Pan-genome Analysis of the Genus Serratia

Zarrin Basharat*, Azra Yasmin

Microbiology & Biotechnology Research Lab, Department of Environmental Sciences, Fatima Jinnah Women University, Rawalpindi 46000, Pakistan.

*Corresponding author
Email: zarrin.ituig@gmail.com

ABSTRACT
Pan-genome analysis is a standard procedure to decipher genome heterogeneity and diversification of bacterial species. Specie evolution is traced by defining and comparing the core (conserved), accessory (dispensable) and unique (strain-specific) gene pool with other strains of interest. Here, we present pan-genome analysis of the genus Serratia, comprising of a dataset of 100 genomes. The isolates have clinical to environmental origin and consist of ten different species from the genus, along with two subspecies of the representative strain Serratia marcescens.

Out of 19430 non-redundant coding DNA sequences (CDS) from the dataset, 972 (5%) belonged to the core genome. Majority of these genes were linked to metabolic function, followed by cellular processes/signalling, information storage/processing while rest of them were poorly characterized. 10,135 CDSs (52.16%) were associated with 8,321 CDSs (42.82%) were singletons or strain specific. The Pan-genome orthologs indicated a positive correlation to the number of genomes whereas negative correlation was obtained for core genome. Genomes were aligned to obtain information about synteny, insertion/inversion, deletion and duplications. This study provides insights into variation of Serratia species and paves way for pan-genome analysis of other bacterial species at genus level.

Keywords: Genome analysis, core genome, pan-genome, Serratia, codon usage, evolution.

1. INTRODUCTION

Genus Serratia consists of rod-shaped, Gram-negative, endospore forming, facultative anaerobic bacteria belonging to the family Enterobacteriaceae (Li et al., 2015). They exist in diverse habitats including soil (Kamensky et al., 2003; Giri et al., 2004; Lavania and Nautiyal, 2013), water (Gavini et al., 1979; Henriques et al., 2013; Kämpfer and Gläser, 2016), plants (Lim et al., 2015; Afzal et al., 2016; Zaheer et al., 2016), nematodes (Abebe-Akele et al., 2015; Vicente et al., 2016), bats (García-Fraile et al., 2015), humans (Roy et al., 2014; Bonnin et al., 2015) etc, and consist of non-pathogenic as well as some pathogenic strains.

Some Serratia species stimulate plant growth through production of specific compounds as siderophores, lipo-chitin, oligosaccharides etc and are considered plant growth promoting rhizobacteria (Koo and Cho, 2009; Purushotham et al., 2012). Their antagonistic behaviour against phytopathogens has been documented due to a diverse array of antimicrobial compound production (De Vleeschauwer and Hofte, 2003; Adam et al., 2016; Roberts et al., 2016). Prodiginine (with five sub-types i.e. prodigiosin, undecylprodigiosin, cycloprodigiosin, cyclononylprodigiosin, butyl-meta-cyclo-heptylprodiginine) is the most well characterized antimicrobial compound in Serratia (Williamson et al., 2006). Serratia marcescens, a typical/representative strain is an opportunistic pathogen and causes nosocomial infections (Gerc et al., 2015). However, some reports of infections by Serratia plymuthica, Serratia liquefaciens, Serratia rubidosa, and Serratia odoriferae also exist (Weise et al., 2014) Members of Serratia marcescens are distinguished from other Enterobacteriaceae family members due to production of enzymes DNase, lipase, and gelatinase which are not produced by the other species (De Vleeschauwer and Hofte, 2003; Li et al., 2015). Prodigiosin, a red pigment with antibacterial, antifungal, antiprotozoal, immunosuppressive and anticancer properties (Williamson et al., 2007) is normally produced by environmental isolates of Serratia marcescens, but not the clinical isolates (Mahlen, 2011).

High throughput methods for genome sequencing with reduced time and cost have resulted in massive DNA-sequencing data in recent years (Mardis, 2008; Goodwin et al., 2016). To encompass the detailed pan genomic repertoire of genus Serratia, a hundred genomes were taken, encoding for all possible lifestyles carried out by the species. The phylogenetic resolution of the genus of interest ranged from...
species to subspecies/serovar. This effort was centred on heterogeneity analysis and evolutionary study of this genus.

