Authors: Draia-Nicolau Ondina & Murray Madeleine M2 Bio-informatique DLAD 2019-2020

Into the genome of Candidatus Blochmannia pennsylvanicus

I – Introduction

Nowadays, next generation sequencing of multiple organisms provides a better understanding of functions, structure and interactions of genes in order to identify those that are unique and those that are conserved among the species. A set of programs and algorithms are also available, allowing the users to obtain this complex data.

We have chosen to work on the genome of an organism called *Candidatus Blochmannia Pennsylvanicus*. The other organisms that are from the same family and enabled us to make the comparative analysis are*: Candidatus Blochmannia chromaiodes str. 640*, *Candidatus Blochmannia Floridanus*, *Enterobacter hormaechei* and *Klebsiella Pneumoniae*.

Candidatus Blochmannia pennsylvanicus, and more precisely, the BPEN strain is an enterobacteria and an endosymbiont, as is *Candidatus Blochmannia chromaiodes* and *Candidatus Blochmannia floridanus*. In opposition, *Enterobacter hormaechei* is an enterobacteria but it is not an endosymbiont. *Klebsiella pneumoniae* is also an enterobacteria without being an endosymbiont. All species are from the same family of **Enterobacteriaceae** but only *Candidatus Blochmannia pennsylvanicus*, *Candidatus Blochmannia chromaiodes* and *Blochmannia floridanus* belong to the same genus. This information is important for the further analysis of those genomes.

Several programs, scripts and methods were deployed to study the genome of *Candidatus Blochmannia pennsylvanicus* and to compare it to the other four genomes of its family. We started studying the basic characteristics of the genomes (number of genes, the average length of those genes, the gene density, the cumulative frequency of dinucleotides, trinucleotides and the RSCU) by using some basic python or pearl scripts or an online tool (CAI calculator).

Then, the MAUVE program added another level of genomic comparison as at aligns multiple genomes to visualise regions of synteny. BLAST analysis was also used in order to realise a functional comparative analysis of those genomes. It also allowed us to check and visualise the core and the pan genomes. However, the main functional annotation of orthologs was based on eggNOG mapper using databanks of orthologous groups.

All those methods allowed us to make different hypothesis and conclusions on the evolutionary history *of Candidatus Blochmannia Pennsylvanicus*.

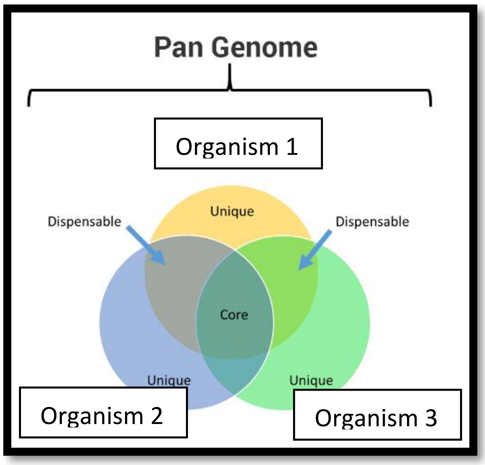


Figure explaining the Core and Pan genome

II – Methods

Genomic signature

Number of genes, gene mean length and gene density were calculated with “NbGen_AvgLen_GenDen.py” to obtain a general idea of the genomic signatures of the five organisms (Table 1).

Evolutionary genealogy of genes: Non-supervised Orthologous Groups

EggNOG was first used to find the orthologous groups in the genome of interest. Both eggNOG diamond and hmmer were run and later compared with a python script. EggNOG diamond uses a protein database to find orthologs while eggNOG hmmer relies on Hidden Markov Model. The python script finds genes that differ between the two methods, as well as the most common orthologous groups found with each method.

Dinucleotide and codon frequencies

A pearl script was created to compute the Dinucleotide frequency ratio between the observed data and the expected data. Also, an online tool was used and enabled us to compute the cumulative frequency of the codon usage and the RSCU (Relative Synonymous Codon Usage). Complete data is available here (under the “Tableaux section”, 324 rows, too long to display in report, called “GeCo_Murray_Draia-Nicolau_2019.xlsx”).

