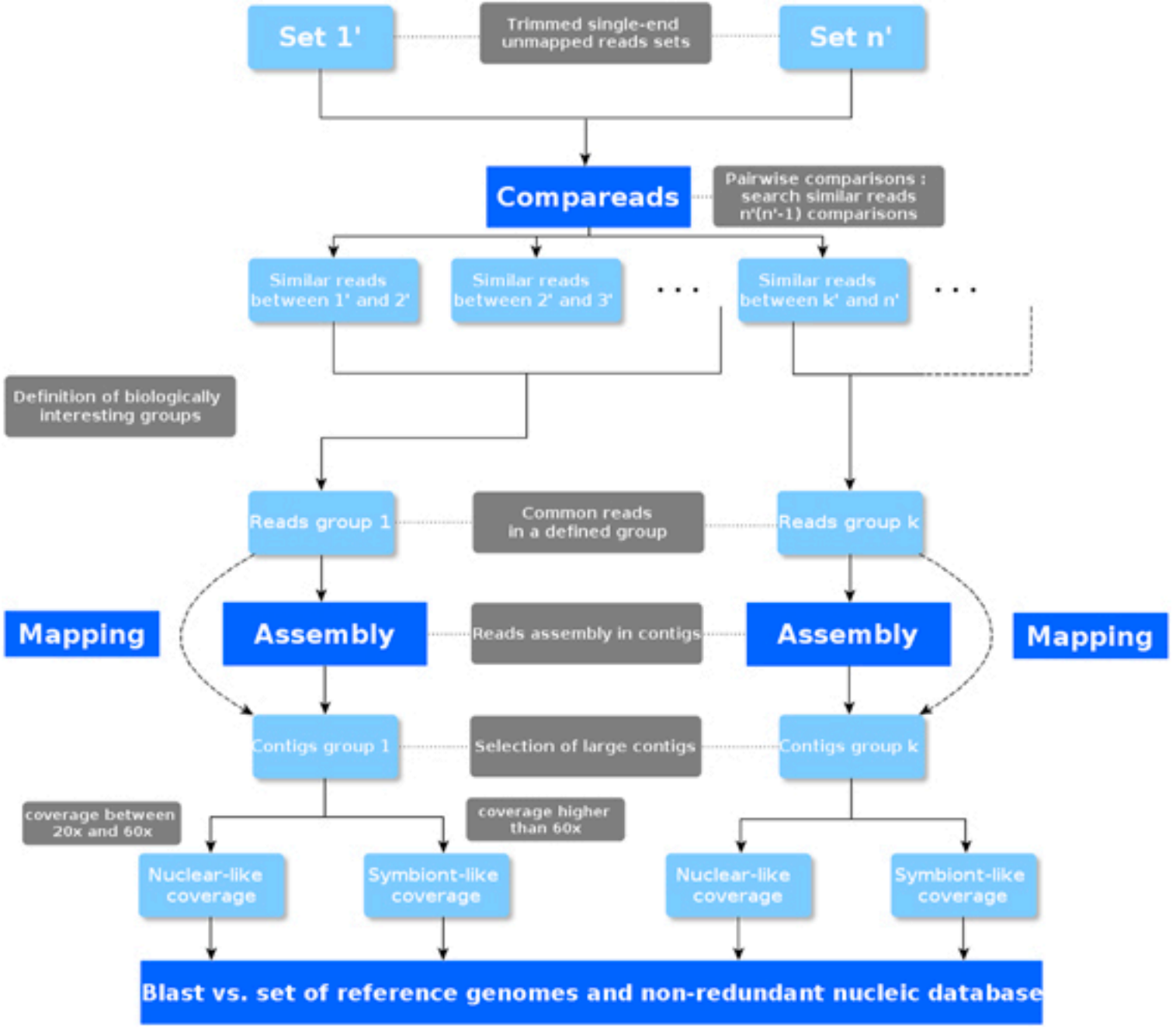
Comparison of unmapped reads

Assembly

Comparison and analyses of contigs

***De novo* assembly and characterization of an aphid symbiont genome**

Identification and analysis of potentially divergent regions of the reference genome