2. MATERIAL AND METHODS

Genomes used in this study were obtained from the public repository ‘NCBI database’ (Table 1). Statistical parameters were calculated for the genomes such as genome size in Mbp (million base pairs), DNA GC content in percentage, Number of CDSs, Mean length of CDSs etc. Pairwise analysis was done via ortholog matrix (parameters: Reference Genome = Serratia marcescens subsp. marcescens Db1, Data type = RBH protein) (Manzano-Marin et al., 2012) as previously implemented for seven Serratia strains. while whole genome alignment was carried out using NUCMER (Delcher et al., 2002; Kurtz et al., 2004; Hasan et al., 2015), to reveal differences caused by duplication, inversion, deletion and insertion. The strain phylogeny was inferred through orthologous average nucleotide identity (OrthoANI) algorithm (Lee et al., 2016). Gene frequency and the gene presence/absence heatmap was constructed for all 100 genomes through UPGMA clustering (Comandatore et al., 2013). A pan-genome is a pool or amalgamation of protein coding genes i.e. CDSs in a specified set of genomes (Vernikos et al., 2015). Set of annotated genomes are taken as a seed substance for the construction of a pan-genome. Genomes are subjected to a pair wise homology search through BLAST or uBLAST. “Reciprocally best hit (RBH)” orthologs are calculated for all possible genome pairs. In case of a partial gene, the reciprocity cannot be defined owing to small length. The gene, therefore, is considered as a singleton/different gene, even though it may have 100% sequence similarity with the orthologous group gene. uCLUST was implemented to circumvent this problem, and genes were regrouped in case of ~95% identity according to Cao et al. (2013).

Pan-genome information was then converted into binary matrix based on the presence (coded as 1) and absence (coded as 0) of genes (Laing et al., 2010; Georgiades et al., 2011). Similarity values were then calculated between two genomes/POGs based on Jaccard coefficient. The obtained similarity matrix was then used for clustering potential CDSs via UPGMA algorithm into non-redundant gene sets, to generate POGs (Contreras-Moreira and Vinuesa, 2013). Frequency of POGs (Lefebure and Stanhope, 2007), functional categorization through COG (Altenhoff and Dessimoz, 2009) and SEED scheme (Disz et al., 2010) was obtained. The pan/core genome curves were then calculated from POGs (Zakham et al., 2012).

3. RESULTS AND DISCUSSION

Similar and dissimilar properties of the genomes were matched through statistical properties of genome size in Mbp (million base pairs), DNA GC content in percentage, Number of CDSs, Mean length of CDSs (Table 2) and simple regression analysis (Table 3). Remarkable differences and similarities were observed. Alignment of two whole genome sequences can be very useful in comparative genomics and it is particularly helpful when alignments are visualized. Gene synteny and genomic structural difference caused by duplication, inversion, deletion and insertion are inferred by this way. For a dot plot, the reference sequence is laid across the x-axis, while the query sequence is on the y-axis. Wherever the two sequences agree, a coloured line or dot is plotted. The forward matches are displayed in red, while the reverse matches are displayed in green. If the two sequences are perfectly identical, a single red line would go from the bottom left to the top right. However, two sequences seldom exhibited this behaviour, and in the studied plots, multiple gaps and inversions were identified for majority genomes. Serratia plymuthica AS12, Serratia liquefaciens ATCC 27592, Serratia nematodiphila WW4 and Serratia proteamaculans S4 showed a significant degree of similarity to the reference strain Serratia marcescens subsp. marcescens Db1 whereas least similarity was observed for Serratia grinesii A2, Serratia subsp. sakuensis and Serratia symbiotica SAf with respect to the reference strain (Fig. 1).

ANI based typing differed for several strains from the traditional typed strain names. An example is of Serratia marcescens (NCBI project accession number: GCF_000336425.1) demarcated as Serratia nematodiphila. ANI and pan-genome studies often challenge the traditionally typed bacterial strain names (Kim et al., 2014; Yi et al., 2014). In this case too, several bacteria were reannotated as new species. A phylogenetic tree was constructed based on ANI and genomes with similar content were clustered together (Fig. 2).
The genome content and genes within the studied bacterial genomes might be altered due to environmental stress, horizontal gene transfer etc. Despite continual variation in whole genome content of bacteria, orthologous genes are shared between multiple genomes. In case of evolutionarily related bacteria, more genes are shared while in case of distantly related species, number of common genes is less. Several strains showed varied pan genomic profile due to presence of significant number of genes that were absent in the other strains and were clustered near to each other depending on the type and number of these genes (Fig. 3). A significant portion of the whole genome (for all studied genomes) belonged to the house keeping/core genes required for essential functions and were present in all of the species while some genes (called singletons) are present only in a single genome. These singletons usually belong to the genomic islands/mobilome with horizontal gene transfer, conjugation, transformation or transduction origin. A “gene frequency plot” helps distinguish the varying trend. A U-shaped plot is usually obtained when most genes are present either in all or just single genome. A U-shaped plot was obtained for the genes of Serratia species.