Multiple alignment of conserved genomic sequences

MAUVE is a program that studies the evolution of organisms by aligning their sequence and highlighting their synthons. In this paper, *Candidatus Blochmannia pennsylvanicus* was always kept as the reference genome. Of course, as the 5 genomes are part of the same phylogenetic family, high similarity between some regions is expected.

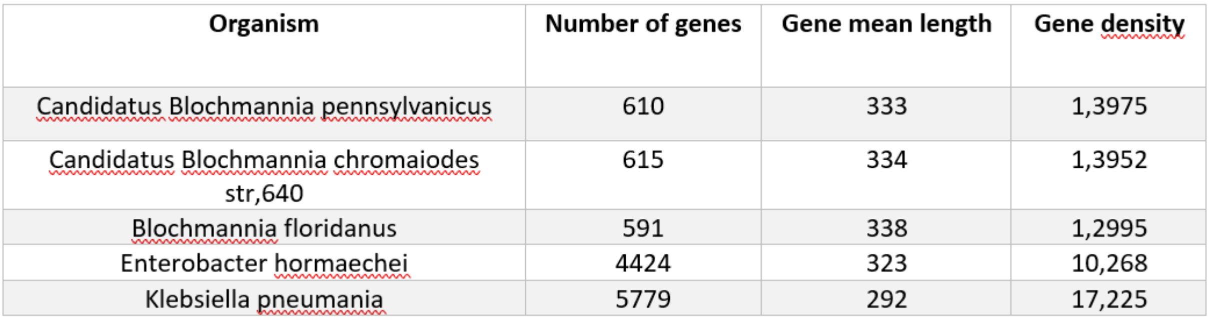
Core and pan genomes

Two scripts were written to create the pan and core genomes. The first script "blastscript.sh" performs protein BLAST (Basic Local Alignment Search Tool) against all species both ways. Only sequences with over 80% coverage length and 30% identity were considered as potential homologs (and only the best hit per sequence was kept). Then the second script "coregenome.sh" filters for bidirectional hits only and counts the number of genes in the pan and core genomes.

III – Results

EggNOG diamond and eggNOG hmmer were run to search for orthologous groups in the genome of interest. 603 orthologs were found with eggNOG diamond compared to 600 with eggNOG hmmer. The main orthologous groups identified by both methods are also very similar with GO:0005575, GO:0005623, GO:0005886, GO:0016020, GO:0044464 and GO:0071944, respectively cellular component, cell, plasma membrane, membrane, obsolete cell part and cell periphery. The main orthologous groups found are therefore associated with cellular components essential for cell survival.

