INTER-FACULTY MASTER PROGRAM on

COMPLEX SYSTEMS and NETWORKS

SCHOOL of MATHEMATICS

SCHOOL of BIOLOGY

SCHOOL of GEOLOGY

SCHOOL of ECONOMICS

ARISTOTLE UNIVERSITY of THESSALONIKI

Master Thesis

Title:

Defining the statistical metrics of a pangenome

Καθορίζοντας τις στατιστικές μετρικές ενός Πανγονιδιώματος

Asterios Mpatziakas

Supervisor: Stefanos Sgardelis, Professor AUTH

Co-supervisors:

Fotis E. Psomopoulos, Researcher INAB CERTH

Theodoros Moysiadis, Researcher INAB CERTH

Thessaloniki, July 2017

INTER-FACULTY MASTER PROGRAM on

COMPLEX SYSTEMS and NETWORKS

SCHOOL of MATHEMATICS

SCHOOL of BIOLOGY

SCHOOL of GEOLOGY

SCHOOL of ECONOMICS

ARISTOTLE UNIVERSITY of THESSALONIKI

Master Thesis

Title:

Defining the statistical metrics of a pangenome

Καθορίζοντας τις στατιστικές μετρικές ενός Πανγονιδιώματος

Αστέριος Μπατζιάκας

ΕΠΙΒΛΕΠΩΝ: Στέφανος Σγαρδέλης

Καθηγητής Α.Π.Θ.

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή την 13η Ιουλίου 2017.

…………………………	…………………………	…………………………
Σ. Σγαρδέλης	Φ. Ψωμόπουλος	Θ. Μωυσιάδης
Καθηγητής Α.Π.Θ.	Ερευνητής, ΙΝΕΒ ΕΚΕΤΑ	Ερευνητής, ΙΝΕΒ ΕΚΕΤΑ

Θεσσαλονίκη, Ιούλιος 2017

………………………………………….

Αστέριος Μπατζιάκας

Πτυχιούχος Μαθηματικός Α.Π.Θ.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευτεί ότι εκφράζουν τις επίσημες θέσεις του Α.Π.Θ.

Abstract

Advances in sequencing techniques have massively increased the publicly accessible genome data and thus enable further and more extensive research opportunities on genome diversity at increasing levels of detail. The concept of the pangenome refers to the union of gene families shared by a set of genomes. There are several studies that have implemented specific pangenome analyses for a variety of organisms, ranging from microbes to viruses and plants, leading to genomic projects of various scales. These projects have led to the advancement of general understanding of evolutionary mechanisms, leading to usable knowledge across multiple sectors such as health, medicine and agriculture. A pangenome can be defined as the identification and construction of three distinct subsets of gene families, the Core genome consisting of all gene families that are shared amongst all genomes, the Dispensable or Accessory genome consisting of gene families present in the majority of the genomes and genes that have presence only in one genome, known as Peripheral or Cloud genome. Other names and overlapping definitions have been used in literature that provide alternate description of a pangenome. However, the essential part of this type of analysis is the use of data in an encompassing way instead of the traditionally linear approaches evident in targeted genome studies.

Currently there is a variety of tools available, enabling several computational aspects of the pangenome approach, the majority of which are primarily aimed towards the study of prokaryote genomes. We present a package written for the statistical programming language *R*, named pasaR, usable in the later stages of such an analysis, i.e. after the construction of the gene families for a given set of genomes, based on information of the full complement of gene families. A complete methodology is proposed, suitable for sets of genomes of varying complexity, optimizing and enriching an assortment of existing measures from micropan, the only R package currently available on CRAN for such studies. Furthermore, we propose a new technique using the Sorensen distance, referred to as fluidity in the context of a pangenome analysis, that allows the identification of distinct subsets of genomes in a given dataset, based on their inferred commonalities at the gene family level. Finally, we demonstrate the methodology using publicly available data from UniProt and additional reference databases.

Keywords: pangenome, genome diversity, comparative genomics, R statistical language

Περίληψη

Η πρόοδος στις τεχνικές sequencing έχει αυξήσει τον δημόσια διαθέσιμο, όγκο της πληροφορίας που αφορά το γονιδίωμα επιτρέποντας περαιτέρω και εις βάθος ερευνητική δραστηριότητα στο ζήτημα της γονιδιακής ποικιλομορφίας. Η έννοια του πανγονιδιώματος (pangenome) αναφέρεται στην ένωση οικογενειών γονιδίων που είναι κοινά ανάμεσα σε κάποια γονιδιώματα . Υπάρχει μια πληθώρα από μελέτες στις οποίες εφαρμόστηκε η ανάλυση του πανγονιδιώματος σε διάφορους οργανισμούς, από μικρόβια σε ιούς και φυτά. Οι μελέτες αυτές έχουν βοηθήσει στην προαγωγή γενικότερης κατανόησης σχετικά με τους εξελικτικούς μηχανισμούς, οδηγώντας σε πρακτική γνώση σε διάφορους τομείς όπως πχ. την υ υγεία, την φαρμακολογία και την γεωργία.

Ενώ υπάρχει μια ποικιλία εργαλείων που είναι διαθέσιμα για την διεξαγωγή μιας ανάλυσης πανγονιδιώματος, η πλειοψηφία αυτών έχει ως κύρια λειτουργία την μελέτη προκαρυωτικών γονιδιωμάτων. Στην παρούσα εργασία παρουσιάζεται ένα λογισμικό γραμμένο στην στατιστική προγραμματιστική γλώσσα R, που ονομάζεται pasaR, το οποίο μπορεί να χρησιμοποιηθεί στα τελευταία στάδια μιας τέτοιας ανάλυσης, δηλαδή μετά την κατασκευή των οικογενειών των γονιδίων για κάποια γονιδιώματα. Προτείνεται μια πλήρης μεθοδολογία για την ανάλυση γονιδιακών δεδομένων διαφορετικής πολυπλοκότητας, βελτιστοποιώντας και εμπλουτίζοντας ήδη υπάρχοντα εργαλεία από το πακέτο micropan, το μοναδικό αντίστοιχο πακέτο διαθέσιμο για την γλώσσα R. Επιπλέον προτείνεται μια καινούργια τεχνική η οποία χρησιμοποιεί την απόσταση Sorensen, γνωστή και ως ρευστότητα (fluidity) στο πλαίσιο της ανάλυσης πανγονιδιώματος, με στόχο την αναγνώριση διακριτών υποομάδων γονιδιωμάτων μέσα σε δοσμένο σύνολο δεδομένων. Τέλος εφαρμόζεται η μεθοδολογία αυτή σε δημόσια διαθέσιμα δεδομένα από τις βάσεις UniProt και Ensembl.

Λέξεις κλειδιά: Πανγονιδίωμα. Γονιδιακή ποικοιλομορφία, συγκριτικά genomics. R statistical language

Contents

Table of figures

Table of Tables

Acknowledgments

I would like to thank all my supervisors for the time and energy invested in this work, with a special mention to Fotis Psomopoulos without whose valuable contributions, remarks and general engagement the result would surely be of lesser quality. Last but not the least, I would like to thank my friends & family for their continuous support.

1. Introduction

The term pangenome is fairly new, being introduced by Tettelin et al. in 2005 (H. Tettelin et al. 2005) to describe a method of comparatively analyzing genetic data of different strains of the Streptococcus Agalactiae microbe in order to explore variability between them and most importantly trying to answer if there is a way to determine the number of genomes that must be sequenced in order to have a full genetic description of a bacterial species. Since then, this concept was applied by various other researchers, whose work ranges from species to phylum level (Vernikos et al. 2015), since it involves a more complete and dynamic way of handling genomic data as opposed to more linear approaches that have been used in the past (The Computational Pan-genomics Consortium et al. 2016, Lapierre and Gogarten (2009)).

The development of high throughput sequencing, vastly increased the genomic data available to researchers and provided new insights in the evolution and physiology mechanisms of various species (Muzzi, Masignani, and Rappuoli 2007). This plethora of data is enabling the pangenome analysis: While early works mainly involved prokaryotic species, with more than forty (40) studies existing for bacterial pangenomes (Rouli et al. 2015), the last years, pangenomic studies have expanded to agronomic plants such as maize (Hirsch et al. 2014), rice (Sun et al. 2017) and soybean (Y.-h. Li et al. 2014), eukaryote microorganisms such as phytoplankton Emiliania (Read et al. 2013) and research in the direction of finding a human pan-genome has been conducted (Li et al. 2010).

