Relation between Insertion Sequences and Genome Rearrangements in *Pseudomonas aeruginosa*

Huda Al-Nayyef[1,3], Christophe Guyeux[1], Marie Petitjean[2], Didier Hocquet[2] and Jacques M. Bahi[1]

[1] FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department Université de Franche-Comté, 16, route de Gray, 25000 Besançon, France.
[2] Laboratoire d'Hygiène Hospitalière, UMR 6249 CNRS Chrono-environnement, Université de Franche-Comté, France.
[3] Computer Science Department, University of Mustansiriyah, Iraq.
{huda.al-nayyef, christophe.guyeux, marie.petitjean, jacques.bahi}@univ-fcomte.fr
dhocquet@chu-besancon.fr

Abstract. During evolution of microorganisms genomes underwork have different changes in their lengths, gene orders, and gene contents. Investigating these structural rearrangements helps to understand how genomes have been modified over time. Some elements that play an important role in genome rearrangements are called insertion sequences (ISs), they are the simplest types of transposable elements (TEs) that widely spread within prokaryotic genomes. ISs can be defined as DNA segments that have the ability to move (cut and paste) themselves to another location within the same chromosome or not. Due to their ability to move around, they are often presented as responsible of some of these genomic recombination. Authors of this research work have regarded this claim, by checking if a relation between insertion sequences (ISs) and genome rearrangements can be found. To achieve this goal, a new pipeline that combines various tools has firstly been designed, for detecting the distribution of ORFs that belongs to each IS category. Secondly, links between these predicted ISs and observed rearrangements of two close genomes have been investigated, by seeing them with the naked eye, and by using computational approaches. The proposal has been tested on 18 complete bacterial genomes of *Pseudomonas aeruginosa*, leading to the conclusion that IS3 family of insertion sequences are related to genomic inversions.

Keywords: Rearrangements, Inversions, Insertion Sequences, Pseudomonas aeruginosa

1 Introduction

The study of genome rearrangements in microorganisms has become very important in computational biology and bio-informatics fields, owing to its applications in the evolution measurement of difference between species [1]. Important

elements in understanding genome rearrangements during evolution are called transposable elements, which are DNA fragments or segments that have the ability to insert themselves into new chromosomal locations, and often make duplicate copies of themselves during transposition process [2]. Indeed, within bacterial species, only cut-and-paste of transposition mechanism can be found, the transposable elements involved in such way being the insertion sequences. These types of mobile genetic elements (MGEs) seem to play an essential role in genomes rearrangements and evolution of prokaryotic genomes [3, 4].

Table 1: P. aeruginosa isolates used in this study.

Isolates	NCBI accession number	Number of genes
19BR	485462089	6218
213BR	485462091	6184
B136-33	478476202	5818
c7447m	543873856	5689
DK2	392981410	5871
LES431	566561164	6006
LESB58	218888746	6059
M18	386056071	5771
MTB-1	564949884	6000
NCGM2.S1	386062973	6226
PA1	558672313	5981
PA7	150958624	6031
PACS2	106896550	5928
PAO1	110645304	5681
RP73	514407635	5804
SCV20265	568306739	6190
UCBPP-PA14	116048575	5908
YL84	576902775	5856

In this research work, we questioned the relation between the movement of insertion sequences on the one hand, and genome rearrangements on the other hand, and tested whether the type of IS family influences this relation. Investigations will focus on inversion operations of rearrangement (let us recall that an inversion occurs within genomes when a chromosome breaks at two points, and when the segment flanked with these breakpoints is inserted again but in reversed order, this event being potentially mediated with molecular mechanisms [5, 6]). To achieve our goal, we built a pipeline system module that combines existing tools together with the development of new ones, for finding putative ISs and inversions within studied genomes. We will then use this system to investigate the structure of prokaryotic genomes, by searching for IS elements at the boundaries of each inversion.