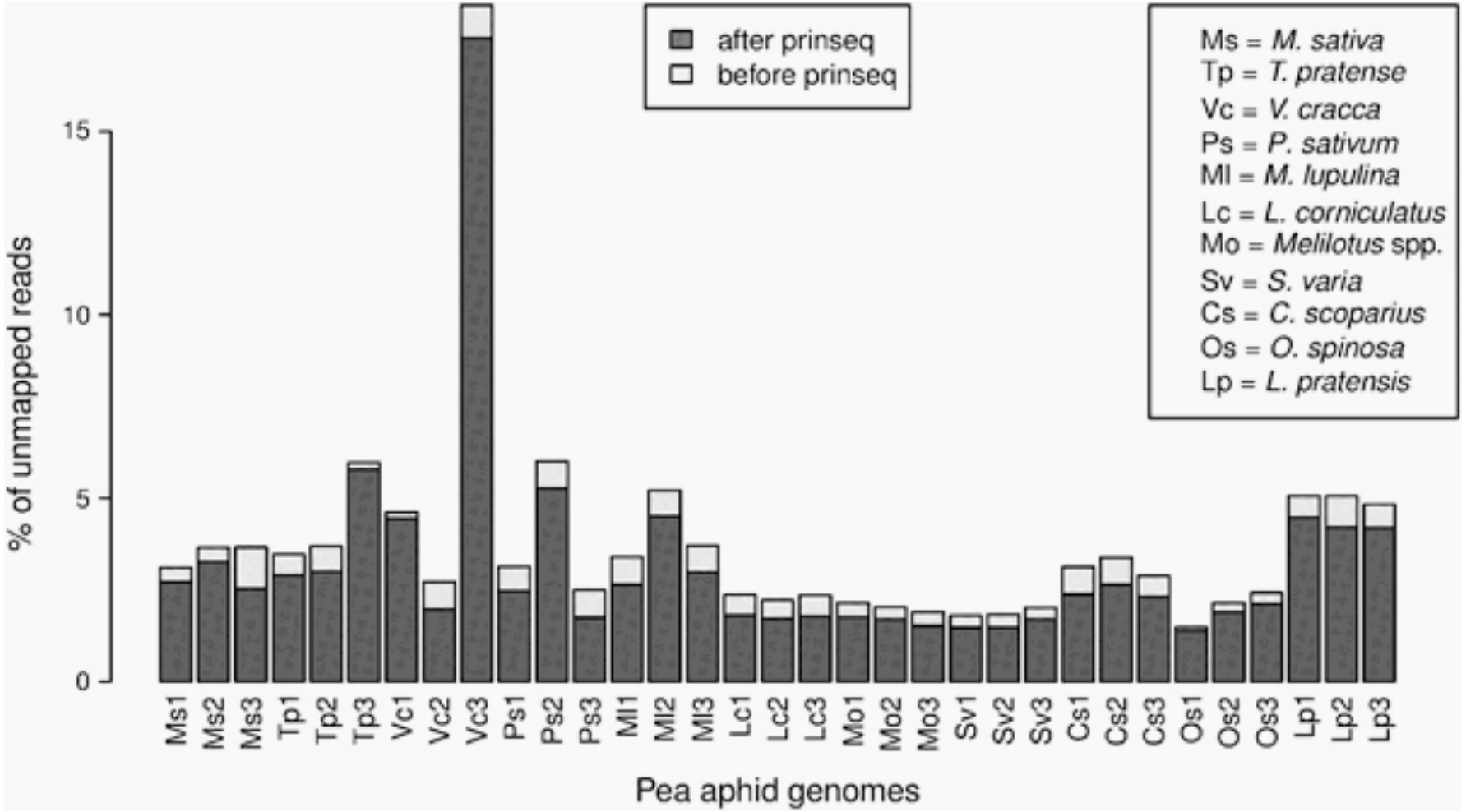


Global overview of the pipeline followed for the analysis of unmapped reads.