In this thesis, a new software package for the R language for statistical computing (R Core Development Team 2016) is presented and showcased on various datasets. This software, named pasaR, is usable for the last stages of a pangenomic analysis, that is after the genomes that are analyzed have been sequenced and the gene families found have been clustered, allowing both exploration of the data and its statistical analysis. Some of the functions used in pasaR are based on existing ones from the only other complete package for pangenomic analysis in R micropan (Snipen & Liland, 2015), however they have been tweaked to optimize for speed. Comparison results are available in appendix: benchmarking.

1.1 Components of a pangenome

Pangenome, also known as supragenome (Tettelin et al. 2008) as a concept refers to the union of gene families shared by a number of genomes of a grouping of organisms (Lapierre and Gogarten 2009). Other more limiting definitions include the one given by Vernikos et al. (Vernikos et al. 2015) i.e. "the entire genomic repertoire of a given phylogenetic clade and encodes for all possible lifestyles carried out by its organisms" or the one by McInerney et al. (McInerney, McNally, and O’Connell 2017) "the collection of gene families that are found to be present in all members of a particular species".

A pangenome consists of three parts (H. Tettelin et al. 2005,Lapierre and Gogarten (2009),Carlos Guimaraes et al. (2015)) :

1. the Core genome consisting of all gene families that shared amongst all genomes examined
2. the Dispensable or Accessory genome consisting of genes present in some of the genomes
3. a subset of the Dispensable genome, genes present only in one genome, known as singletons or ORFans and might be species or strain specific

Other names and overlapping definitions are used in the literature, to describe the pangenome.

A pan-genome can be characterized as closed when as genome sample grows it's size approaches a constant number and open when new gene families are detected with every new genome sample (Golicz, Batley, and Edwards 2016). A closed pangenome is observed in species that exist in isolated and sparse ecological niches, while an open pangenome is a sign of flexible genetic content in the cases of the same species pangenome (Carlos Guimaraes et al. 2015) or a sign of genome diversity and non-coherence in the case where multiple species are examined. MCirney et al. offer a useful schematic representation of a pangenome.

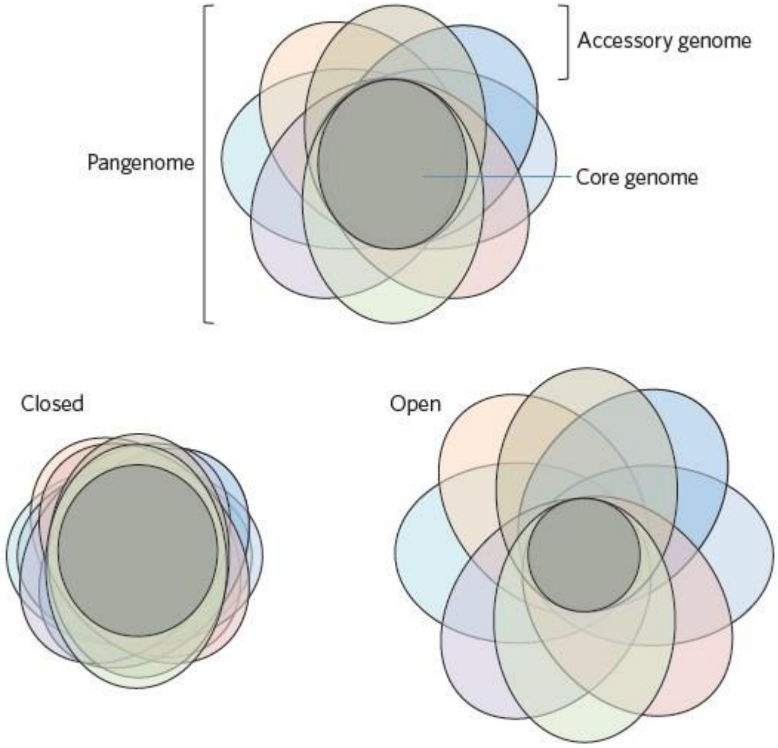


Figure 1 "Schematic representation of pangenomes as Venn diagramms" (McInerney, McNally, and O'Connell 2017)

1.2 Gene "homology"

The building block for a pangenomic analysis, is a dataset of clustered gene families from some genomes of interest. The first step to produce such a dataset is the identification of homologous sequences between the genomes using tools such as BLAST, FASTA or HMMR3 (Pearson 2013) and computing in a pairwise manner a similarity measure between all sequences of interest. Another technique, faster than the identification of all similarities but less accurate (Dalquen and Dessimoz 2013), is the bidirectional best hit (BBH) where only best matched pairs of genes are kept in the results. The final step is clustering the homologue genes, using algorithms such as Marcovian Cluster Algorithm (MCL) or CFinder (Rhee and Mutwil 2014). A lack of a community standard for the dataset construction for a pangenomic analysis should be noted.

1.3 Relevant software review

The increased interest on the pan-genome, has led to the availabity of various software for various aspects of such an analysis. These programs, written in various programming languages, are either stand-alone or complimentary to existing ones and the majority is aimed for prokaryote organisms. The results of an extensive literature review on pangenomic related software published until 2016 is presented on table 1. A software is characterized as a full pipeline when it provides functionalities than cover all steps from ortholog detection to pangenomic analysis results.

Table 1 Software for pangenomic analysis

Title	Species	Methods	Standalone
EDGAR (Blom et. al, 2009)	Prokaryote	Full pipeline: Core & pangenome size analysis	Yes
PanCGHweb (Bayjanov et. al, 2010)	Bacteria	Genotyping through pangenome data	Yes - online
CAMBer (Wozniak,Wong & Tiryun, 2011)	Bacteria	Core, accessory genome analysis and pangenome size	Yes
Panseq (Laing et. al, 2011)	Bacteria	Core and accessory genome analysis	Yes, online & offline versions
PGAT (Brittnacher et. al, 2011)	Bacteria	Ortholog prediction & Presence/absence gene analysis	Yes - online
PGAP (Zhao et al., 2011)	Bacteria	Full pipeline: pangenome profile analysis & exponential fit	Yes
PanDaTox (Amitai & Sorek, 2012)	E. Coli	Specific app mainly to investigate toxicity of organisms to E. Coli using pangenome analysis	Yes - online
PANNOTATOR (Santos et. Al, 2013)	Bacteria	Pipeline for pangenome annotation but not analysis	Yes - online
Pancake (Ernst&Rahmann,2013)		Full pipeline for pangenome exploration implemented through pooling similar genomic subsequence	Yes
GET_HOLOGUES (Conteras & Vinueas, 2013)	Microbes	Full pipeline: tools for pangenome creation, overview & statistical analysis	Yes
eCAMBer (Wozniak,Wong & Tiryun, 2014)	Bacteria	Core, accessory genome analysis and pangenome size	Yes - online

Table 2 (continued) Software for pangenomic analysis

Title	Species	Methods	Standalone
ITEP (Benedict et al., 2014)	Microbes	Full pipeline: pangenome analysis & exploration of subsets of interest	Yes, python/BASH scripts also available
SplitMem (Marcus et al., 2014)	N/A	de Brujjin Graphs pangenome representation	Yes
PanGP (Zhao et al., 2014)	Bacteria	Full pipeline: pangenome profile analysis with sampling capabilities	Yes
Spine/AGEnt/ClustAGE (Ozer, 2014)	Bacteria	Full pipeline for core and accessory genome detection and annotation	Yes - online
Harverst (Trangen et al., 2014)	Microbes	Full pipeline for core genome alignment and visualization	Yes
Roary (Page et al., 2015)	Prokaryote	BLASTP and MCL, graphs for cluster relationships	No (Perl)
Pan-tetris (Hennig, Bernhart & Niselt 2015)	Microbes	"Super-genome" based alighnment and visualization	Yes
Micropan (Snipen & Liland, 2015)	Microbes	Full pipeline: tools for pangenome creation, overview & statistical analysis	No (R)
BFT (Holley, Wittler & Stoye, 2016)	-	Indexing scheme of the pangenome through de brujin graphs	Yes
PanTools (Sheikhizadeh et. al, 2016)	Microbes	Full pipeline: De Brujjin Graphs & pangenome comparison	Yes - online
PanX (Ding, Baumdicker & Neher, 2016)	Bacteria	Full pipeline: Analysis & Visualization	Yes - online

1.4 Software development perspectives

One of the main outcomes of this thesis is a software package for the R language. The package, named pasaR is open source software licensed under the GNU general public license v3.0 (Free Software Foundation 2007), meaning that the user has access to all source code of the software and can use, modify or distribute the package at will. The software is available through the popular (Perez-Riverol et al. 2016) code repository service Git Hub, in the address https://github.com/ampatzia/pasaR . All tools available through the package where written using the style guide proposed by Wickham (2015), are documented using standard R procedure and a minimal reproducible example of use is provided.