The contributions of this article can be summarized as follows. (1) A pipeline for insertion sequences discovery and classification is proposed. It uses unannotated genomes and then combines different existing tools for ORF predictions and clustering. It also classifies them according to an international IS database specific to bacteria. Involved tools in this stage are, among others, Prodigal [7], Markov Cluster Process (MCL) [8], and ISFinder[1] [9]. (2) We then use two dif-

[1] www-is.biotoul.fr

ferent strategies to check the relation between ISs and genomic rearrangements. The first one used a well-supported phylogenetic tree, then genomes of close isolates are drawn together, while the questioned relation is checked with naked eye. In the second strategy, inversion cases are thoroughly investigated with ad hoc computer programs. And (3), the pipeline is tested on the set of 18 complete genomes of *Pseudomonas aeruginosa* provided in Table 1. After having checked left and right inversion boundaries according to different window sizes, the probability of appearance of each type of IS family is then provided, and biological consequences are finally outlined.

The remainder of this article is organized as follows. The proposed pipeline for detecting insertion sequences in a list of ORFs extracted from unannotated genomes is detailed in Section 2. Rearrangements found using drawn genomes of close isolates is detailed in Section 3, while a computational method for discovering inversions within all 18 completed genomes of *P. aeruginosa* and results are provided in Section 3.2. This article ends by a conclusion section, in which the contributions are summarized and intended future work is detailed.

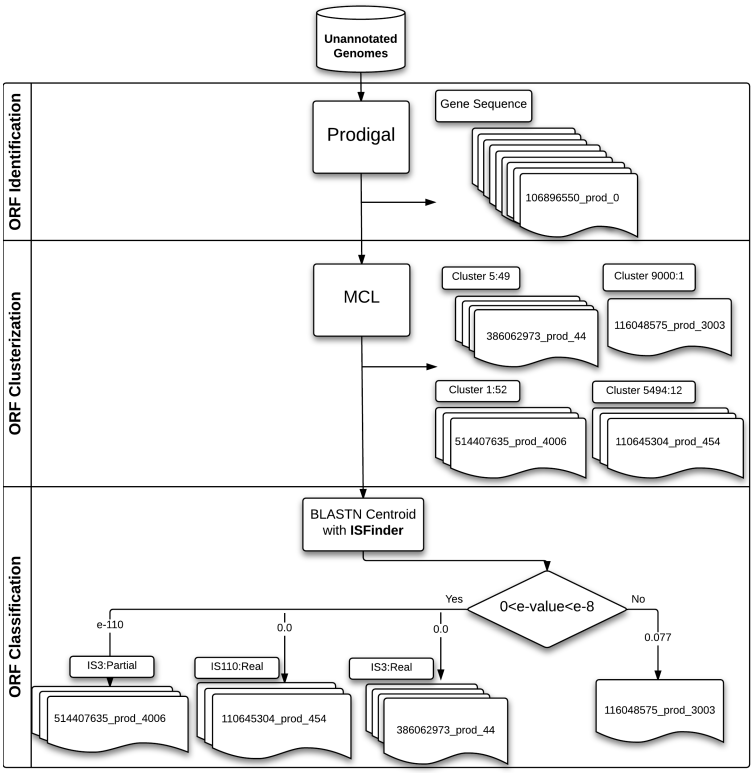


Fig. 1: Pipeline for detecting IS clusters in genomes of *P.aeruginosa*

2 Methodology for IS detection

In a previous work [10], we have constructed a pipeline system that combines three annotation tools (BASys [11], Prokka [12], and Prodigal [7]) with OASIS [13], that detected IS elements within prokaryotic genomes. This pipeline produces various information about each predicted IS, each IS is bordered by an *Inverted-Repeat* (IR) sequence, number of ORFs in each IS family and group, etc. As we are now only interested in detecting which ORFs are insertion sequences, we have developed a new lightweight pipeline that focuses on such open reading frames.. This pipeline, depicted in Figure 1, relies on ISFinder database [9], the up-to-date reference for bacterial insertion sequences. The main function of this database is to assign IS names and to provide a focal point for a coherent nomenclature. This is also a repository for ISs that contains all information about insertion sequences such as family and group.

The proposed pipeline can be summarized as follows.

Step 1: ORF identification. Prodigal is used as annotation tool for predicting gene sequences. This tool is an accurate bacterial and archaeal gene finding software provided by the Oak Ridge National Laboratory [7]. Table 1 lists the number of the predicted genes in each genome.

Step 2: ORF clustering. The Markov Cluster Process (MCL) algorithm is then used to achieve clustering of detected ORFs [8, 14].