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

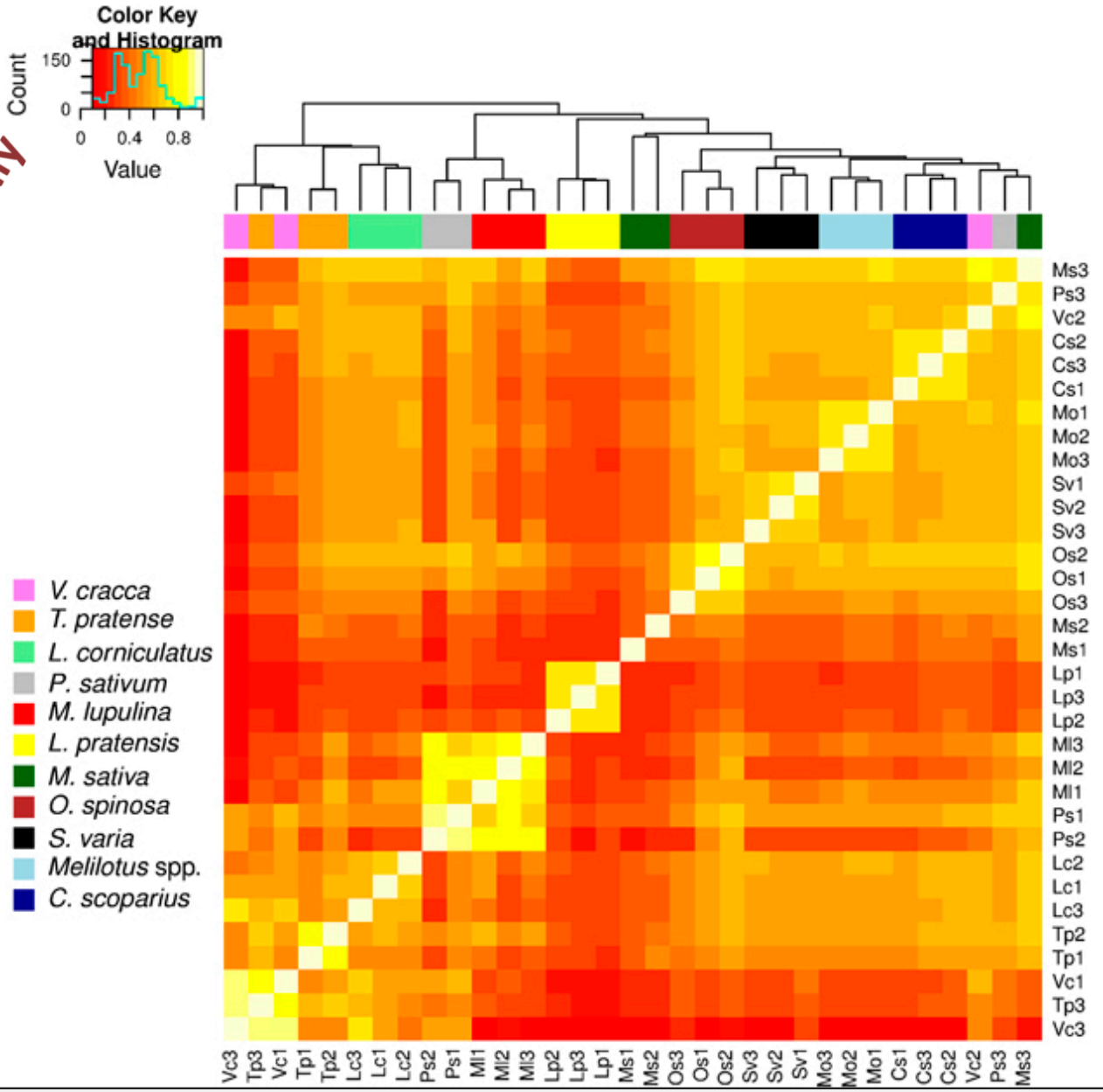
- The coverage of the *A. pisum* nuclear genome 14.3 × (min=10.6 × and max=19.96 ×),
- mitochondrial genome 946.0 × (min=257.09 × and max=3245.60 ×)
- its obligate symbiont genome, 748.8 × on average (min=138.08 × and max=1509.03 ×)
- The coverage of the facultative symbiont genomes depended strongly on the individual host and varied from 0 × to 117.7 × * .

* Good data correlation: Their presence of a given symbiont as detected by diagnostic PCR was confirmed by >2 × coverage of reads that mapped against the reference genome.



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.

Do unmapped reads contain biologically meaningful information ?



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.

Where do these sequences come from?

- reads from the same biotope were pooled
- unique sequences removed
- assambled contjointly by biotype, using the Assambler ABySS

⇒ 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

⇒ 57% of them having a nuclear-like coverage (20-60x)

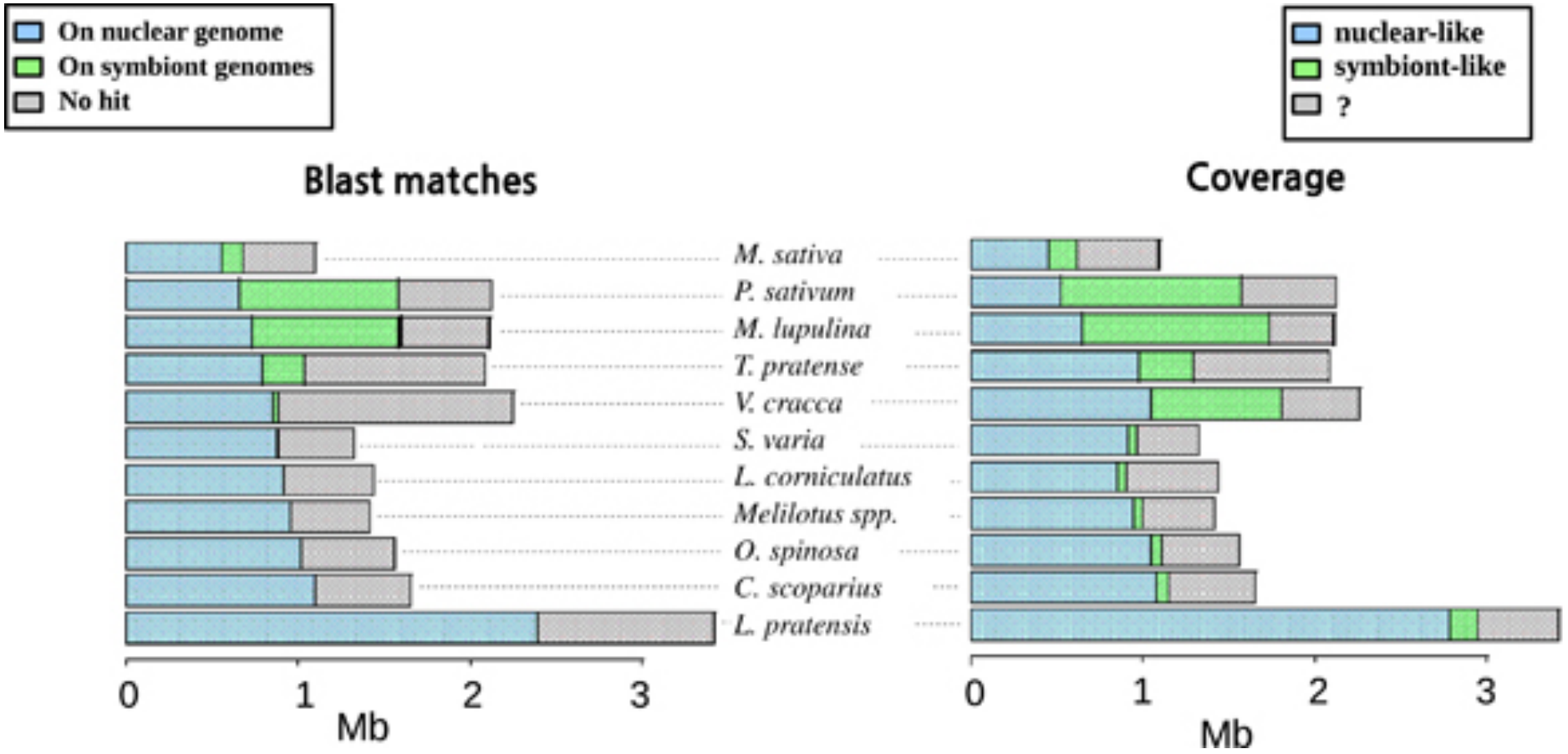
⇒ 14% of contigs had coverage >60 × , which would be consistent with an origin from bacterial symbionts, the mitochondrion or repeated sequences