After formulating pan-genome, all CDSs were grouped into pan-genome orthologous groups (POGs). POGs enclosed a minimum of one CDS (called singleton POGs). Extremely conserved POGs were present in all 100 genomes, and these constituted the core genome. For 100 genomes of Serratia, 19,430 POGs (~4%) were obtained from 485,201 CDSs for pan-genome and 972 POGs (~5%) were obtained from 485,201 CDSs for the core region. Previously, a comparative study by Li et al. (2015) reported that ~4% of the predicted CDSs for pan-genome were shared among ten Serratia genomes (Li et al., 2015). This number was reduced to 4% for 100 genomes. This illustrates the need for increased number of genomes for determining pan and core genome instead of just a few genomes. It helps elucidate the conserved and varying regions in a better way. Some notable examples of bacteria which appear to have an open pan-genome include E. coli (Collins and Higgs, 2012), Streptococcus agalactiae (Tettelin et al., 2005), Prochlorococcus (Kettler et al., 2007), Listeria (den Bakker et al. 2010) etc whereas closed pan-genome has been deciphered for Clostridium difficile (Scaria et al., 2010), with a sample size of 26 genomes rendered sufficient for discernment of the entire pan-genome. However, this result may depend on the way the extrapolation was done and may not indicate a qualitative difference between Clostridium difficile and other bacteria. Expanding the idea of open and close pan-genome, it should be noted that mostly environmental bacteria have an open pan-genome while other bacteria exhibit closed pan-genome (Vernikos et al., 2015), but at genus scale, bacteria from diverse type of habitats and varying characteristics are included and it is difficult to comment upon the closedness or openness of its pan-genome. Vernikos et al. (2015) have suggested a combinatorial method for pan-genome prediction, where average values are taken for all data types in the set. This is attempted by first calculating the new, core and shared genes at the nth genome through the following function, with C being the total number of comparisons and N total number of genomes.

\[ C = \frac{N!}{(n - 1)! \cdot (N - n)!} \]

Scalability issue has been addressed by developing a method for sub-sampling the number of comparisons to be performed between N genomes through random selection and comparison. Each strain is however, ensured to undergo equal number of comparisons. Some additions in the metric are required which might account for the environmental/pathogenic isolates, habitat, genome size, total number of genes, threshold for calculation of ortholog relatedness and related parameters for calculation of pan-genome openness or closedness. We postulate that a pan-genome should always be open in bacteria because the genomes are continuously evolving, either due to horizontal gene transfer, environmental impact or new gene arousal or deletion simply according to the neutral Wright–Fisher model (Wright 1969; Baumdicker et al., 2010)

Frequency of POGs, functional categorization through COG and SEED scheme was studied which illustrated a visible difference between pan and core genome for specific set of proteins (related to particular cell functions) for all studied genomes (Fig. 4). Functional categories according to SEED and COG with a cut-off of 50% and 100% were studied and variation of core genes to the pan genome was highly contrasting. The contrasting categories according to SEED scheme included DNA/RNA/protein coding machinery and co-factor/vitamin/prosthetic group genes as well pigment synthesis genes more localized within core genome region. A small portion of the core while a large portion of the pan-genome demonstrated no match to the database contents. According to COG functional categorization,
amino acid transport/metabolism, translational machinery, ribosomal structure proteins, chaperones and post translational modification apparatus contributed significantly to the core portion as compared to the miniscule portion of these genes in the pan-genome region. Only a tiny portion of core genome with a cut-off less than 100% and pan-genome regions could not be related to the database contents. POGs (shown as boxplots) signify the distribution of values of genome data set obtained through random selection of the genomes on X-axis, 100 times each. Boxplot and trend line illustrations the shift in number of pan-genome and core genome (Fig. 5) gene groups to the number of genomes. The pan-genome curve shows a positive correlation between the number of POGs and the number of genomes. Pan-genome analysis of 100 genomes with the cut-off setting at 100% resulted in core genome that included only the genomes with all of the certain genes. In contrast, the core genome curve showed a negative correlation where the number of POGs decreased as the number of genomes increased. In the case of core genome, it was calculated by using a cut-off setting of 80 and 85 and 100%. The cut-off setting specified the calculation of core genome by using genomes with certain genes that were above 80, 85 and 100% of the genomes in the entire data set. Overall, we saw that the number of POGs that belonged to the core-genome increased as the cut-off percentage got smaller.