Step 3: Clusters classification. The IS family and group of the centroid sequences of each cluster is determined with ISFinder database. BLASTN program is used here: if the e-value of the first hit is equal to 0, then the cluster of the associated sequence is called a "Real IS cluster". Otherwise, if the e-value is lower than 10^{-8}, the cluster is denoted as "Partial IS". At each time, family and group names of ISs that best match the considered sequence are assigned to the associated cluster. In Table 2 summarizes founded IS clusters found in the 18 genomes of *P. aeruginoza*.

Table 2: Summary of detected IS clusters

	No. of Clusters	Max. size of Cluster	Total no. of IS genes
Partial IS	94	57	362
Real IS	66	49	238
Total IS Cluster	160	-	600

3 Rearrangements in *Pseudomonas aeruginosa*

At the nucleotide level, genomes evolve with point mutations and small insertions and deletions [15], while at genes level, larger modifications including duplication, deletion, or inversion, of a single gene or of a large DNA segment, affect genomes by large scale rearrangements [16, 17]. The pipeline detailed previously

investigated the relations between insertion sequences and these genome rearrangements, by using two different methods that will be described below.

3.1 Naked eye investigations

In order to visualize the positions of IS elements involved in genomic recombination that have occurred in the considered set of *Pseudomonas*, we have first designed Python scripts that enable us to humanly visualize close genomes. Each complete genome has been annotated using the pipeline described in the previous section, and the strict core genome has been extracted. This latter is constituted by genes shared exactly once in each genome. Thus polymorphic nucleotides in these core genes have been extracted, and a phylogeny using maximum likelihood (RAxML [18, 19] with automatically detected mutation model) has been inferred. (Figure 2).

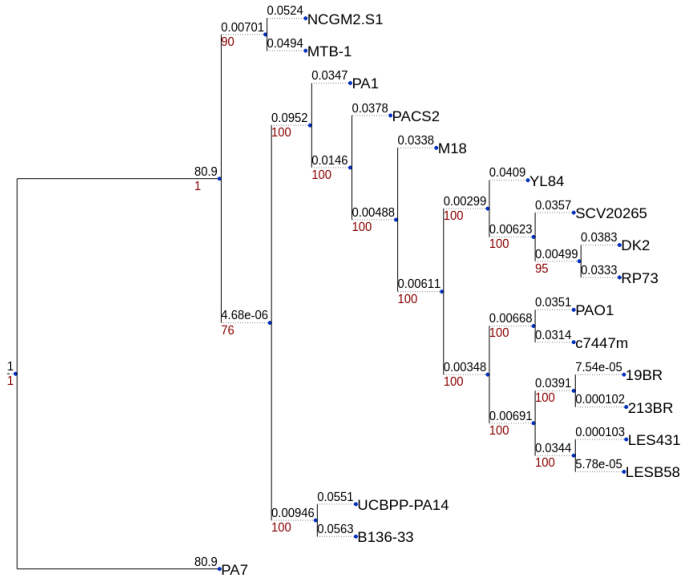
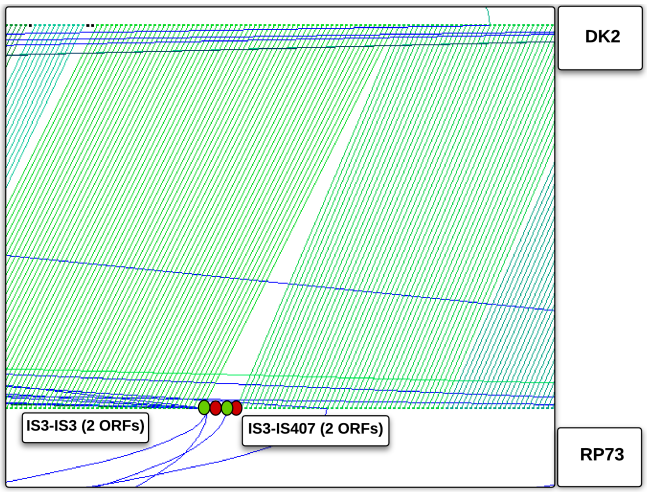
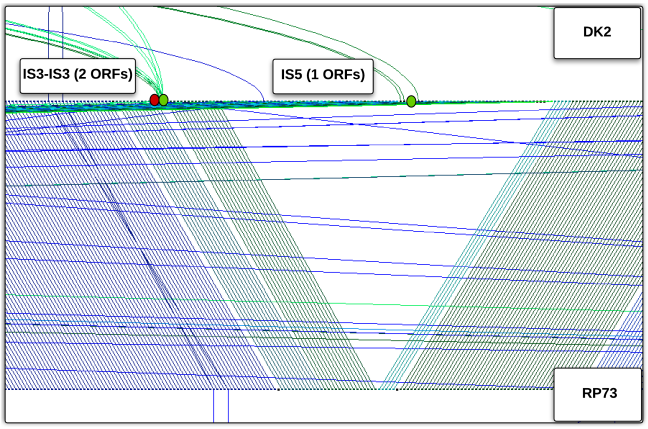