⇒ Contigs with coverage <20 × (29%) could correspond to sequences from other microbes (including unreported symbionts) that are in low abundance in the aphid host.

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

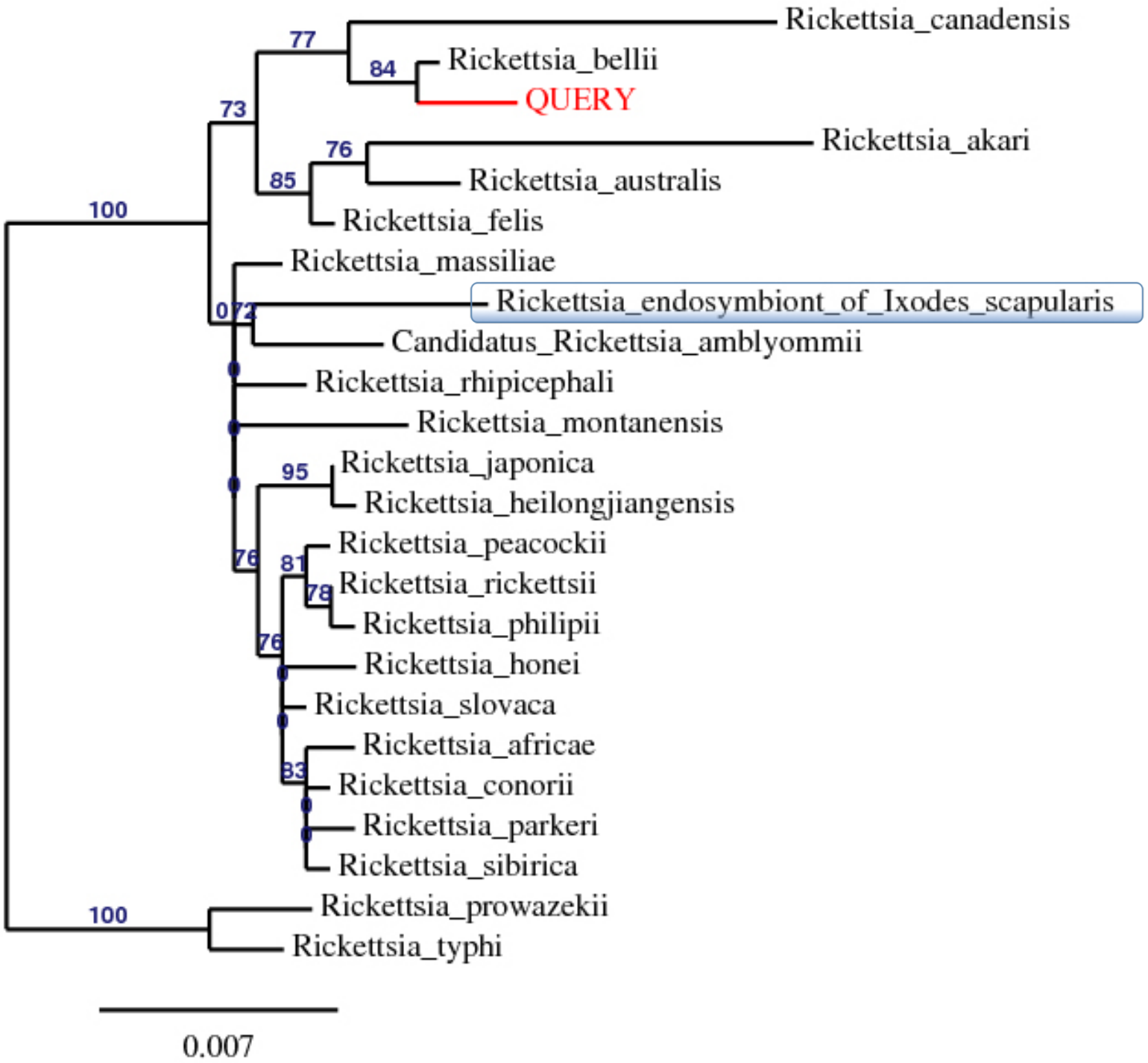
For each biotype, the number of unmapped reads in million (*n* reads) used for the assembly is indicated along with several statistics describing the properties for two contig length cutoffs (100 bp and 1 kb), namely, the number of obtained contigs (nb), their cumulative length (assbl. Mb), the percentage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