The decisions presented above offer several positive traits:

a) R as a scripting language that has all components available for free, enables fast reproducible results,

b) The combination of public availability of the code and the explicit permission of modification enables community accessibility, evaluation and opportunities of scientific collaboration or reuse

c) The software is ready to use, with minimum setup required from the user

d) Repository services such as Git Hub, offer automated software versioning, collaboration and issues tracking tools thus allowing existing users of the software a clear overview of the status of the software, recent changes made

The practices presented comply with a number of recommendations from researchers and software developers that promote open science and software instead of more traditional approaches (McKiernan et al. 2016; Jiménez et al. 2017).

2. Theory and Methods

In this chapter, mathematical tools that will be used in the analysis part of the thesis, are extensively presented along with relevant proofs and details of usage.

2.1 Heap's Law.

In their seminal work Tettelin et al. (Tettelin et al. 2008), proposed the utilization of a power law model for the estimation of the pan-genome size, replacing the exponential decay model used up to that point. In many natural cases, an attribute *n* grows in concurrence to a power law of the number *N* of the objects under examination, something that can be expressed as $n \sim N^{\gamma}$, $0 < \gamma < 1$. This empirical law has been used in various scientific areas and in the context of pan-genomics, originated from linguistics information retrieval (Heaps 1978), where it is known as Heaps' and to a lesser extend Herden - Heaps' Law. Examples of quantities that follow a power law include word frequencies inside big bodies of text, populations of cities and the magnitude of earthquakes (Newman 2004).

Two parameters are needed to describe a power law, the exponent written as γ or more commonly as $\alpha = 1 - \gamma$ and a constant k. Then the power law can be written as,

$$\begin{aligned} n = k * N^{(\gamma-1)} = k * N^{-a}, \\ 0 < \gamma < 1, \\ \alpha = 1 - \gamma \end{aligned}$$

describing the number *n* of genes i.e. the pan-genome, of *N* genomes. It follows from this equation, that the number of genes observed are decreasing, as the number of genomes sampled increase. For $\alpha > 1$ ($\gamma < 0$), *n* diverges to a constant as *N* increase and we call the pan-genome **closed**. For $\alpha \geq 1$, *n* is not bounded and increases as *N* increases and the pan-genome is **open** .

2.2 Chao lower bound estimator.

A common problem faced in field biology is the estimation of the size of an organism population based on data gathered by observing or capturing members of that population in different locations also known as sites. Considering genomes as the components of population, genes the "individuals" and genome clusters the sites one can use ecological tools in order to find the pan-genome size. One such estimator, using the method of moments, was proposed by Chao (Chao 1987) and can be used on pan-genomic analyses as it provides a conservative population estimation (Snipen, Almøy, and Ussery 2009), suitable even for data where species are observed with unequal probability. The proof of the estimator as given from Chao is presented below.

Consider the Pan-genome size of a number of genomes to be N, composed by $i = 1,2,..,N$ genes and *j= 1,2,...,t* clusters. Let p_{ij} be the probability of gene *i* to participate in cluster t and assume that $p_{ij} = p_i$ for $j = 1,..t$ and p_i for $i = 1,...,N$ are sampled from a probability distribution F. The population could be written as a $X = (Xij)$ matrix sized *(N,t)*, with $X_{ij} = [i \in j]$ using the Iverson Bracket. It is also assumed that X_{ij}, i and j are mutually independent. The number of distinct observations is denoted by

$$S = \sum_{i=1}^{N} [\sum_{j=1}^{t} X_{ij} \geq 1]$$

and the number of individuals observed exactly k times in t clusters is

$$f_k == \sum_{i=1}^{N} [\sum_{j=1}^{t} X_{ij} = k], k = 0,1,...,t$$

If f_0 are the number of unobserved genomes, then it follows that the pan-genome size equals $N = S + f_0$. The number of observations $(f_0, f_1, ..., f_t)$, have a multinomial joint, unconditional distribution function:

$$p(f_0, f_1, \dots, f_t) = \binom{N}{f_0\ f_1\ \dots\ f_t} \prod_{i=0}^{t} [\,\theta(F)]_i^f \quad (1)$$

where

$$\theta_i(F) = \int_0^1 \binom{t}{i} p^i (1-p)^{t-i} dF(p)$$

From (1), Chao provides the following estimator

$$E(f_i) = N \int_0^1 \binom{t}{i} p^i (1-p)^{t-i} dF(p), \quad i = 0,1,\dots,t \quad (2)$$

which for a sufficiently large t and small p can be rewritten as

$$E(f_i) \approx N \int_0^1 \frac{(tp)^i\ e^{-tp}}{i!} dF(p), \quad i = 0,1,\dots,t \quad (3)$$

Considering a cumulative distribution in the $[0, t]$ space

$$G(u) = \frac{\int_0^u x\, e^{-x} dF(\frac{x}{t})}{\int_0^t x\, e^{-x} dF(\frac{x}{t})}$$

and combining it with (3), we get

$$E(f_0) \approx N \int_0^1 e^{-tp}\, df(p) \approx E(f1) \int_0^1 u^{-1}\, dG(u) \quad (5)$$

Consequently, the kth moment of G, m_k is,

$$\mu_k = \int_0^t u^k\, dG(u) = \frac{\int_0^1 (\,tx)^{k+1} e^{-ix} dF(x)}{\int_0^1 (\,tx)^{-ix} dF(x)} \approx (k+1)! \frac{E(f_{k+1})}{E(f_1)} \quad (6)$$

$E(f_i)$ can be replaced by f_i, to obtain an estimator of m_k (if $f_1 \neq 0$)

$$m_k = (k+1)! \frac{f_{k+1}}{f_1}$$

Combining (5) and (6), and Jensen's inequality

$$E(f_0) \geq \frac{E(f_1)}{\mu_1} = \frac{E(f_1^2)}{2E(f_2)}$$

Thus, a lower bound of the population size $\hat{N}$ is:

$$\hat{N} = S + \frac{f_1^2}{2f_2}$$

An approximation formula of the lower bound with would be:

$$\hat{N}_{min} \approx \hat{N}_1 = S + \frac{f_1^2}{2f_2}(\frac{\frac{1-m_1}{t}}{\frac{1-m_2}{tm_1}})$$

An asymptotic variance estimator of the above quantity is

$$\hat{\sigma}^2 = f_2((\frac{1}{4})(\frac{f_1}{f_2})^4 + (\frac{f_1}{f_2})^3 + \frac{1}{2}(\frac{f_1}{f_2})^2)$$

with a 95% confidence interval of

$$[S + \frac{(\hat{N} - S)}{C}, S + (\hat{N} - S)C]$$

where

$$C = exp(1.96(log(\frac{1 + \hat{\sigma}^2}{(\hat{N} - S)^2}))^{\frac{1}{2}})$$

treating $log(\hat{N} - S)$ as an approximately normal random variable.

2.3 Binomial Mixture model for pan genome size estimation

Let the pan-genome size of sample number of a grouping of genomes, composed by $i = 1,2,..,G$ genes and *j= 1,2,...,t* clusters, to be N and $\hat{N}$ be the true pan-genome size, that resulting from all genomes of the grouped organisms, already included or not. If γ_0 is the unobserved number of gene families existing in all genomes, then it follows that $\hat{N} = N + \gamma_0$ and it is clear that an estimation of γ_0 allows the estimation of $\hat{N}$.