Fig. 2: Phylogeny of *P. aeruginosa* based on mutations on core genome

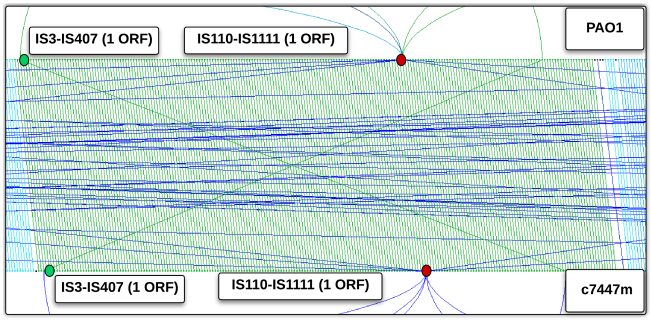
For each close isolates, a picture has then been produced using our designed Python script, for naked eye investigations. Real and Partial IS are represented with a red and green circles,respectively. Additionally, DNA sequences representing the same gene have been linked either with a curve (two same genes in the same isolate) or with a line (two same genes in two close isolates). Example of recombination events are given below.



(a) Insertion events of IS sequences have occurred in this set of 18 *P. aeruginosa* species. For instance, when comparing DK2 and RP73, we have found that IS3-IS3 (2 ORFs) and IS3-IS407 (2 ORFs too) have been inserted inside RP73.



(b) Deletions of insertion sequences can be found too, IS5 (Partial IS) is present in the genome of DK2, while it is deleted in the close isolates RP73.



(c) A duplication occurs in the insertion sequence type IS110-IS1111 that contains one ORF (Real IS), as there are 6 copies of this insertion sequence in both PAO1 and C7447m genomes.

Fig. 3: Examples of genomic recombination events: Insertion, Deletion, and Duplication.

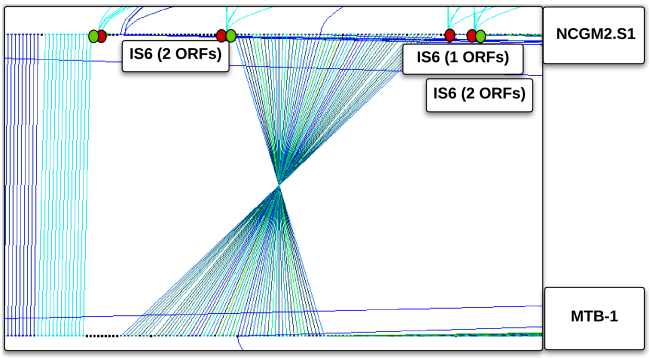


Fig. 4: The surrounding insertion sequences are within the same IS family (IS6) in the NCGM2.S1 genome. We have found too that insertion sequences are not always exactly at the beginning and end positions of the inversion, but they are overrepresented near these boundaries.

We will focus now on the link between large scale inversions and ISs as shown in Figure 4, by designing another pipeline that automatically investigate the inversions.

3.2 Automated investigations of inversions

The proposal is now to automatically extract all inversions that have occurred within the set of 18 genomes under consideration, and then to investigate their relation with predicted IS elements. The proposed pipeline is described in Figure 5.