Core-genome size of this bacteria is small (972 genes for ~5Mbp/100 genomes) as compared to Campylobacter jejuni (1042 genes for ~1.6 Mbp/130 genomes) (Meric et al., 2014) but sufficiently large as compared to the Bacillus species (143 genes for ~4.5 Mbp/172 genomes) (Collins and Higgs, 2012). The Bacillus specie might have reduced core as number of genomes were large and they are different from their own subspecie i.e. Geobacillus (527 genes for ~4.5 Mbp/29 genomes). This might be due to small sample size in case of Geobacillus (Bezuindt et al., 2016). Overall, this observation is a bit interesting and warrants further investigation. Codon usage bias analysis and fitting of pan-genome data using various models like random permutation model, power law regression model, exponential regression model etc would also be of aid in better understanding evolution of bacteria. The authors are currently involved in studying phage and mobilome impact on pan-genomics of Serratia. We expect that similar efforts at genus scale for other bacteria would also be undertaken by our group in the future.

4. CONCLUSION

We have presented the pan-genome analysis of Serratia species and this has allowed us to gain an improved understanding of this versatile genus. Individual genomes of the Serratia are made up of core and pan CDS. Mining the accessory genomes and alignment of species has shed light on the genome conservation. Compared to other bacterial species, Serratia does not depict exceptional genetic conservation or diversity. Standardization of pan-genome parameters would lead to understanding the details of this and contribute to knowledge of different phenotypes compared to molecular genotypes. Incorporation of expression data, metabolic pathways and mobile elements etc is also desired. With the development of new approaches and better tools on horizon, future of bacterial pan-genomics seems bright.

NOTE: Supplementary data consisting of core and pan genome orthologs is available upon request from the authors.

ACKNOWLEDGEMENTS
The authors are thankful to Alexandra Elbakyans for providing literature support. Furthermore, we are grateful to all those scientists, research labs and universities who made computational biology tools, data and web servers freely available for public use.

FUNDING
This study did not receive any particular funding.

CONFLICT OF INTEREST
The authors declare that they have nothing to disclose which might pose a potential conflict of interest.
Table 1. Characteristics of the studied 100 *Serratia* genome species.