Snipen, Almøy and Ussery (2009), propose a model that relates γ_0 to the sum of gene families $\gamma_1, \gamma_2, \ldots, \gamma_G$ of each genome present in the families. Considering the population pan-genome size $\hat{N}$ as constant and assuming independence between gene families, allows the consideration of $\gamma = \{\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_G\}$ as a multinomial vector. Let $\theta = \{\theta_0, \theta_1, \theta_2, \ldots, \theta_G\}$ be the multinomial probabilities of detecting a gene family in 0, ... G genomes. Using these assumptions, the expected value of γ_0 is,

$$E(\gamma_0) = \hat{N}\,\theta_0$$

and

$$E(N) = \hat{N}\,(1 - \theta_0),$$

which can be combined to

$$E(\gamma_0) = E(N)\frac{\theta_0}{1 - \theta_0}$$

which can be simplified, by using N instead of E(N), to:

$$E(\gamma_0) = N\frac{\theta_0}{1 - \theta_0} \quad (1)$$

Consequently, by estimating θ_0, the value of γ_0 can be computed through (1). Assuming a degree of smoothness over the probability distribution and using a binomial mixture model (Hand,1989) we can continue with the estimation of θ.This model is composed by

$$\theta_g = \sum_{k=1}^{K} \pi_k \, f(g; p_k), \;\; g = 0, \ldots, G \quad (2)$$

with π_k called the mixing proportion and the binomial probability mass function (PMF) of detection probability p_k

$$f(g; \pi_k) = \binom{G}{g} p_k^g (1 - p_k)^{G-g}, k = 1, \ldots, K \quad (3)$$

It should be noted that $\sum_{k=1}^{K} \pi_k$ is always one, also an assumption that $p_1 = 1$, i.e. there is always a core genome presented, is applied. We end up with a model where the aforementioned multinomial distribution is explained by K binomial PMFs. Hence, the next step is to estimate the parameters of these PMFs, something that can be accomplished by maximizing the following zero-truncated log-likelihood function:

$$l(p, \pi | k) = \sum_{g=1}^{G} \gamma_g \, log(\frac{\theta_g}{1 - \theta_0}) + C \quad (4)$$

where $p_g = \frac{\theta_g}{1-\theta_0}$ is the probability of an element of $g = 1, \ldots, G$ from the multinomial vector $\gamma_+ = (\gamma_1 + \gamma_2 + \ldots + \gamma_G)$ over a fixed N, $\theta_0, \ldots, \theta_G$ are dependent on $\pi \; and \; p$ as described in (2,3) and C is a independent constant. For arbitrary choices of K and maximizing (4), $k = 1, \ldots, K$ estimations of p and π occur which can be denoted p_k and π_k; these can then be used in equations one to three, in reverse order, to get a prediction of $\hat{\gamma}_0$. Finally, the optimal number of components is determined by choosing the minimal value of the Bayesian Information Criterion (BIC)

$$BIC(K) = 2(k - 1) logN - 2l(\pi, k | K)$$

The number of free parameters, differs to the formal by one due to the assumption of $p_1 = 1$. These computations lead to the desired results: the pan-genome size $\hat{N} = N + \hat{\gamma}_0$ and the core size $\hat{\gamma} = \hat{N} \, \hat{\pi}_1$.

2.4 Genomic Fluidity

Genomic Fluidity was proposed by Kislyuk, Haegernan, Bergman & Weitz (2011) as a more robust alternative to core and pan genome size estimation techniques to assess the similarity of a group of genomes.

The genomic fluidity of a group of N genomes is:

$$\varphi = \frac{2}{N(N-1)} \sum_{k,l=1\ldots N} \frac{U_k + U_l}{M_k + M_l}, \quad k < l$$

with U_k,U_l are the number of gene families that are respectively unique in the k,l genomes and M_k,M_l all the gene families in the k,l genomes. In plain language, it is the average of sum of the unique gene families between pairs of genomes divided by the sum of all gene families between pairs of genomes. A useful aspect of the genomic fluidity is its intuitive interpretation: A group with φ =0.2, has on average 80% shared genes and 20% of the genes are unique.

In other contexts, the same measure is known as Sorensen distance (M. M. Deza and Deza 2009) and before averaging is a "true" mathematical distance.

2.5 Hierarchical Clustering using Fluidity as a distance

As the Sorensen distance is a distance function, it can be used as measure of cluster proximity in the sense of agglomerative or also known as bottom-up hierarchical clustering. While the process of hierarchical clustering can be considered trivial or common knowledge it will be briefly presented for the sake of completeness: Let the number of genomes under study be N with $\phi_i j$ the complete set of all pairwise fluidity values for these genomes, $\frac{n(n-1)}{2}$ in number. In the beginning of the process, all genomes are considered individual clusters. Then using a linkage method, for example Average or Complete, the least dissimilar genomes are found and fused, resulting to $N-1$ clusters and the dissimilarities between the remaining clusters is recalculated.

This process is repeated until N genomes are pooled in one (1) cluster (James et al. 2007). In the present work, the Ward method of linkage known also as Ward2 is used which produces clusters by minimizing intra-cluster variance. While this method was originally developed to be used with Euclidean distances, it can be also used with other distance metrics (Murtagh and Legendre 2014, Miyamoto et al. (2016)).

There are numerous indices that can be used to determine the optimal number of clusters (Kovács, Legány, and Babos 2005). In the context of the pan-genome study, non-exhaustive empirical evidence, in our experiments, shows that the gap statistic (Tibshirani, Walther, and Hastie 2001) and the Dunn index (Dunn 1973) produces adequate splits of the data, however the Silhouettes Index (Rousseeuw 1987) and the Variation of Information Index (Meila 2007) are provided.

3. Analysis

In this chapter, a pangenomic analysis workflow, based on the theoretical tools introduced in chapter two, is proposed. The workflow is implemented via the pasaR R package, on five (5) datasets of various genomic sizes.

3.1 Proposed workflow

The input of the workflow is datasets containing clustered gene families from different genomes, for example the output of a sequence clustering pipeline with the default settings (Blast and MCL). First the sample core genome, pangenome size and number of orfan genes should be computed to assess the dataset quality and genome coherence. Further insights can be achieved through visualizing cluster spread for genomes, genome participation per cluster and gene participation per cluster.

Continuing, the openness of the pangenome can be evaluated using Heap's Law and then estimate the actual pangenome size using the Chao estimator or Binomial Mixtures. Binomial mixtures models also provide estimates about the core genome size and the number of underlying components that compose the pangenome, the component mixture probabilities and corresponding detection probabilities.

Finally, either the whole sample or an estimation of the fluidity, with the use of the first one recommended, can be computed. If the fluidity score is smaller than an arbitrarily chosen threshold, then the user can choose to use agglomerative clustering based on fluidity, in order to determine existent coherent subsets of dataset. The process is depicted in the picture below.