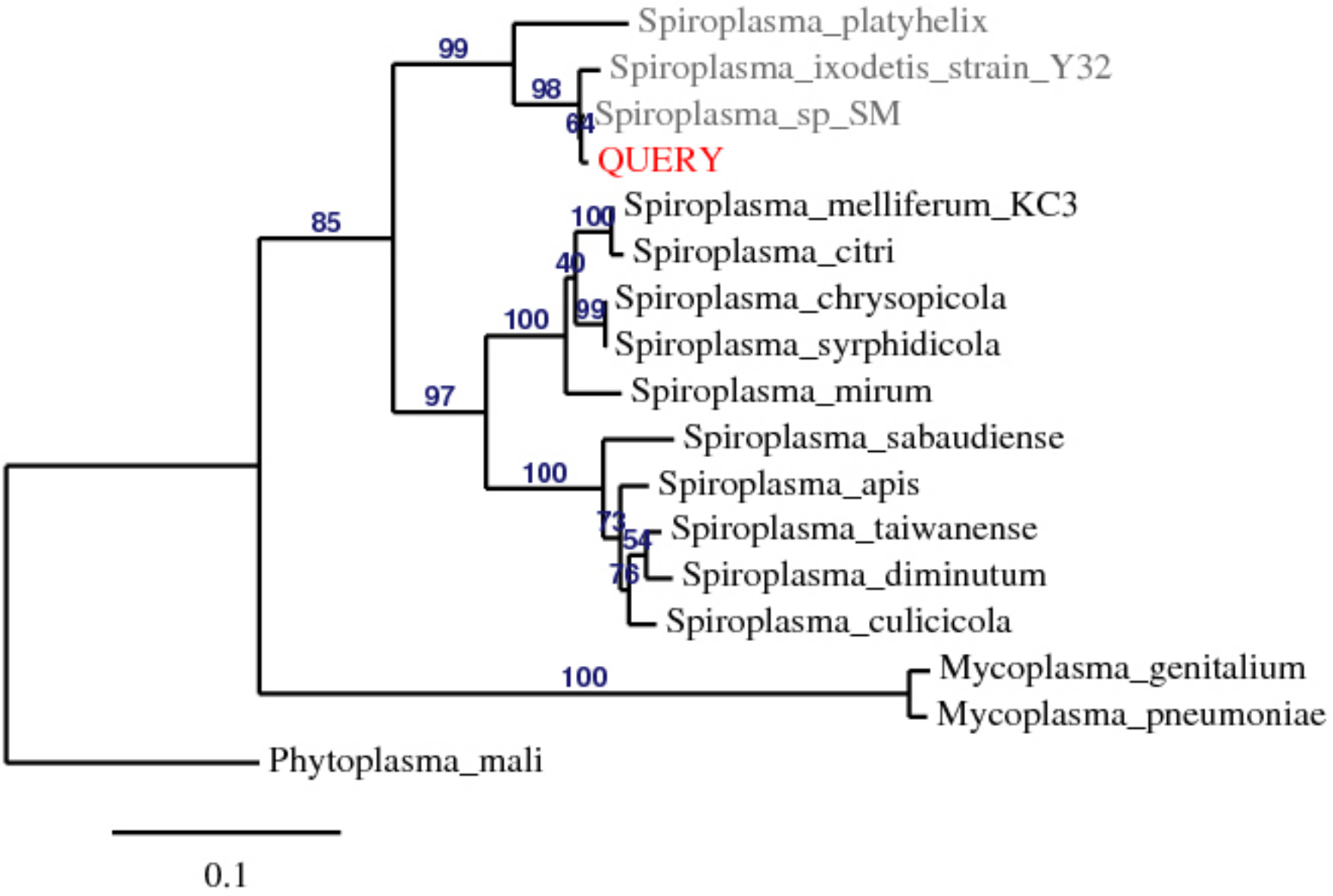


Analysis of contigs >1 kb in terms of blast matches and read coverage.