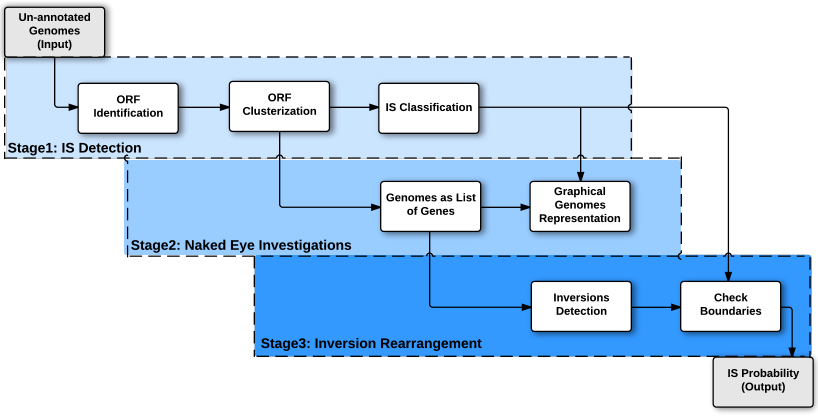
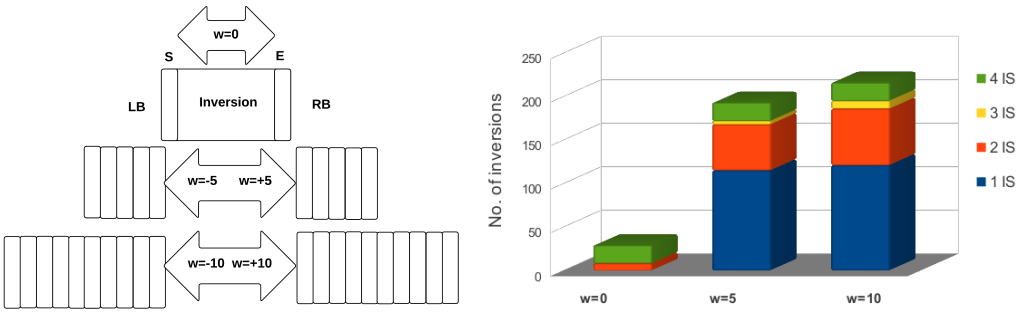


Fig. 5: Pipeline for detecting the role of ISs in inversions

Table 3: Small and large inversions detected from all genomes

Genome 1	Start	Stop	Genome 2	Start	Stop	Length (genes)
19BR	1001	1002	213BR	5094	5095	2
19BR	2907	2920	213BR	2933	2946	14
19BR	684	685	LES431	4689	4690	2
19BR	850	978	LES431	4393	4521	129
PAO1	997	998	c7447m	1977	1978	2
LESB58	4586	4587	LES431	2347	2348	2
DK2	2602	2603	RP73	2516	2517	2
DK2	1309	1558	RP73	3489	3738	250
DK2	260	261	SCV20265	3824	3825	2
DK2	2846	2852	SCV20265	3065	3071	7
M18	3590	3591	PACS2	1920	1921	2
M18	3194	3579	PACS2	2076	2461	386
MTB-1	5581	5582	B136-33	4742	4743	2
UCBPP-PA14	4820	4821	B136-33	2871	2872	2
NCGM2.S1	1053	1307	MTB-1	4507	4761	255
NCGM2.S1	1742	1743	MTB-1	4882	4883	2
PA1	95	96	B136-33	2691	2692	2
PA1	1334	1491	B136-33	1286	1443	158
PACS2	94	97	PA1	495	498	4
PACS2	970	1206	PA1	2220	2456	237
SCV20265	45	46	YL84	721	722	2
SCV20265	261	462	YL84	306	507	202
UCBPP-PA14	259	260	B136-33	3507	3508	2
YL84	721	722	M18	43	44	2
YL84	768	983	M18	5555	5770	216
YL84	721	722	PAO1	44	45	2
YL84	1095	1264	PAO1	5192	5361	170

– **Step1**: Convert genomes from the list of predicted coding sequences in the list of integer numbers, by considering the cluster number of each gene.
– **Step2**: Extract sets of inversion from all input genomes. 719 inversions have been found (see Table 3).
– **Step 3**: Extract IS clusters (Partial and Real IS) using the first pipeline, as presented in a previous section.



(a) Left and Right Boundary using window (b) No. of inversions for three different window size

Fig. 6: Using different window size within all inversions

– **Step 4**: Investigate boundaries of each inversion (starting S and ending E positions), by checking the presence of insertion sequences within a window ranging from $w = 0$ up to 10 genes. Between 0 and 4 insertion sequences have been found at the boundaries of each inversion, (Figure 6).