| Serial no. | NCBI project | Accession no. | Strain name | Name | Status | No. of contigs | Genome size (bp) | DNA G+C content (%) | No. of CDS | No. of rRNA genes | No. of tRNA genes |
|------------|--------------|---------------|-------------|------|--------|---------------|------------------|---------------------|------------|------------------|------------------|
| 1          | GCF_00001808 | 5.1           | 968         | *Serratia proteamaculans* 968 | COMPLETE | 2           | 549365          | 55.0238           | 3          | 506              | 22               |
| 2          | GCF_00021499 | 5.1           | AS12        | *Serratia sp.* AS12          | COMPLETE | 1           | 544300          | 55.9644           | 4          | 497              | 22               |
| 3          | GCF_00021423 | 5.1           | AS9         | *Serratia plymuthica* AS9    | COMPLETE | 1           | 544288          | 55.9611           | 5          | 497              | 22               |
| 4          | GCF_00021480 | 5.1           | AS13        | *Serratia sp.* AS13          | COMPLETE | 1           | 544254          | 55.9605           | 8          | 497              | 22               |
| 5          | GCF_00033086 | 5.1           | FG194       | *Serratia marcescens* FG194  | COMPLETE | 1           | 489821          | 58.8617           | 3          | 436              | 22               |
| 6          | GCF_00033642 | 5.1           | WW4         | *Serratia marcescens* WW4    | COMPLETE | 2           | 524470          | 59.5470           | 3          | 483              | 22               |
| 7          | GCF_00047854 | 5.1           | RVH1        | *Serratia plymuthica* RVH1   | COMPLETE | 1           | 551432          | 59.1983           | 3          | 502              | 21               |
| 8          | GCF_00076853 | 5.2           | 4Rx13       | *Serratia plymuthica* 4Rx13  | COMPLETE | 2           | 549373          | 56.1339           | 2          | 493              | 22               |
| 9          | GCF_00043882 | 5.1           | Si3         | *Serratia plymuthica* Si3    | COMPLETE | 1           | 546730          | 56.1953           | 2          | 493              | 22               |
| 10         | GCF_00042208 | 5.1           | ATCC 27592(Type) | *Serratia liquefaciens ATCC 27592 | COMPLETE | 2           | 528279          | 55.3295           | 4          | 489              | 24               |
| 11         | GCF_00053215 | 5.1           | Db11        | *Serratia marcescens* subsp. marcescens Db1 | COMPLETE | 1           | 511802          | 59.5163           | 2          | 467              | 22               |
| 12         | GCF_00082877 | 5.1           | SM39        | *Serratia marcescens* SM39   | COMPLETE | 3           | 532602          | 59.7424           | 2          | 489              | 22               |
| 13         | GCF_00069599 | 95.1          | FS14        | *Serratia sp.* FS14          | COMPLETE | 1           | 524987          | 59.4623           | 1          | 477              | 21               |
| 14         | GCF_00074756 | 5.1           | SCBI        | *Serratia sp.* SCBI          | COMPLETE | 2           | 510189          | 59.5932           | 9          | 474              | 21               |
| 15         | GCF_00097524 | 5.1           | HUMV-21     | *Serratia liquefaciens*      | COMPLETE | 1           | 532665          | 55.1901           | 2          | 491              | 22               |
| 16         | GCF_00000600 | 5.1           | DSM 4576(Type) | *Serratia fonticola*         | COMPLETE | 1           | 600051          | 53.6128           | 2          | 532              | 22               |
| 17         | GCF_00012215 | 5.1           | CAV1492     | *Serratia marcescens*        | COMPLETE | 6           | 582840          | 58.6126           | 2          | 5397             | 22               |
| 18         | GCA_0006485 | 75.1          | V4          | *Serratia plymuthica*        | CHROMOSOME | 1           | 551353          | 56.1942           | 2          | 5073             | 3                |
| 19         | GCF_0009880 | 45.1          | LR5/4       | *Serratia ureilytica*        | CONTIG    | 1           | 539136          | 59.183            | 4          | 498              | 22               |
| 20         | GCF_00052192 | 5.1           | BIDMC 44    | *Serratia marcescens* BIDMC 44 | SCAFFOLD | 2           | 534166          | 59.4347           | 8          | 482              | 22               |
| 21         | GCF_00034799 | 5.1           | S4          | *Serratia sp.* S4            | SCAFFOLD  | 2           | 545474          | 59.4987           | 1          | 495              | 23               |
| 22         | GCF_00078391 | 65.1          | FDA_MicroDB  | *Serratia marcescens*        | CONTIG    | 1           | 524825          | 59.8771           | 2          | 478              | 22               |
| 23         | GCA_00103440 | 5.1           | UCL88       | *Serratia marcescens*        | SCAFFOLD  | 1           | 518749          | 59.6027           | 5          | 472              | 22               |