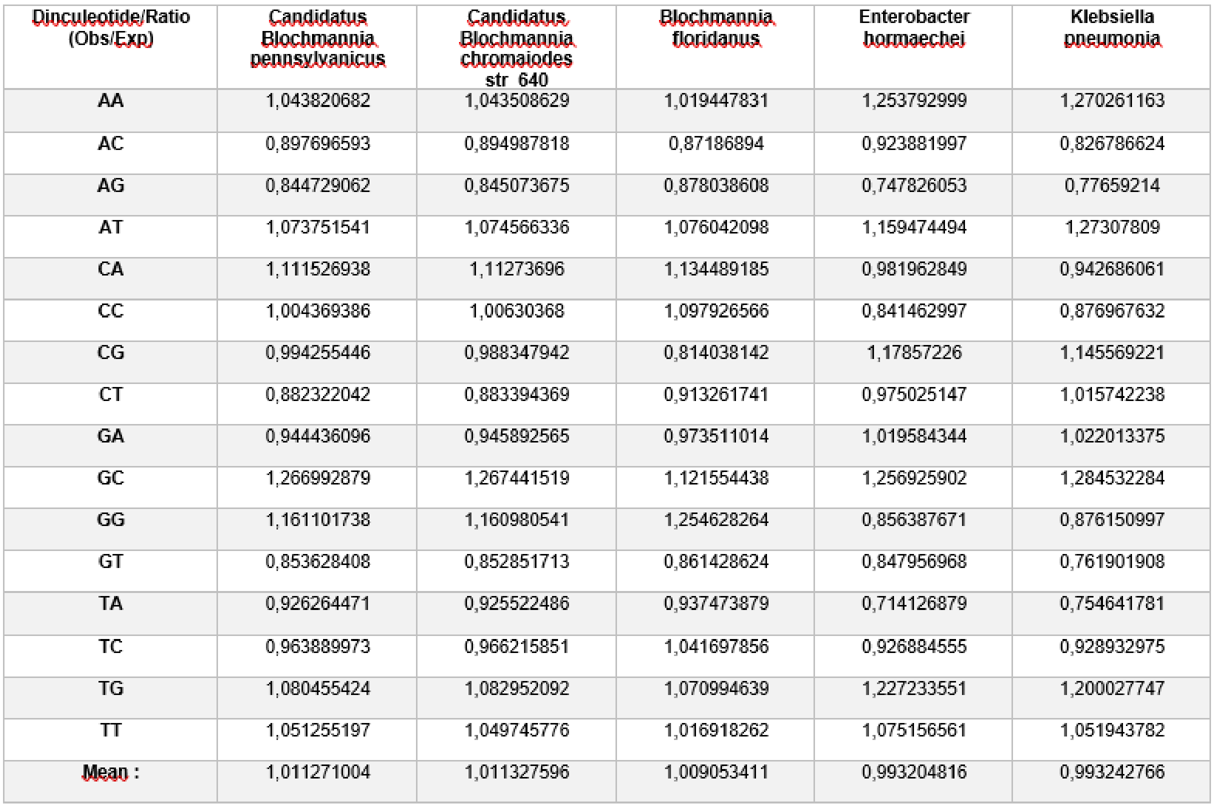
To better comprehend *Candidatus Blochmannia pennsylvanicus*, four other genomes were chosen for comparison. All five genomes are more or less related belonging to the same family. *Enterobacter hormaechei* and *Klebsiella pneumonia* possess a lot more genes than *Candidatus Blochmannia pennsylvanicus, Candidatus Blochmannia chromaiodes* and *Blochmannia floridanus* increasing their gene density 10-fold. However, gene mean length is constant across the five genomes ranging from 292 to 338 base pairs (Table 1).



Organism	Number of genes	Gene mean length	Gene density
Candidatus Blochmannia pennsylvanicus	610	333	1,3975
Candidatus Blochmannia chromaiodes str,640	615	334	1,3952
Blochmannia floridanus	591	338	1,2995
Enterobacter hormaechei	4424	323	10,268
Klebsiella pneumania	5779	292	17,225

Table 1: Number of genes, Average gene length and gene density of the 5 genomes

Genomes can start being compared from estimating dinucleotide frequency (Table 2). Results are also represented as a graph in Figure 1.



Dinculeotide/Ratio (Obs/Exp)	Candidatus Blochmannia pennsylvanicus	Candidatus Blochmannia chromaiodes str 640	Blochmannia floridanus	Enterobacter hormaechei	Klebsiella pneumonia
AA	1,043820682	1,043508629	1,019447831	1,253792999	1,270261163
AC	0,897696593	0,894987818	0,87186894	0,923881997	0,826786624
AG	0,844729062	0,845073675	0,878038608	0,747826053	0,77659214
AT	1,073751541	1,074566336	1,076042098	1,159474494	1,27307809
CA	1,111526938	1,11273696	1,134489185	0,981962849	0,942686061
CC	1,004369386	1,00630368	1,097926566	0,841462997	0,876967632
CG	0,994255446	0,988347942	0,814038142	1,17857226	1,145569221
CT	0,882322042	0,883394369	0,913261741	0,975025147	1,015742238
GA	0,944436096	0,945892565	0,973511014	1,019584344	1,022013375
GC	1,266992879	1,267441519	1,121554438	1,256925902	1,284532284
GG	1,161101738	1,160980541	1,254628264	0,856387671	0,876150997
GT	0,853628408	0,852851713	0,861428624	0,847956968	0,761901908
TA	0,926264471	0,925522486	0,937473879	0,714126879	0,754641781
TC	0,963889973	0,966215851	1,041697856	0,926884555	0,928932975
TG	1,080455424	1,082952092	1,070994639	1,227233551	1,200027747
TT	1,051255197	1,049745776	1,016918262	1,075156561	1,051943782
Mean :	1,011271004	1,011327596	1,009053411	0,993204816	0,993242766

Table 2: Dinucleotide frequency Ration Observed/Expected of the 5 genomes

A few dinucleotides have a lower frequency than expected by change alone for all five genomes: AC, AG, GT and TA. On the other hand, some dinucleotides have a higher frequency than expected: AA, AT, GC, TG and TT.