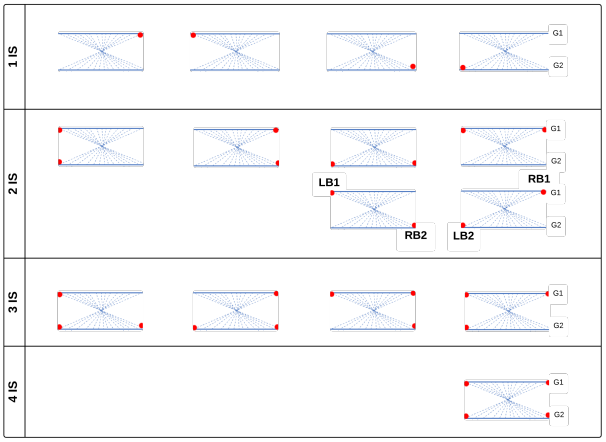


Fig. 7: Different cases of IS inversions

– **Step 5**: Finally, compute the presence probability for each IS families and groups near inversions. (Figure 7).

As presented in Figure 8, there is no major problem in dealing with small inversions because the small inversions having small ratio of increment as compared with big inversions (*i.e.*, during window size increment of inversion boundaries, the small inversions, which have length lower than 4 genes, have small increase ratios compared to large inversions).

Table 4 details the roles of IS in largest inversions found within two close isolates.

Table 4: Summary of large inversion sets within closed genomes

Genome 1	Genome 2	Inversions no.	Largest inversion	IS family	Boundary	Window
19BR	213BR	9	14	IS110	LB1-RB2	w=0
PAO1	c7447m	2	2	IS3	(LB1-RB2)/(LB2-RB1)	w=0
LES431	LESB58	3	2	IS3	(LB1-RB2)/(LB2-RB1)	w=0
DK2	RP73	93	250	Tn3	LB1-RB2	w=3
UCBPP-PA14	B136-33	7	2	IS3	(LB1-RB2)/(LB2-RB1)	w=0
NCGM2.S1	MTB-1	91	255	IS5	LB1-RB2	w=5

The IS family of type IS3 always have the most probability of appearance with left and right boundaries of inversions.(Figure 9)

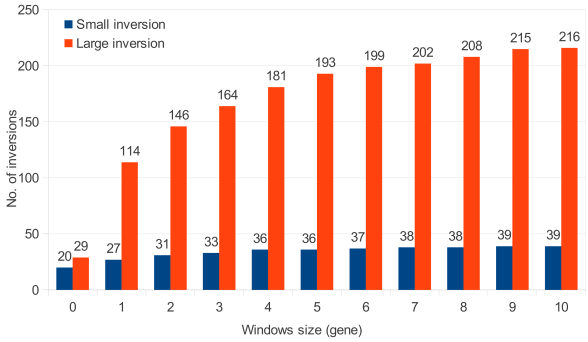
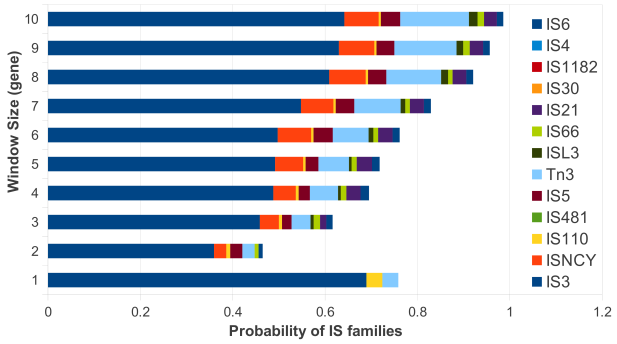
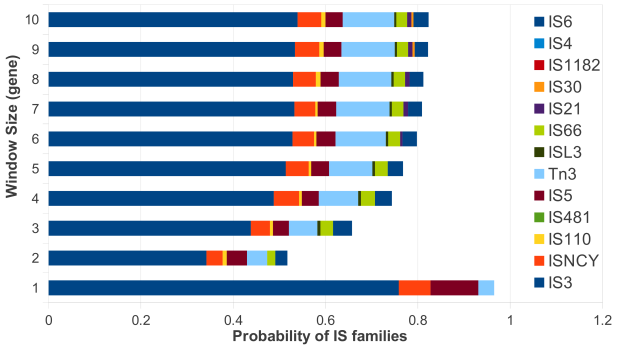


Fig. 8: Small inversions ($\leqslant 4$ genes) vs. large inversions (> 4 genes).



(a) Left Boundary



(b) Right Boundary.