| 24         | GCF_00074339 | 5.1           | CDC 831-60(Type) | *Serratia marcescens*        | SCAFFOLD  | 2           | 513164          | 59.7836           | 4          | 472              | 22               |
| ID       | Organism 1                     | Organism 2                     | Scaffolds | GCF_ID | GCA_ID | Start | End   | Length | Closure | Summary  |
|----------|-------------------------------|-------------------------------|-----------|--------|--------|-------|-------|--------|---------|----------|
| GCF_00016359  | DSM 4582(Type)                | Serratia odorifera DSM 4582   | SCAFFOLD  | 53     | 54     | 526738| 526738| 33     | 543744  | 4901     | 3      | 70    |
| GCF_00078397  | FDAARGOS_62                   | Serratia marcescens           | CONTIG    | 4      | 477    | 516411| 516411| 4       | 595697  | 22      | 92    |
| GCA_00034377  | BWH56                        | Serratia marcescens           | SCAFFOLD  | 10     | 5174   | 562004| 562004| 10     | 583597  | 22      | 84    |
| GCF_00073867  | DSM 21420(Type)               | Serratia marcescens           | CONTIG    | 2      | 480    | 522492| 522492| 2      | 595360  | 24      | 91    |
| GCF_00026427  | LCT-SM213                    | Serratia marcescens LCT-SM213| SCAFFOLD  | 11     | 463    | 506732| 506732| 11     | 591519  | 5       | 73    |
| GCF_00052190  | BIDMC 50                     | Serratia marcescens           | SCAFFOLD  | 4      | 469    | 512065| 512065| 4      | 595622  | 24      | 85    |
| GCF_00073921  | NGS-ED-1015                  | Serratia marcescens           | CONTIG    | 7      | 468    | 512536| 512536| 7      | 599127  | 8       | 78    |
| GCF_00069916  | Fk01                         | Serratia liquefaciens         | SCAFFOLD  | 28     | 4851   | 528013| 528013| 28     | 557457  | 9       | 80    |
| GCA_00030267  | BWH57                        | Serratia marcescens           | SCAFFOLD  | 13     | 5217   | 565925| 565925| 13     | 582771  | 26      | 84    |
| GCA_00034399  | UCl87                        | Serratia marcescens           | SCAFFOLD  | 12     | 472    | 533118| 533118| 12     | 595811  | 15      | 85    |
| GCF_00080387  | RM66262                      | Serratia marcescens           | CONTIG    | 19     | 444    | 488225| 488225| 19     | 600528  | 9       | 79    |
| GCF_00044245  | LCT-SM166                    | Serratia marcescens LCT-SM166| SCAFFOLD  | 13     | 4751   | 507002| 507002| 13     | 591237  | 24      | 92    |
| GCA_00010586  | Ah0650_Sm1                   | Serratia marcescens subsp.    | CONTIG    | 20     | 4791   | 520165| 520165| 20     | 595269  | 24      | 92    |
| GCA_00078364  | FDAARGOS_79                  | Serratia marcescens           | CONTIG    | 14     | 5543   | 538778| 538778| 14     | 597084  | 31      | 88    |
| GCF_00000755  | 90-166                       | Serratia marcescens           | CONTIG    | 63     | 568439 | 548439| 548439| 63     | 591129  | 10      | 85    |
| GCF_00063371  | BIDMC 80                     | Serratia marcescens           | SCAFFOLD  | 37     | 4861   | 527423| 527423| 37     | 596099  | 6       | 79    |
| GCF_00063355  | PH1a                         | Serratia marcescens           | CONTIG    | 36     | 489    | 524384| 524384| 36     | 592326  | 4       | 81    |
| GCF_00063333  | Hqt                          | Serratia marcescens           | CONTIG    | 60     | 489    | 524309| 524309| 60     | 592339  | 5       | 81    |
| GCF_00047764  | AU-AP2C                      | Serratia fonticola AU-AP2C    | CONTIG    | 34     | 460    | 499981| 499981| 34     | 537544  | 6       | 79    |
| GCA_00016337  | 420_SMAR                     | Serratia marcescens           | CONTIG    | 37     | 478    | 520638| 520638| 37     | 594396  | 9       | 74    |
| GCA_00016497  | 454_SMAR                     | Serratia marcescens           | CONTIG    | 33     | 479    | 520966| 520966| 33     | 59446   | 13      | 77    |
| GCF_00046807  | AU-P3(3)                     | Serratia fonticola AU-P3(3)   | CONTIG    | 44     | 470    | 502312| 502312| 44     | 537538  | 5       | 65    |
| GCF_00082188  | SAF                          | Serratia symbiotica           | CONTIG    | 32     | 3997   | 358774| 358774| 32     | 520799  | 35      | 74    |
| GCF_00044237  | LCT-SM262                    | Serratia marcescens LCT-SM262| SCAFFOLD  | 55     | 474    | 516015| 516015| 55     | 559932  | 5       | 80    |
| GCF_00063369  | BIDMC 81                     | Serratia marcescens           | SCAFFOLD  | 33     | 465    | 504622| 504622| 33     | 596697  | 11      | 65    |
| GCA_00062992  | 20_SPLY                      | Serratia liquefaciens         | CONTIG    | 40     | 474    | 516450| 516450| 40     | 554322  | 8       | 68    |
| GCF_00073447  | YDC963                       | Serratia marcescens           | CONTIG    | 54     | 462    | 507364| 507364| 54     | 599700  | 5       | 74    |