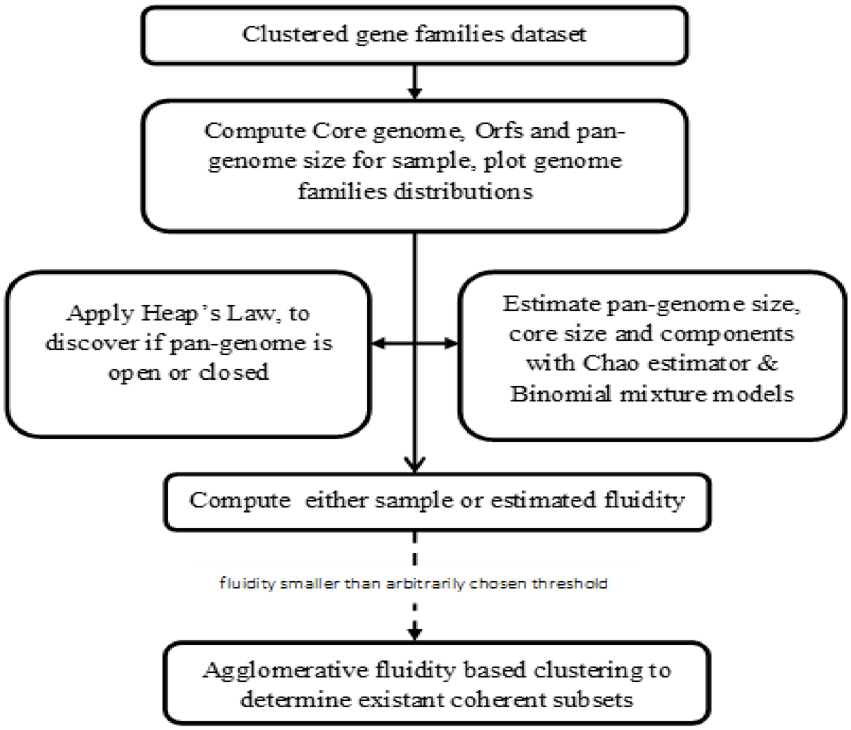

Clustered gene families dataset
Compute Core genome, Orfs and pan-genome size for sample, plot genome families distributions
Apply Heap's Law, to discover if pan-genome is open or closed
Estimate pan-genome size, core size and components with Chao estimator & Binomial mixture models
Compute either sample or estimated fluidity
fluidity smaller than arbitrarily chosen threshold
Agglomerative fluidity based clustering to determine existant coherent subsets

Figure 2 Proposed analysis flow

3.2 Dataset summary

Five (5) different datasets with different sizes from various sources ares used:

1. The first dataset was made publicly available by the authors of the **R** package *micropan* (Snipen and Liland 2017) and contains seven (7) strains of the Mycoplasma Pneumoniae bacteria. It provides a good example of a small, closed bacterial pangenome.

2. The second dataset was made using publicly available data from Ensembl (Aken et al. 2016), and contains eighty one (81) strains in total, distributed across the following four (4) bacterial species: twelve (12) strains of Streptococcus pneumoniae, thirteen (13) of Streptococcus Pyogenes, thirty nine (39) of Bacillus cereus and seventeen (17) of Bacillus thuringiensis. It was produced by a standard clustering pipeline with the default settings (BLAST and MCL).

3. The third dataset made from the same genomic data used in the second, however only the best bidirectional hits where kept during the homology detection.

4. The fourth dataset was made using publicly available data from the Ensembl database, and contains twenty-two (22) strains from three distinct groups: twelve strains (12) of Streptococcus pneumoniae, six (6) strains of the Buchnera Aphidicola proteobacteria, four species (4) of the Pyrococcus Genus p. abyssi, p. furiosus, p horikoshii and p. kodakarensis. It was produced by a standard clustering pipeline with the default settings (BLAST and MCL).

5. The last dataset was constructed using publicly available data from UniProt (The Uniprot Consortium 2017), Plaza (Proost et al. 2015) and Pico-Plaza (Vandepoele et al. 2013) and contains ninety five (95) genomes of organisms with photosynthetic abilities. It was originally presented in (Psomopoulos, Kintsakis, and Mitkas 2016)

Frequency tables of Genome participation in clusters for all the datasets are available on the appendix.

3.3 Mycoplasma pneumoniae

The sample pangenome size of the Mycoplasma Pneumoniae Genomes is 1210 gene families, the sample core size is 1100 gene families and the number of orfan genes is 33. Fitting the Mycoplasma Genomes according to Heap's Law, results to the estimation of a closed pangenome a = 1.42766, with an intercept of k = 59.59479. Using the Chao estimator, a pangenome size of $n = 1258$ C.I. 95% = (1212,2456), with variance of $s^2 = 710.30025$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 3 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.0742364	0.6652024	1.0000000
Mixing.prop	0.0779073	0.0576295	0.8644632

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
3 components	1096	1268	1143.909

Finally the sample fluidity is $\phi = 0.0199089$ with $s = 0.006602$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.0200534$ with $\sigma = 0.0064887$.

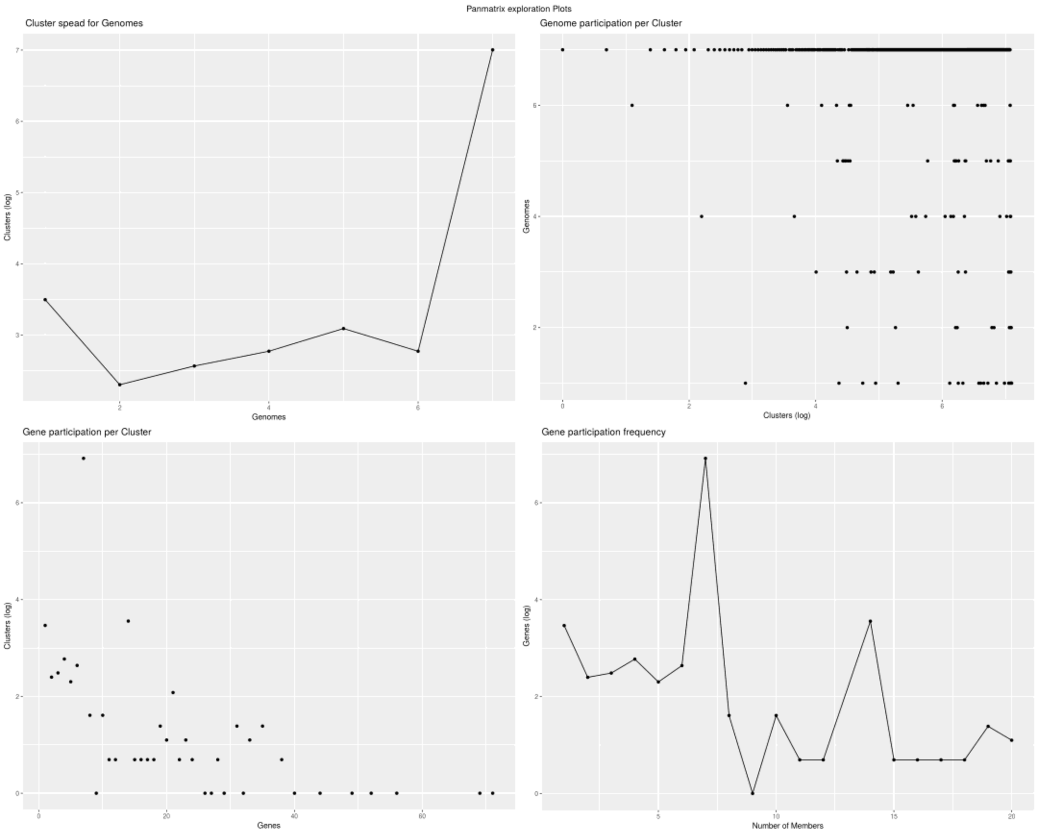


Figure 3 Summary plots and information for the sample Mycoplasma Pneumoniae pangenome

3.4 Four Species dataset

The sample pangenome size of the four (4) species genome collection is 149721 gene families, no sample core is found and the number of orfan genes is 93296. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.23614$, with an intercept of $k = 3989.0162$. Using the Chao estimator, a pangenome size of $n = 338349$ C.I. 95% = (156401,5475997), with variance of $s^2 = 3256678.02761$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 9 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.008268	0.07223	0.2384
Mixing.prop	0.9199	0.06976	0.00839

	Comp_4	Comp_5	Comp_6
Detection.prob	0.4696	0.6578	0.8015
Mixing.prop	0.001377	0.00044	0.00008865

	Comp_7	Comp_8	Comp_9
Detection.prob	0.8585	0.954	1
Mixing.prop	0.0000001162	0.00000004843	0.0000006824

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
9 components	0	282357	435058

Finally the sample fluidity is $\phi = 0.9162145$ with $s = 0.1092225$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.9146962$ with $\sigma = 0.1089366$.