Interestingly, the large genomes of *Klebsiella pneumonia and Enterobacter hormaechei* behave in similar ways compared to the smaller genomes of the other three organisms. They possess higher frequencies of AA, AT, CG and TG, but lower frequencies of CA, CC, GG and TA. Apart from GG and TA, those dinucleotides have frequencies of about 1 for the three smaller genomes, suggesting that frequencies of AA, AT, CG, TG, CA and CC are impacted by genome size and gene density. Furthermore, CA, CC, GG and TA may be specific to gene coding regions as their frequencies decrease in low gene density organisms.

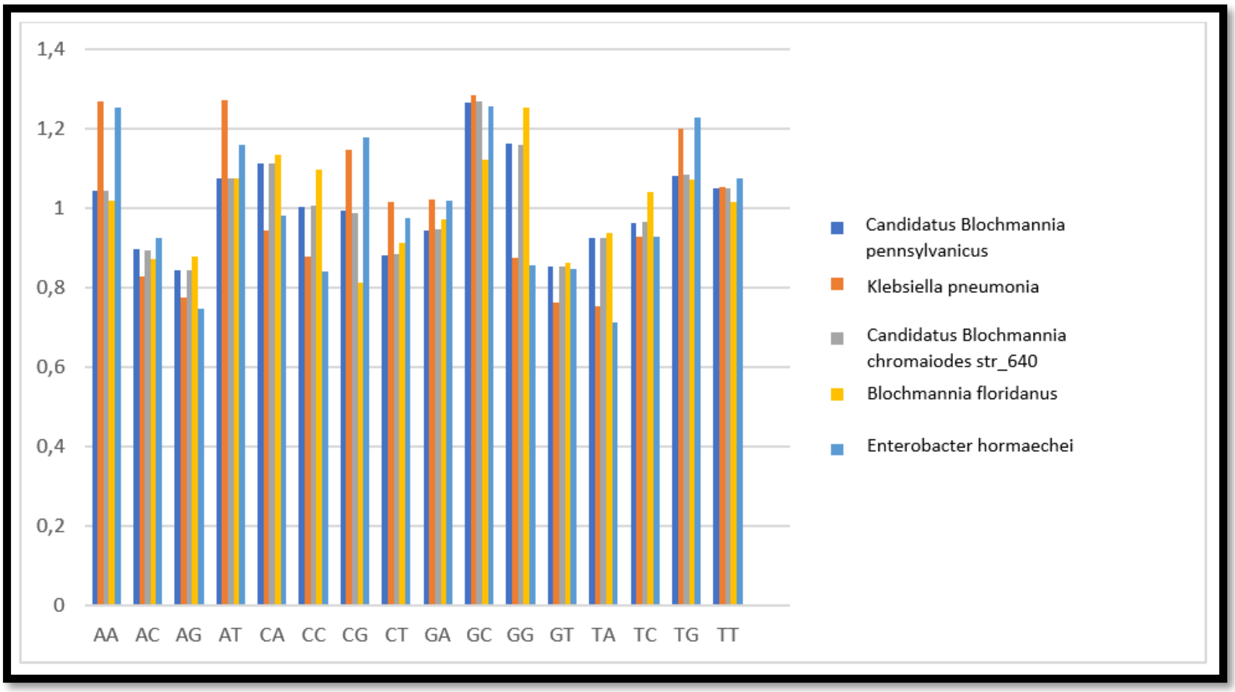


Figure 1: Graph of the Dinucleotide frequency ratio (Obs/Expected)

As proteins are coded from trinucleotides commonly known as codons, the frequency of codon usage and the Relative Synonymous Codon Usage (RSCU) can be visualized on Figures 2 and 3 respectively. For some reason, the program could not use the sequence of Enterobacter hormaechei.
According to Figure 2, TGG which translates for Tryptophan (represented in deep blue) is the rarest codon within all five genomes. The genomes appear to follow similar trends in codon frequency except for *Klebsiella pneumoniae*. Indeed, the latter possesses a large number of AGT (Serine), GAT (Aspartic acid), CTG (Leucine) in comparison with the four others for example.