Basharat and Yasmin, 2016
| Accession | Version | Species | Strain/Name | Type | SMAR | Start | End | Length |
|-----------|---------|---------|-------------|------|------|-------|-----|--------|
| GCA_00106471 | 5.1 | Serratia marcescens | 370_SMAR | CONTIG | 52 | 509632 | 59.4628 | 478 | 5 | 44 |
| GCF_00034220 | 5.1 | Serratia marcescens | VGH107 | CONTIG | 83 | 559124 | 59.4641 | 5157 | 6 | 85 |
| GCF_00029236 | 5.1 | Serratia marcescens | W2.3 | CONTIG | 72 | 523797 | 59.2388 | 4891 | 5 | 74 |
| GCA_00106917 | 5.1 | Serratia marcescens | 286_SMAR | CONTIG | 49 | 512046 | 59.7706 | 475 | 4 | 66 |
| GCA_00106433 | 5.1 | Serratia marcescens | 287_SMAR | CONTIG | 54 | 512281 | 59.7672 | 4755 | 5 | 68 |
| GCA_00106314 | 5.1 | Serratia marcescens | 280_SMAR | CONTIG | 55 | 513427 | 59.7665 | 474 | 3 | 66 |
| GCF_00074335 | 5.1 | Serratia sp. Ag2 | 534757 | CONTIG | 110 | 522653 | 59.7665 | 496 | 13 | 72 |
| GCF_00086448 | 5.1 | Serratia symbiotica str. Tucson | Tucson | SCAFFOLD | 388 | 278921 | 47.9491 | 3334 | 22 | 51 |
| GCF_00045661 | 5.1 | Serratia marcescens | EGD-HP20 | CONTIG | 62 | 508300 | 59.7603 | 470 | 5 | 81 |
| GCA_00106312 | 5.1 | Serratia marcescens | 276_SMAR | CONTIG | 108 | 516539 | 59.5680 | 479 | 9 | 64 |
| GCA_0007662 | 5.1 | Serratia marcescens | 294_SMAR | CONTIG | 75 | 511696 | 59.7699 | 4791 | 5 | 46 |
| GCA_00106485 | 5.1 | Serratia marcescens | 398_SMAR | CONTIG | 64 | 519846 | 59.4513 | 479 | 5 | 50 |
| GCA_00106532 | 5.1 | Serratia marcescens | 534_SMAR | CONTIG | 89 | 509923 | 59.7199 | 479 | 3 | 35 |
| GCA_00106434 | 5.1 | Serratia marcescens | 290_SMAR | CONTIG | 61 | 515483 | 59.7716 | 474 | 5 | 54 |
| GCA_00066228 | 5.1 | Serratia marcescens | 198.repi_SMA R | CONTIG | 77 | 519999 | 59.9309 | 479 | 2 | 43 |
| GCA_00106332 | 5.1 | Serratia marcescens | 410_SMAR | CONTIG | 74 | 520281 | 59.4431 | 480 | 11 | 52 |
| GCF_00030089 | 5.1 | Serratia plymuthica | A30 | CONTIG | 78 | 555402 | 59.1027 | 5179 | 4 | 65 |
| GCF_00076695 | 5.1 | Serratia marcescens | 198 | CONTIG | 66 | 510009 | 59.8129 | 474 | 15 | 84 |
| GCF_00074336 | 5.1 | Serratia sp. Ag2 | 532481 | CONTIG | 109 | 522741 | 59.927 | 492 | 16 | 69 |
| GCF_0004893 | 5.1 | Serratia marcescens | AB42536490-isolate1 | SCAFFOLD | 117 | 533487 | 59.1927 | 500 | 4 | 64 |
| GCF_0004883 | 5.1 | Serratia marcescens | MC6001 | SCAFFOLD | 131 | 536762 | 59.089 | 504 | 7 | 61 |
| GCF_0004889 | 5.1 | Serratia marcescens | MC459 | SCAFFOLD | 129 | 537306 | 59.089 | 505 | 5 | 66 |
| GCA_0006483 | 5.1 | Serratia marcescens | 395_SMAR | CONTIG | 79 | 519953 | 59.4580 | 479 | 9 | 37 |
| GCA_00060662 | 5.1 | Serratia marcescens | 1241_SMAR | CONTIG | 117 | 517553 | 59.3411 | 472 | 6 | 29 |
| GCF_0004887 | 5.1 | Serratia marcescens | MC460 | SCAFFOLD | 139 | 537197 | 59.0638 | 503 | 8 | 62 |
| GCF_0004885 | 5.1 | Serratia marcescens | MC6000 | SCAFFOLD | 134 | 533316 | 59.1966 | 505 | 7 | 61 |
| GCF_0004881 | 5.1 | Serratia marcescens | MC620 | SCAFFOLD | 138 | 528980 | 59.825 | 492 | 6 | 68 |
| GCA_00106737 | 5.1 | Serratia | 790_SMAR | CONTIG | 159 | 568713 | 59.3041 | 5342 | 7 | 50 |
|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 5.1 | GCA_00106223 | Serratia marcescens | CONTIG | 139 | 510346 |
| 5.1 | GCF_00073488 | A2 | Serratia grimesii | CONTIG | 120 | 513306 |
| 5.2 | GCF_00049675 | DD3 | Serratia sp. | CONTIG | 123 | 527353 |
| 5.1 | GCA_00106065 | 1242_SMAR | Serratia marcescens | CONTIG | 144 | 513922 |
| 5.1 | GCA_00106701 | 706_SMAR | Serratia marcescens | CONTIG | 125 | 510620 |
| 5.1 | GCA_00106119 | 7098.rep2_SMAR | Serratia marcescens | CONTIG | 135 | 512623 |
| 5.1 | GCF_00047631 | UTAD54 | Serratia fonticola UTAD54 | CONTIG | 133 | 595342 |
| 5.2 | GCA_00033636 | C-1 | Serratia sp. | SCAFFOLD | 133 | 565614 |
| 5.1 | GCF_00041891 | MC458 | Serratia marcescens MC458 | SCAFFOLD | 156 | 540209 |
| 5.1 | GCF_00026104 | PRI-2C | Serratia plymuthica PRI-2C | CONTIG | 104 | 391916 |
| 5.1 | GCA_00106445 | 31_SMAR | Serratia marcescens | CONTIG | 164 | 509539 |
| 5.1 | GCA_00106098 | 1218_SMAR | Serratia marcescens | CONTIG | 154 | 512689 |
| 5.1 | GCA_00106145 | 1186_SMAR | Serratia marcescens | CONTIG | 204 | 515778 |
| 5.1 | GCA_00106808 | 907_SMAR | Serratia marcescens | CONTIG | 185 | 492060 |
| 5.1 | GCA_00106122 | 1219_SMAR | Serratia marcescens | CONTIG | 156 | 513737 |
| 5.1 | GCA_00106593 | 684_SMAR | Serratia marcescens | CONTIG | 239 | 547263 |
| 5.1 | GCF_00063333 | Hin | Serratia sp. Hin | CONTIG | 247 | 612335 |
| 5.1 | GCF_00063335 | Hgw | Serratia sp. Hgw | CONTIG | 240 | 612602 |
| 5.1 | GCA_00106694 | 698_SMAR | Serratia marcescens | CONTIG | 193 | 521300 |
| 5.1 | GCA_00106472 | 374_SMAR | Serratia marcescens | CONTIG | 197 | 518509 |
Table 2. General statistical values (mean, median, standard deviation) for genome size and other properties for the studied genus. Intergenic regions are spacers or regions between genes, shorter intergenic regions mean more compact arrangement of genes and hence, genome.