=> 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.

Three biotypes contained a sizeable proportion of sequences with a putative symbiotic origin:


Rickettsia_canadensis
77
Rickettsia_bellii
84
QUERY
73
Rickettsia_akari
76
Rickettsia_australis
85
Rickettsia_felis
100
Rickettsia_massiliae
Rickettsia_endosymbiont_of_Ixodes_scapularis
Candidatus_Rickettsia_amblyommii
Rickettsia_rhipicephali
Rickettsia_montanensis
Rickettsia_japonica
95
Rickettsia_heilongjiangensis
Rickettsia_peacockii
76
81
Rickettsia_rickettsii
78
Rickettsia_philipii
76
Rickettsia_honei
Rickettsia_slovaca
Rickettsia_africae
83
Rickettsia_conorii
Rickettsia_parkeri
Rickettsia_sibirica
Rickettsia_prowazekii
100
Rickettsia_typhi
0.007


Spiroplasma_platyhelix
Spiroplasma_ixodetis_strain_Y32
Spiroplasma_sp_SM
QUERY
Spiroplasma_melliferum_KC3
Spiroplasma_citri
Spiroplasma_chrysopicola
Spiroplasma_syrphidicola
Spiroplasma_mirum
Spiroplasma_sabaudiense
Spiroplasma_apis
Spiroplasma_taiwanense
Spiroplasma_diminutum
Spiroplasma_culicicola
Mycoplasma_genitalium
Mycoplasma_pneumoniae
Phytoplasma_mali
99
98
64
85
100
40
99
100
97
100
73
54
76
100
0.1

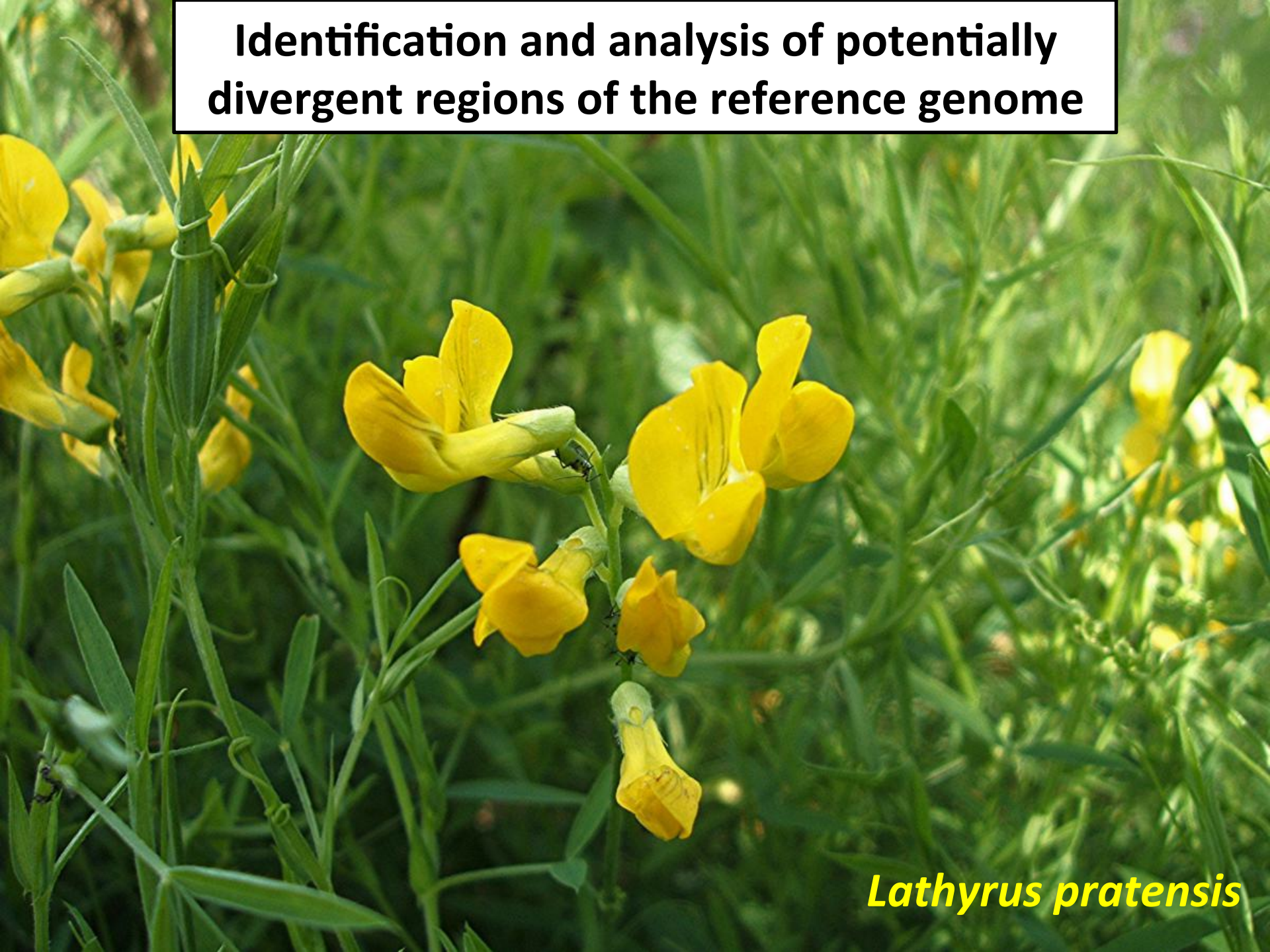
De novo assembly and characterization of an aphid symbiont genome

