Gene xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

PanGeT: Pan-genomics tool

Iyyappan Yuvaraj [a,1], Jayavel Sridhar [b,c,1], Daliah Michael [a], Kanagaraj Sekar [a,*]

[a] Department of Computational and Data Sciences, Indian Institute of Science, Bangalore 560012, India
[b] Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai 625021, India
[c] Department of Biotechnology (DDE), Madurai Kamaraj University, Madurai 625021, India

ARTICLE INFO

Article history:
Received 27 May 2016
Received in revised form 5 November 2016
Accepted 10 November 2016
Available online xxxx

Keywords:
Pan-genome
Core genes
Dispensable genes
Unique genes
Species
Genome
Proteome
CDS

ABSTRACT

A decade after the concept of Pan-genome was first introduced; research in this field has spread its tentacles to areas such as pathogenesis of diseases, bacterial evolutionary studies and drug resistance. Gene content-based differentiation of virulent and a virulent strains of bacteria and identification of pathogen specific genes is imperative to understand their physiology and gain insights into the mechanism of genome evolution. Subsequently, this will aid in identifying diagnostic targets and in developing and selecting vaccines. The root of pan-genomic studies, however, is to identify the core genes, dispensable genes and strain specific genes across the genomes belonging to a clade. To this end, we have developed a tool, "PanGeT – Pan-genomics Tool" to compute the 'pan-genome' based on comparisons at the genome as well as the proteome levels. This automated tool is implemented using LaTeX libraries for effective visualization of overall pan-genome through graphical plots. Links to retrieve sequence information and functional annotations have also been provided. PanGeT can be downloaded from http://pranag.physics.iisc.ernet.in/PanGeT/ or https://github.com/PanGeTv1/PanGeT.

1. Introduction

The term "Pan-genome" was first introduced by Tettelin et al. (2005) a decade ago. It defines the complement of genes from genomes belonging to a clade, which includes core, dispensable and strain specific genes. In recent decades, the availability of huge genome data has facilitated and accelerated the pace of pan-genomic studies (Medini et al., 2005; Puigbò et al., 2014). It has gained popularity owing to its potential in understanding genome evolution and in identifying potential molecular targets (Ho et al., 2012; Delany et al., 2013). It is known that bacteria undergo minor evolutionary changes by acquiring DNA sequences from plasmids, phages and other genomes of the same species, which enhances their survival, adaptation and virulence (Boto, 2015). Many changes were noticed in the physiology of the pathogens after acquiring the genetic pool from external sources. The development of drug resistance by novel microbial genetic elements in many microbes, pose the greatest threat to therapeutic solutions, thereby highlighting the importance of the acquisition and reduction of specific DNA sequences in a pathogen. Identification and analysis of such genome fluidity will contribute to the clinical management, epidemiology and development of new diagnostic methods (Ahmed et al., 2008). Studies have revealed that such strain specific genes were applicable for the identification of vaccine candidates and diagnostic and antimicrobial targets (Muzzi et al., 2007). Thus, close monitoring of the distribution of virulent gene content in pathogenic strains of a species is mandatory for their effective containment (Sugawara et al., 2013).

Over the last decade, few computational tools like Pan-genome-tree (Snipen and Ussery, 2010), PanSeq (Laing et al., 2010), GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013) and PGAP (Zhao et al., 2012) were developed for pan-genomic studies. However, it is noticed that most of these tools use BLASTP (Altschul et al., 1990) for the multiple proteome comparisons, which does not take into account the non-transcribed or un-translated regions, for comparative analysis. Moreover, the existing pan-genome tools require numerous dependencies for installation and are restricted to compare only a limited number of genomes. Therefore, to address these issues, we have proposed a new robust, user friendly and time inexpensive effective tool to identify core genes, dispensable genes and strain specific genes.

2. Materials and methods

2.1. Implementation

PanGeT has been conceived and developed to facilitate researchers to compare entire genomes or proteomes together in order to compute

Abbreviations: CDS, Coding DNA Sequences; RBH, Reciprocal Best Hit; H-value, Homology value; *E*-value, expected value.