| Property                        | Mean   | Median  | SD  |
|---------------------------------|--------|---------|-----|
| Genome size (Mbp)               | 5.23   | 5.21    | 0.4 |
| GC content (%)                  | 57.95  | 59.27   | 2.62|
| No. of CDS                      | 4852.01| 4817    | 335.95|
| Mean length of CDS              | 940.4  | 949.08  | 38.08|
| Mean length of intergenic region| 149.23 | 146.11  | 13.5 |
Table 3. Regression analysis of genome properties. For regression analysis of genome size to the number of CDSs among 100 *Serratia* strains, number of CDSs within this species were positively correlated to genome size ($r^2=0.95$) and the coefficient of determination ($R^2=0.9$) meant that 90% of data points supported the predicted regression line ($y = 719 + 790x$).

| Property                        | Genome size (Mbp) | GC content (%) | No. of CDS | Mean length of CDS | Mean length of intergenic region |
|---------------------------------|-------------------|----------------|------------|--------------------|----------------------------------|
| Genome size (Mbp)               |                   |                |            |                    |                                  |
| GC content (%)                  | $Y=53+0.87x$      | $R^2=0.02$     | $r^2=0.13$ |                    |                                  |
| No. of CDS                      | $Y=719+790x$      | $R^2=0.9$      | $r^2=0.95$ |                    | $Y=55+0.00054x$                  |
| Mean length of CDS              | $Y=643+57x$       | $R^2=0.36$     | $r^2=0.6$  | $Y=32+0.028x$      |                                  |
| Mean length of intergenic region| $Y=207+11x$       | $R^2=0.11$     | $r^2=0.33$ |                    |                                  |
Fig. 1. NUCMER alignment plot of studied 10 species, classified according to ANI scheme and including (A) *Serratia plymuthica* AS12, (B) *Serratia fonticola* DSM 4576, (C) *Serratia grinesii* A2, (D) *Serratia liquefaciens* ATCC 27592, (E) *Serratia nematodiphila* WW4, (F) *Serratia odorifera* DSM 4582, (G) *Serratia proteamaculans* S4, (H) *Serratia* subsp sakuensis and (I) *Serratia symbiotica* SAf with respect to the reference genome specie *Serratia marcescens* subspecie *marcescens* Db11. Rearrangement of genomic regions was observed in the genomes, in term of collinearity and their localization on the negative or positive strand.
Fig. 2. Phylogeny of studied genomes based on ANI.
Fig. 3. Gene presence/absence matrix of all CDSs of 100 Serratia genomes. Black depicts presence while gray shows absence of a gene. Specie names are shown on the right side. The strains clustered on the top i.e. SAI, Tucson, DD3, Ag1, Ag2, H1n and h1w show presence of significant number of genes that are absent in the other strains.
Fig. 4. Functional categories according to seed scheme. (B) Functional categories by COG scheme. Cut-off of 50% and 100% for core genome is shown.
Fig. 5. Pan (top) and core-genome (bottom) curves, calculated from POGs of 100 *Serratia* genomes. Core genome cut-off is 100% in this figure.