- 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.
- 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)
- Assambler choice SPAdes
- The final assembly contained 509 contigs >500 bp (2442 bp on average), totaling 1.2 Mb of sequence.

Sequences of nuclear origin (*A. pisum*)**:**

- All biotypes possessed contigs with a putative nuclear origin. Some of these contigs were similar between several biotypes or even between all biotypes.
- Contigs were clustered the 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.

Identification and analysis of potentially divergent regions of the reference genome



Lathyrus pratensis

L. pratensis biotype was particularly enriched in sequences with a putative nuclear origin:

- 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)

= > assembled from reads that were too divergent to map in the first place.

- **1137 (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 (377 regions summing to 337 kb), and 'normal-to-high-coverage' regions (760 regions, 663 kb).

Conclusions:

- The **direct pairwise comparisons of read sets**, before assembly, enabled the **rapid identification of similar read sets** and highlighted atypical samples and biotypes.
- For **contigs of symbiont origin**, this revealed notably the **misspecification 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.

Conclusions:

- This analysis **allowed 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.

Future prospects:

- 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.

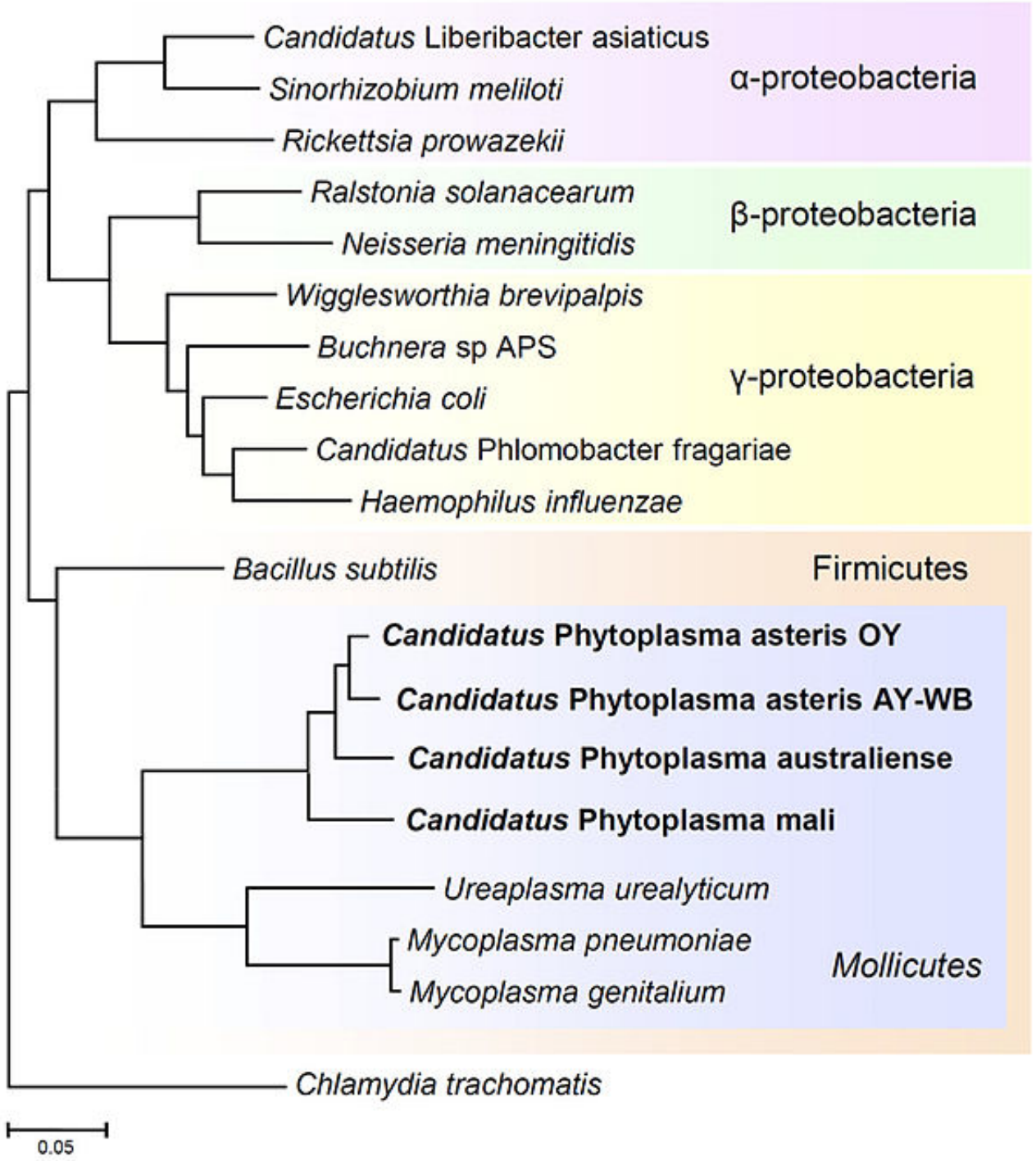
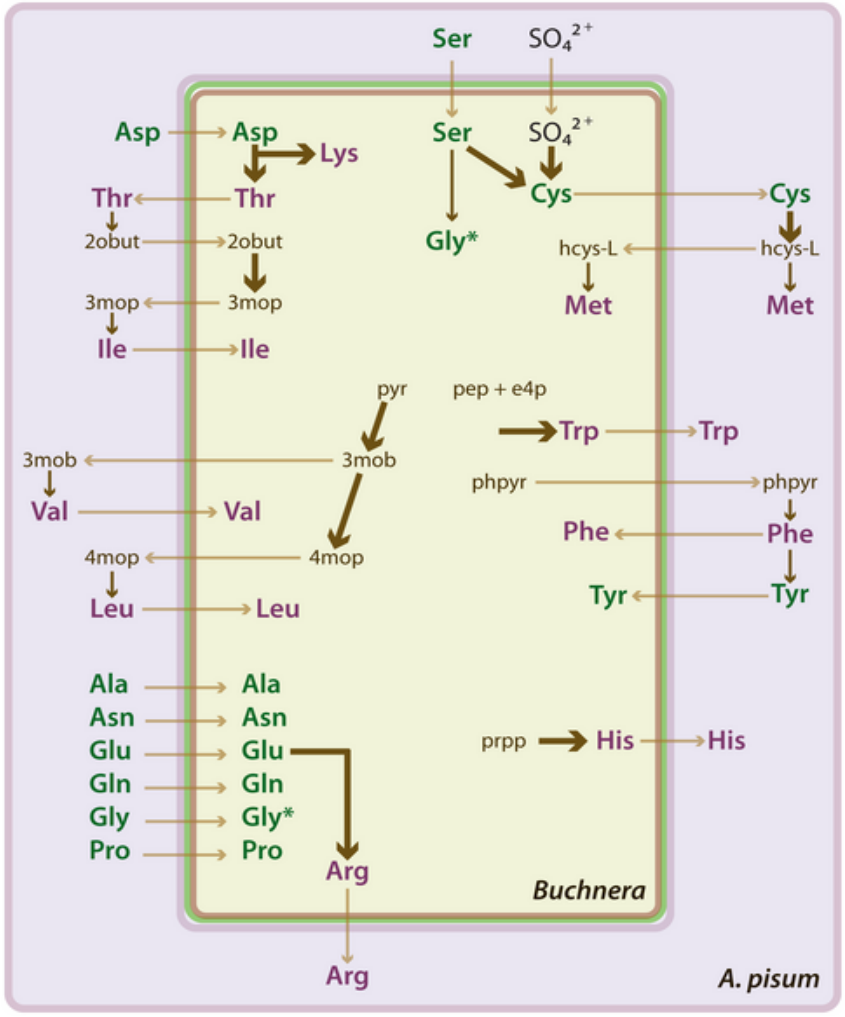

Candidatus Liberibacter asiaticus
Sinorhizobium meliloti
Rickettsia prowazekii
α-proteobacteria
Ralstonia solanacearum
Neisseria meningitidis
β-proteobacteria
Wigglesworthia brevipalpis
Buchnera sp APS
Escherichia coli
Candidatus Phlomobacter fragariae
Haemophilus influenzae
γ-proteobacteria
Bacillus subtilis
Firmicutes
Candidatus Phytoplasma asteris OY
Candidatus Phytoplasma asteris AY-WB
Candidatus Phytoplasma australiense
Candidatus Phytoplasma mali
Ureaplasma urealyticum
Mycoplasma pneumoniae
Mycoplasma genitalium
Mollicutes
Chlamydia trachomatis
0.05

Figure 9. Amino acid relations of the pea aphid Acyrthosiphon pisum and its symbiotic bacterium Buchnera aphidicola.



The International Aphid Genomics Consortium (2010) Genome Sequence of the Pea Aphid Acyrthosiphon pisum. PLoS Biol 8(2): e1000313. doi: 10.1371/journal.pbio.1000313
http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1000313