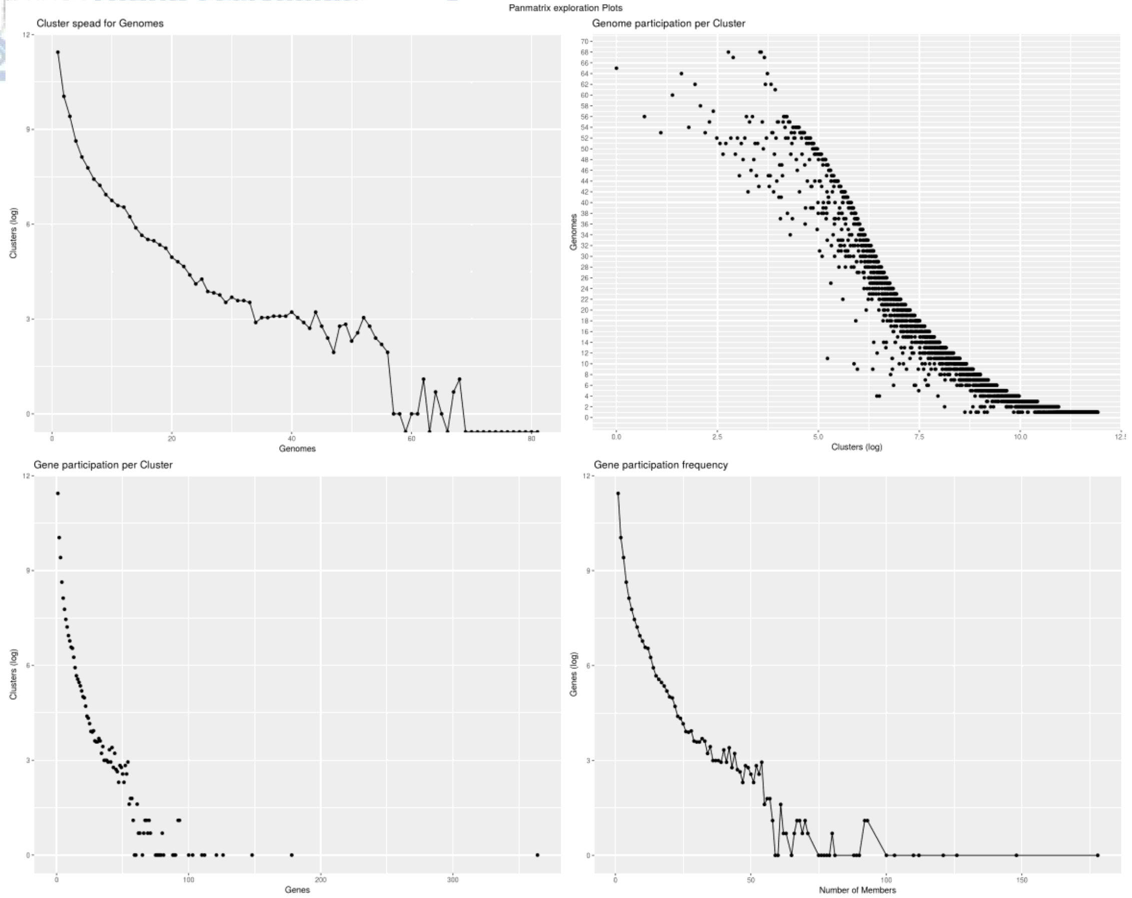


Figure 4 Summary plots and information for the 4 species dataset, as produced by a standard sequence clustering pipeline with default settings (BLAST and MCL)

3.4.1 Bacilus Cereus

The dataset consists of thirty nine (39) strains of a single species i.e. Bacilus Cereus.The sample pangenome size of the genome collection is 102964 gene families, the sample core is found to be 10 and the number of orfan genes is 67714. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a =$ 0.22679, with an intercept of $k =$ 4702.29435. Using the Chao estimator, a pangenome size of $n =$ 242660 C.I. 95% = (107835,4109459), with variance of s^2= 2482078.26512 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 8 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7	Comp_8
Detection.prob	0.01044	0.08666	0.2731	0.5959	0.8615	0.9682	0.9986	1
Mixing.prop	0.9117	0.0744	0.01153	0.001826	0.0004516	0.0001111	0.00000176	0.0000001573

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
8 components	0	262445	261593

Finally the sample fluidity is $\phi = 0.8783143$ with $s = 0.102614$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.8775756$ with $\sigma = 0.1086773$.

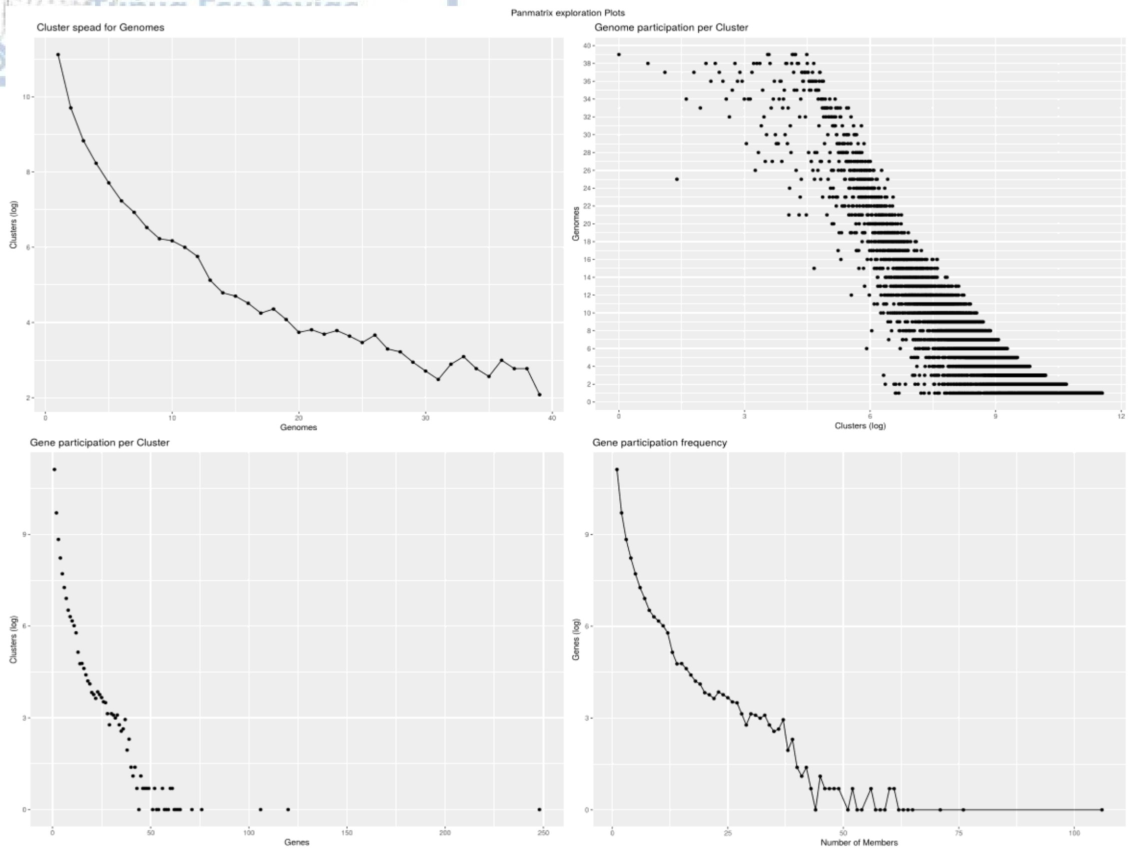


Figure 5 Summary plots and information for the Bacillus Cereus sample, as produced by a standard sequence clustering pipeline with default settings (BLAST and MCL)

3.4.2 Bacillus thuringiensis

The dataset consists of seventeen (17) strains of a single species, i.e. of Bacillus thuringiensis. The sample pangenome size of the genome collection is 53243 gene families, sample core size is discovered to be 91 and the number of orfan genes is 35740. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.24956$, with an intercept of $k = 4837.87588$. Using the Chao estimator, a pangenome size of $n = 148494$ C.I. 95% = (55963,3388574), with variance of $s^2 = 2464843.37952$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 5 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.001084	0.1189	0.3849
Mixing.prop	0.9842	0.01399	0.001565

	Comp_4	Comp_5
Detection.prob	0.8321	1
Mixing.prop	0.0001726	0.00006187

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
5 components	102	1655919	127952

Finally the sample fluidity is $\phi = 0.8276564$ with $s = 0.1399459$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.8243079$ with $\sigma = 0.1468341$.