Candidatus Blochmannia pennsylvanicus, Candidatus Blochmannia chromaiodes and *Candidatus Blochmannia floridanus* were indeed expected to be quite similar as they belong to the same genus.

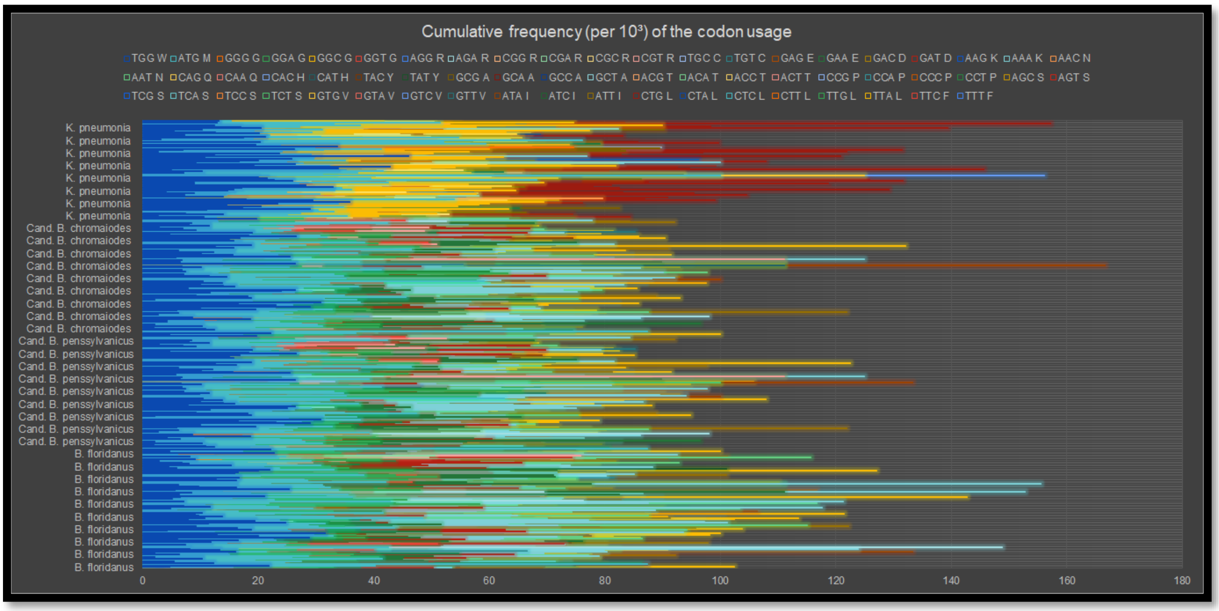


Figure 2: Cumulative frequency of the codon usage

The genetic code is degenerative allowing amino acids to be coded by multiple codons (there are 64 codons for only 20 amino acids). The **RSCU** computes codon usage and can be used to assess codon preference of an organism for an amino acid. On Figure 3, glycine for example can be coded by either of the following: GGG (orange), GGA (green), GGC (yellow) or GGT (light red). GGA is the preferred codon for glycine in all organism but *Klebsiella pneumoniae*, who once again stands out preferring GGT. Overall, there is a clear separation in colour between *Klebsiella pneumoniae* and the other three organisms, with yellow (GGG, GGC and TTA) instead of green (TTG and TCT). In addition, there is a small difference between *Candidatus Blochmannia floridanus* and the other two organisms of the genus, with more lysine than phenylalanine.