Fig. 9: IS distribution using different windows size.

4 Conclusion

We designed a pipeline that detects and classify all ORFs that belong to IS. It has been done by merging various tools for ORF prediction, clusterization, and by finally using ISFinder database for classification.

This pipeline has been applied on a set of *Pseudomonas aeruginosa*, showing an obvious improvement in ORFs detection that belong to insertion sequences. Furthermore, relations between inversions and insertion sequences have been emphasized, leading to the conclusion that the so-called IS3 family has the largest probability of appearance inside inversion boundaries.

In future works, we intend to investigate more deeply the relation between ISs and other genomic recombination such as deletion and insertion. We will then focus on the implication of other types of genes like rRNA (rrnA, rrnB, rrnC, rrnD) in *P. aeruginosa* recombination [20]. By doing so, we will be able to determine genes that are often associated with deletion, inversion, etc. The pipeline will be finally extended to eukaryotic genomes and to other kinds of transposable elements.

References

1. Ying Chih Lin, Chin Lung Lu, Ying-Chuan Liu, and Chuan Yi Tang. Spring: a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic acids research*, 34(suppl 2):W696–W699, 2006.
2. Jennifer S Hawkins, HyeRan Kim, John D Nason, Rod A Wing, and Jonathan F Wendel. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in gossypium. *Genome research*, 16(10):1252–1261, 2006.
3. Patricia Siguier, Jonathan File, and Michael Chandler. Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology*, 9(5):526 – 531, 2006.
4. Casey M. Bergman and Hadi Quesneville. Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, 8(6):382–392, 2007.
5. Mark Kirkpatrick. How and why chromosome inversions evolve. *PLoS biology*, 8(9):e1000501, 2010.
6. José M Ranz, Damien Maurin, Yuk S Chan, Marcin Von Grotthuss, LaDeana W Hillier, John Roote, Michael Ashburner, and Casey M Bergman. Principles of genome evolution in the drosophila melanogaster species group. *PLoS biology*, 5(6):e152, 2007.
7. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):119, 2010.
8. Van Dongen and Stijn Marinus. Graph clustering by flow simulation. *University of Utrecht*, 2000.
9. Patricia Siguier, Jocelyne Pérochon, L Lestrade, Jacques Mahillon, and Michael Chandler. Isfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, 34(suppl 1):D32–D36, 2006.
10. Huda Al-Nayyef, Christophe Guyeux, and Jacques Bahi. A pipeline for insertion sequence detection and study for bacterial genome. *Lecture Notes in informatics (LNI)*, P-235:85–99, 2014.

11. Gary H Van Domselaar, Paul Stothard, Savita Shrivastava, Joseph A Cruz, AnChi Guo, Xiaoli Dong, Paul Lu, Duane Szafron, Russ Greiner, and David S Wishart. Basys: a web server for automated bacterial genome annotation. *Nucleic acids research*, 33(suppl 2):W455–W459, 2005.
12. T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, 2014.
13. David G Robinson, Ming-Chun Lee, and Christopher J Marx. Oasis: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic acids research*, 40(22):e174–e174, 2012.
14. Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584, 2002.
15. Miguel Garcia-Diaz and Thomas A Kunkel. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends in biochemical sciences*, 31(4):206–214, 2006.
16. Matthew Hurles. Gene duplication: the genomic trade in spare parts. *PLoS biology*, 2(7):e206, 2004.
17. Sebastian Proost, Jan Fostier, Dieter De Witte, Bart Dhoedt, Piet Demeester, Yves Van de Peer, and Klaas Vandepoele. i-adhore 3.0fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research*, 40(2):e11–e11, 2012.
18. Alexandros Stamatakis. Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
19. Bassam Alkindy, Jean-François Couchot, Christophe Guyeux, Arnaud Mouly, Michel Salomon, and Jacques M. Bahi. Finding the core-genes of chloroplasts. *Journal of Bioscience, Biochemistery, and Bioinformatics*, 4(5):357–364, 2014.
20. CK Stover, XQ Pham, AL Erwin, SD Mizoguchi, P Warrener, MJ Hickey, FSL Brinkman, WO Hufnagle, DJ Kowalik, M Lagrou, et al. Complete genome sequence of pseudomonas aeruginosa pao1, an opportunistic pathogen. *Nature*, 406(6799):959–964, 2000.