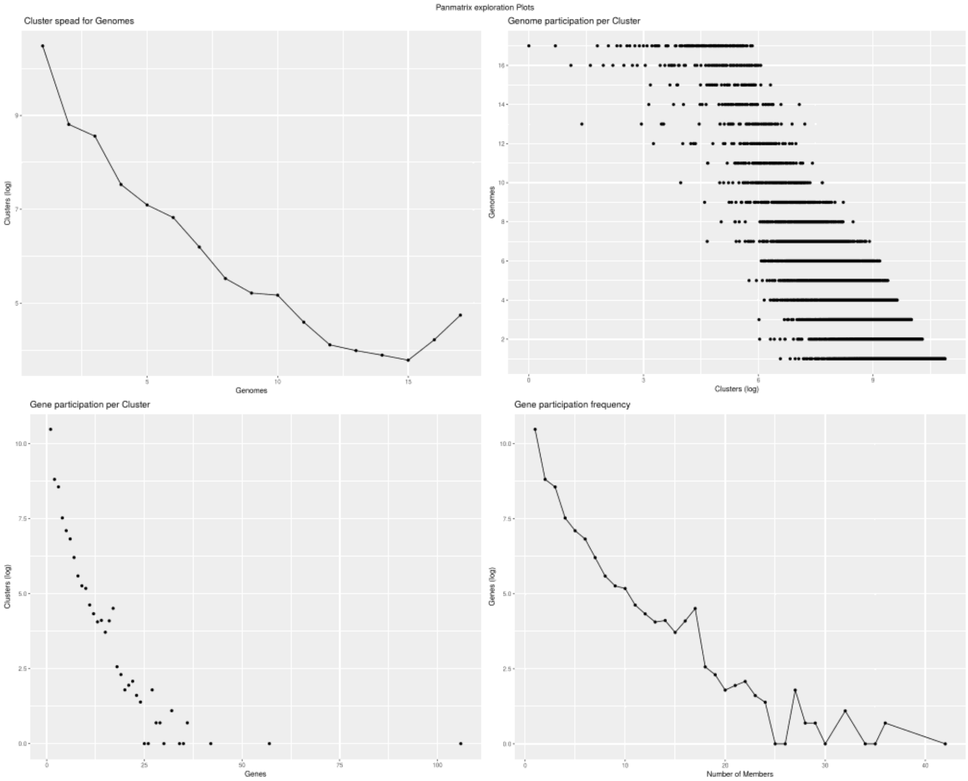


Figure 6 Summary plots and information for the Bacillus Thurigensis sample, as produced by a standard sequence clustering pipeline with default settings (BLAST and MCL)

3.4.3 Streptococcus pneumoniae

The dataset consists of twelve (12) strains of a single species i.e., Streptococcus pneumoniae. The sample pangenome size of the genome collection is 10331 gene families, sample core size is 145 and the number of orfan genes is 5659. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.56344$, with an intercept of $k = 1995.30055$. Using the Chao estimator, a pangenome size of $n = 17854$ C.I. 95% = (10695,165629), with variance of $s^2 = 74229.37295$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 5 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5
Detection.prob	0.05686	0.2867	0.6145	0.9113	1
Mixing.prop	0.8568	0.08809	0.03544	0.01729	0.002334

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
5 components	42	17998	31306

Finally the sample fluidity is $\phi = 0.6473961$ with $s = 0.0946125$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.6467789$ with $\sigma = 0.0945409$.

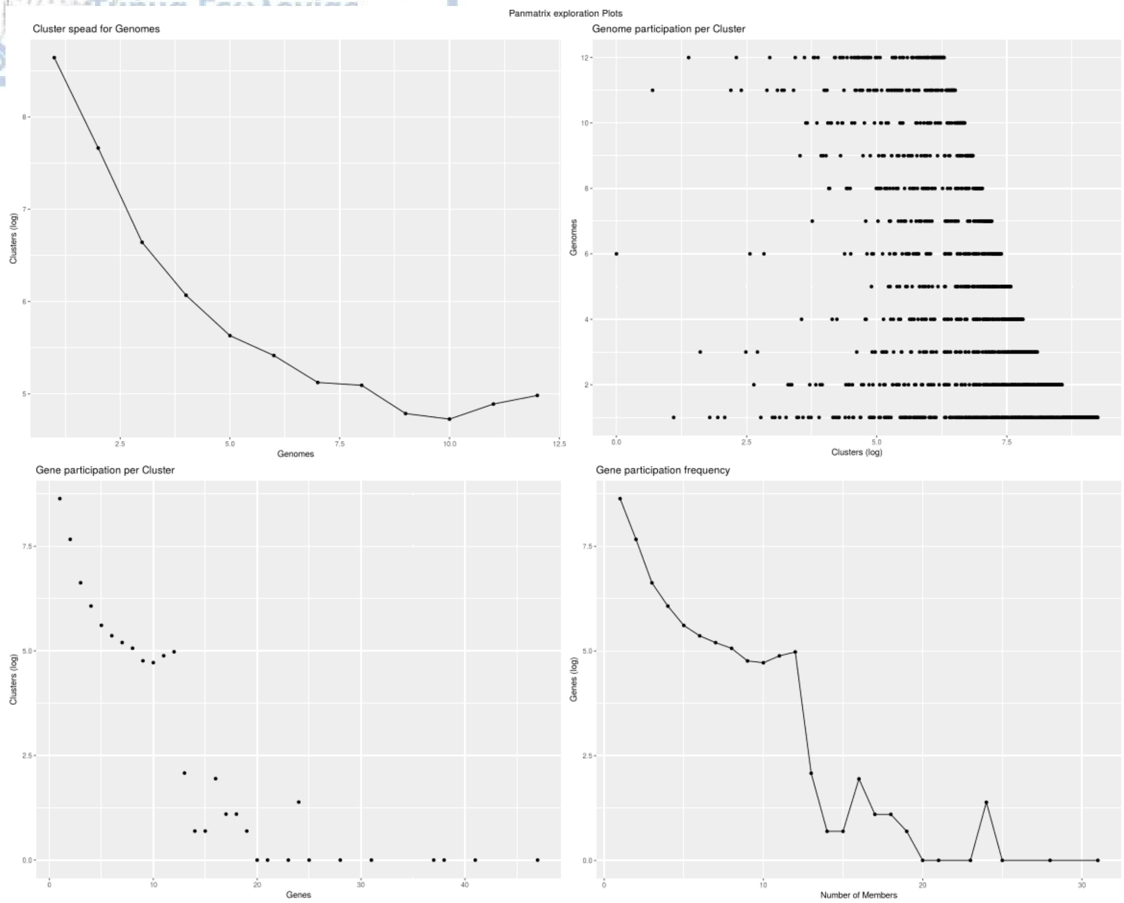


Figure 7 Summary plots and information for the Streptococcus Pneumoniae sample, as produced by a standard sequence clustering pipeline with default settings (BLAST and MCL)

3.4.4 Streptococcus Pyogenes

The dataset consists of twelve strains (12) of a single species, i.e. Streptococcus pneumoniae. The sample pangenome size of the genome collection is 9952 gene families, sample core genome is 103 and the number of orfan genes is 5333. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.55862$, with an intercept of $k = 1839.18593$. Using the Chao estimator, a pangenome size of $n = 16288$ C.I. 95% = (10285,130620), with variance of $s^2 = 54407.78325$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 5 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.05937	0.2896	0.6279
Mixing.prop	0.8673	0.08618	0.02648

	Comp_4	Comp_5
Detection.prob	0.8807	1
Mixing.prop	0.01597	0.004077

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
5 components	67	16379	30178

Finally the sample fluidity is $\phi = 0.6659675$ with $s = 0.1054033$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.6604302$ with $\sigma = 0.1132254$.