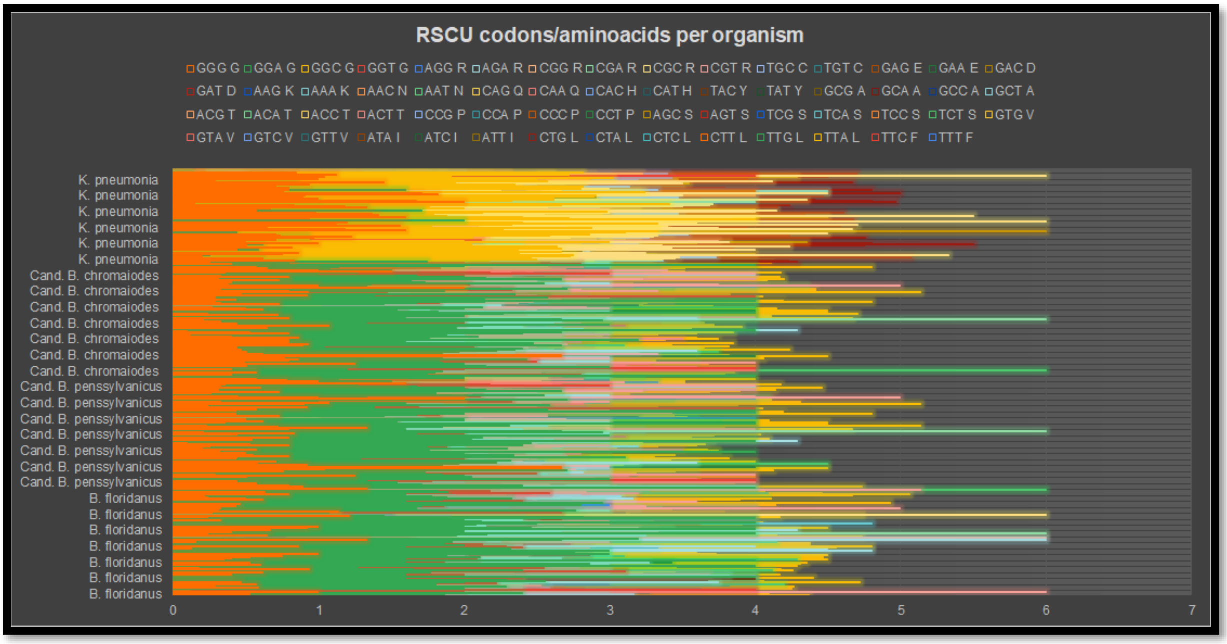


Figure 3: RSCU codons/amino acids per organism

From results above, it is expected that *Candidatus Blochmannia pennsylvanicus*, *Candidatus Blochmannia chromaiodes* and *Blochmannia floridanus* would align relatively well. Mauve confirmed conserved synteny between *Candidatus Blochmannia pennsylvanicus* and *Blochmannia floridanus* (Figure 4b), as well as between *Candidatus Blochmannia pennsylvanicus* and *Candidatus Blochmannia chromaiodes* (Figure 4c). Multiple conserved regions were also found with *Enterobacter hormaechei* (Figure 4d) and *Klebsiella pneumonia* (Figure 4e), but the order of the synthons was not conserved as it was the case with the others. This may be explained by their larger genomes that tolerate genomic rearrangement.