* Corresponding author.
E-mail addresses: sekar@cds.iisc.ac.in, sekar@physics.iisc.ernet.in (K. Sekar).

[1] Contributed equally.

http://dx.doi.org/10.1016/j.gene.2016.11.025
0378-1119/

Table 1
List of twenty *Salmonella* serovars analyzed for their pan-genome content. [$]*Salmonella enterica* strains taken from GenBank, [*]Accession number and gene annotations reported in NCBI-RefSeq, [@]Core (Reciprocal Best Hits exclusive of paralogous and partial homologues) and [%]Unique genes identified using PanGeT, Gene Family algorithm (PGAP-GF) and Multiparanoid algorithm (PGAP-MP) with PGAP. [!]Core and Unique genome fragments were computed using PanSeq. [~]Core Amino Acid gene clusters identified by GET_HOMOLOGUES tool.

S. no.	Serovar[$]	RefSeq Id[*]	Total CDS[*]	Comparison of pan-genome prediction tools											
				PanGeT BLASTN mode		PanGeT BLASTP mode		PGAP-GF		PGAP-MP		PanSeq		GET_HOMOLOGUES	
				Core (RBH-hits)[@]	Unique[%]	Core (RBH-hits) [@]	Unique[%]	Core[@]	Unique[%]	Core[@]	Unique[%]	Core fragments[!]	Unique fragments[!]	Core[~]	Unique[~]
1	*Salmonella enterica* Typhimurium LT2	NC_003197	4451	2559	31	2586	17	2712	38	2756	32	11,334[!] (Fragments)	117	2515 (protein)/2557 (DNA)	Does not provide unique gene data
2	*Salmonella enterica* Typhi CT18	NC_003198	4111		31		15	2686	30	2721	29		81		
3	*Salmonella enterica* Typhi Ty2	NC_004631	4352		12		30	2685	66	2726	46		3		
4	*Salmonella enterica* Paratyphi A ATCC 9150	NC_006511	4096		1		3	2689	29	2722	7		23		
5	*Salmonella enterica* Choleraesuis SC B67	NC_006905	4385		18		28	2691	84	2740	57		54		
6	*Salmonella enterica* arizonaeserovar 62_z4_z23	NC_010067	4467		829		689	2688	880	2746	747		176		
7	*Salmonella enterica* Newport SL254	NC_011080	4605		111		77	2707	139	2748	79		27		
8	*Salmonella enterica* Heidelberg SL476	NC_011083	4666		150		128	2708	177	2749	131		34		
9	*Salmonella enterica* Schwarzengrund CVM19633	NC_011094	4538		177		157	2705	222	2752	171		56		
10	*Salmonella enterica* Paratyphi A AKU 12601	NC_011147	4079		1		3	2688	18	2724	4		0		
11	*Salmonella enterica* Agona SL483	NC_011149	4580		100		111	2697	171	2745	120		50		
12	*Salmonella enterica* Dublin CT 02021853	NC_011205	4530		46		67	2700	118	2738	74		36		
13	*Salmonella enterica* Paratyphi C RKS4594	NC_012125	4577		15		36	2698	87	2743	50		7		
14	*Salmonella enterica* Typhimurium SL1344	NC_016810	4446		0		8	2712	25	2756	11		12		
15	*Salmonella enterica* Gallinarumpullorum RKS5078	NC_016831	4325		32		39	2690	100	2734	63		3		
16	*Salmonella enterica* Typhi_P_stx_12	NC_016832	4690		0		75	2691	185	2731	147		0		
17	*Salmonella enterica* Typhimurium 14028S	NC_016856	5315		12		360	2711	495	2756	449		5		
18	*Salmonella enterica* Typhimurium ST4_74	NC_016857	4625		1		20	2710	51	2758	23		0		
19	*Salmonella enterica* Typhimurium T000240	NC_016860	4713		80		79	2716	142	2764	141		12		
20	*Salmonella enterica* Typhimurium UK_1	NC_016863	4451		5		9	2709	34	2757	9		0		