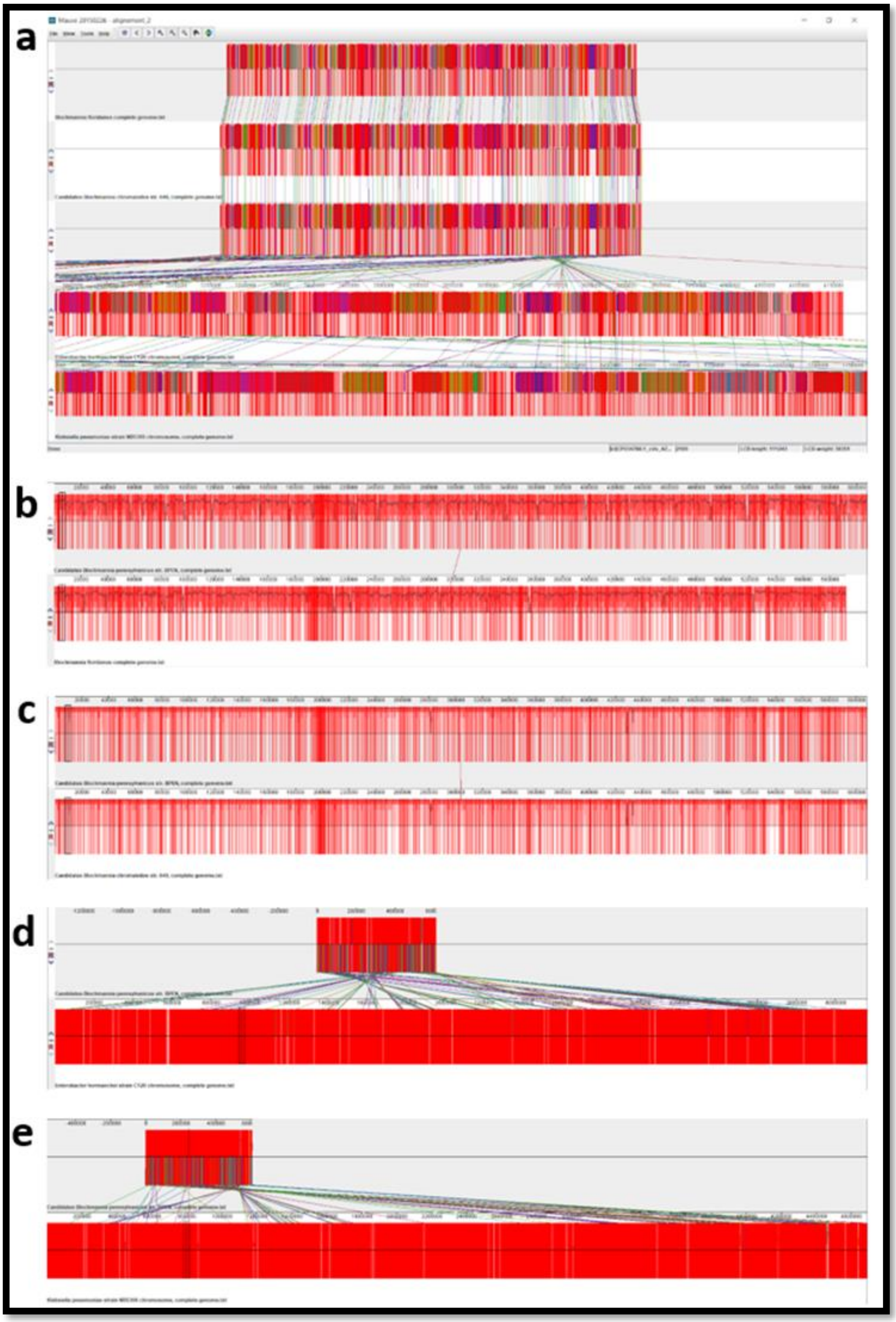


Figure 4: Mauve synteny analysis of the 5 genomes
(a) overview of the five genomes using *Candidatus Blochmannia pennsylvanicus* as a reference
(b) Alignment of *Candidatus Blochmannia pennsylvanicus* and *Blochmannia floridanus*
(c) Alignment of *Candidatus Blochmannia pennsylvanicus* and *Candidatus Blochmannia chromaiodes*
(d) Alignment of *Candidatus Blochmannia pennsylvanicus* and *Enterobacter hormaechei*
(e) Alignment of *Candidatus Blochmannia pennsylvanicus* and *Klebsiella pneumonia*

Using BLAST, one can also align two genomes to found homologs between two organisms. The number of homologs were saved to draw the pan and core genome curves displayed in Figure 5. All core genomes decrease from adding new genomes while pan genomes increase as expected. Interestingly, the pan genomes of *Enterobacter hormaechei* and *Klebsiella pneumoniae* quickly grow when the genes of the other species are added.

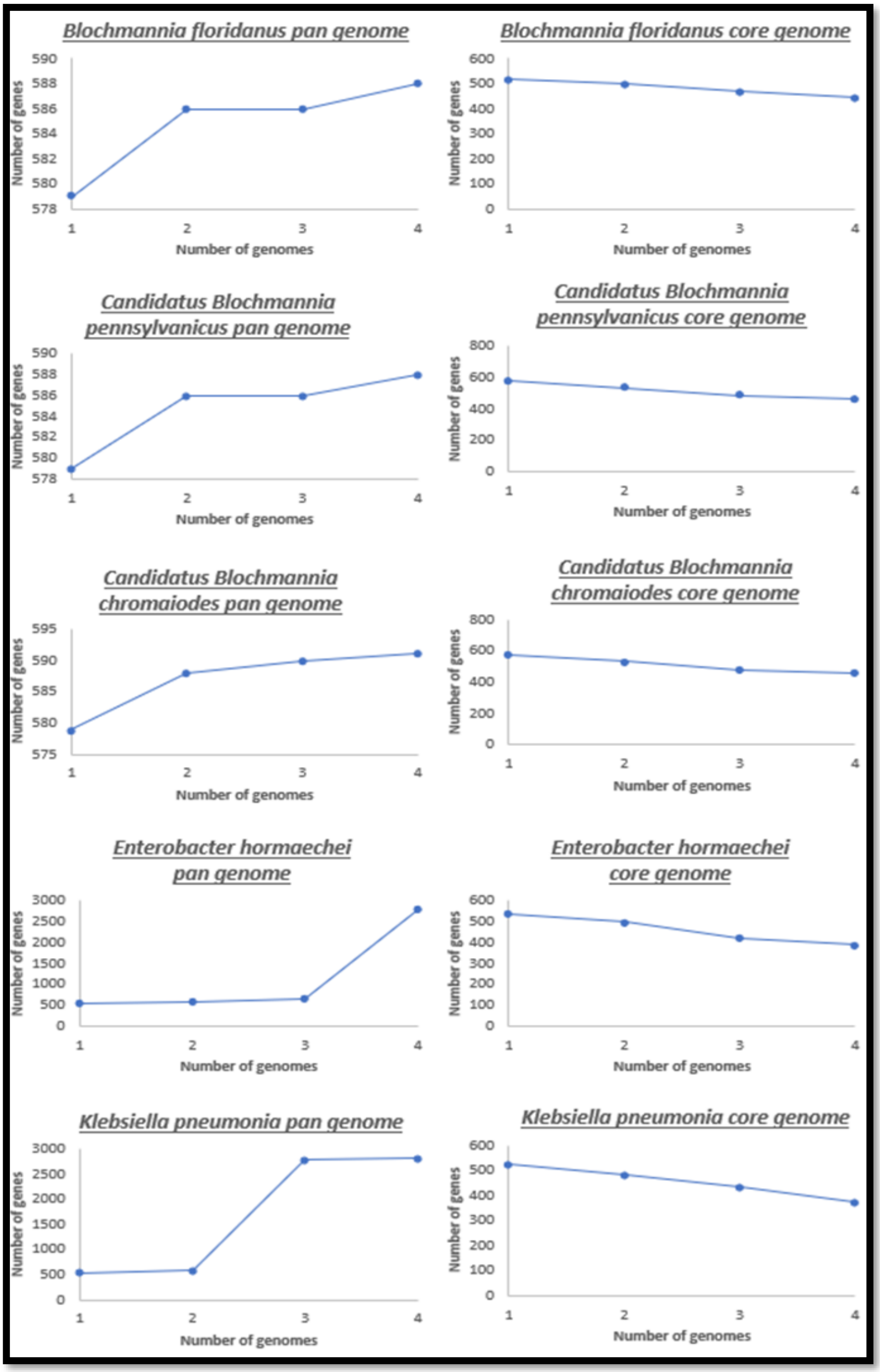


Figure 5: Pan and core genome analysis of the 5 genomes

IV – Conclusion

The analysis of those five organisms suggest the genomes of *Enterobacter hormaechei* and *Klebsiella pneumoniae* accumulated genetic material throughout their evolutionary history, allowing genomic and chromosomic rearrangement compared to the other members of their family. Although their gene density was lowered, they retain similar functions to the other organisms as many homologs were found in the core genome with BLAST. Interestingly, *Klebsiella pneumonia* and (and also, *Enterobacter hormaechei*) stands out from the other four from its codon usage and codon preferences, which is normal as they come from a different species of the same family as the reference genome, *Candidatus Blochmannia Pennsylvanicus str.BPEN*.

The other two organisms, *Candidatus Blochmannia floridanus* and *Candidatus Blochmannia chromaiodes str.640* are quite similar to the reference genome, suggesting that they didn't diverge much as they are the endosymbiont of the same organism so they live in almost the same environnement.

- Supplementary data online availability:
  Scripts and raw data are publicly available on github.

References

(n.d.). Retrieved from EGGNOG Mapper: http://eggnogdb.embl.de/#/app/emapper

(n.d.). Retrieved from EBI QuickGo: https://www.ebi.ac.uk/QuickGO/

(n.d.). Retrieved from CAI Calculator: http://genomes.urv.es/CAIcal/

What is a pangenome? (n.d.). Retrieved from http://www.10wheatgenomes.com/what-is-a-pan-genome/