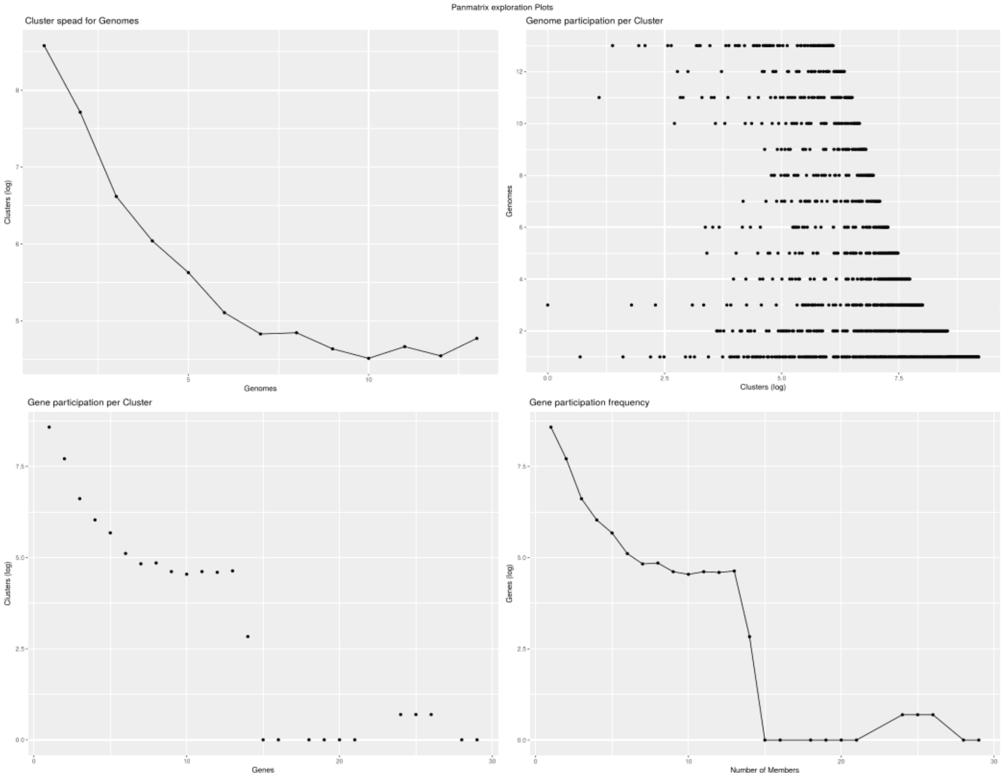


Figure 8 Summary plots and information for the Streptococcus Pyogenes sample, as produced by a standard sequence clustering pipeline with default settings (BLAST and MCL)

3.5 Four species dataset with best bi-directional hits

The sample pangenome size of the four species variant bacterial genome collection consists of 185188 gene families and the number of orfan genes consists of 128528 genes, while the sample core genome was found to be zero.

Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.23614$, with an intercept of $k = 3989.0162$. Using the Chao estimator, a pangenome size of $n = 495538$ C.I. 95% = (194765, 10242219), with variance of $s^2 = 6927692.13345$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 11 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.008195	0.09414	0.3215
Mixing.prop	0.9577	0.0386	0.002725

	Comp_4	Comp_5	Comp_6
Detection.prob	0.4947	0.6099	0.6513
Mixing.prop	0.0003989	0.0005557	0.00002408

	Comp_7	Comp_8	Comp_9
Detection.prob	0.673	0.8367	0.8972
Mixing.prop	0.000001487	0.000004839	0.000001412

	Comp_10	Comp_11
Detection.prob	0.9481	1
Mixing.prop	0.000001056	0.000001654

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
11 components	1	364366	447447

Finally the sample fluidity is $\phi = 0.9374495$ with $s = 0.1027686$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.9374038$ with $\sigma = 0.1049131$.

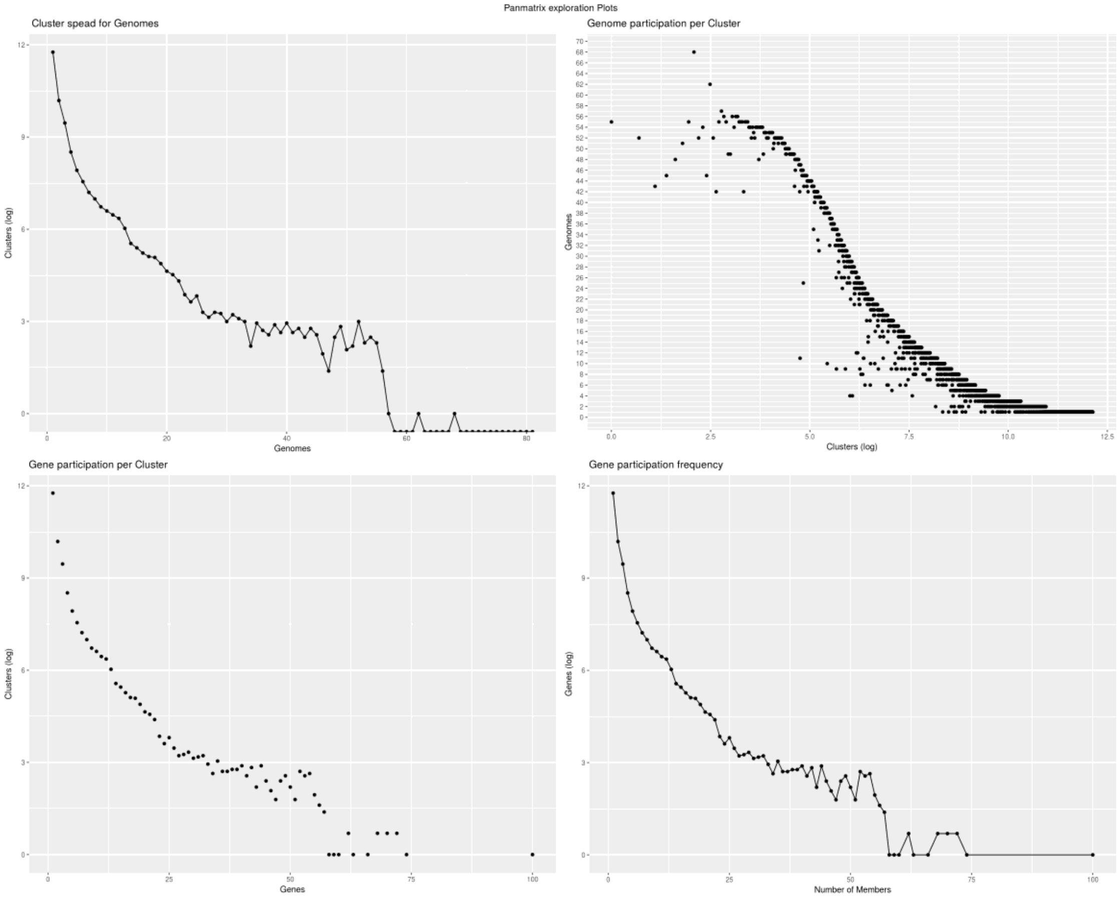


Figure 9 Summary plots and information for the four bacterial species dataset, produced by a standard sequence clustering pipeline with the default settings (Blast and MCL), but maintaining only the best bidirectional hits during the homology detection

3.5.1 Bacilus Cereus with best bi-directional hits

The dataset consists of the thirty-nine (39) strains of a single species i.e. Bacillus Cereus. The sample pangenome size of the genome collection is 124218 gene families, sample core was found to be 7, and the number of orfan genes is 90692. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $\alpha = 0.22679$, with an intercept of

$k =$ 4702.29435. Using the Chao estimator, a pangenome size of $n =$ 354004 C.I. 95% = (131047, 7856400), with variance of s^2= 5509793.80747 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 8 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.01038	0.1081	0.3855
Mixing.prop	0.9587	0.03676	0.003735

	Comp_4	Comp_5	Comp_6
Detection.prob	0.758	0.9295	0.9911
Mixing.prop	0.0005879	0.0001625	0.000009597

	Comp_7	Comp_8
Detection.prob	1	1
Mixing.prop	0.0000008876	0.000001278

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
8 components	0	343840	256962

Finally the sample fluidity is $\phi =$ 0.908099 with $s =$ 0.1013906, while a population estimate through the use of permutations gives $\hat{\phi} =$ 0.9099123 with $\sigma =$ 0.1028378 .

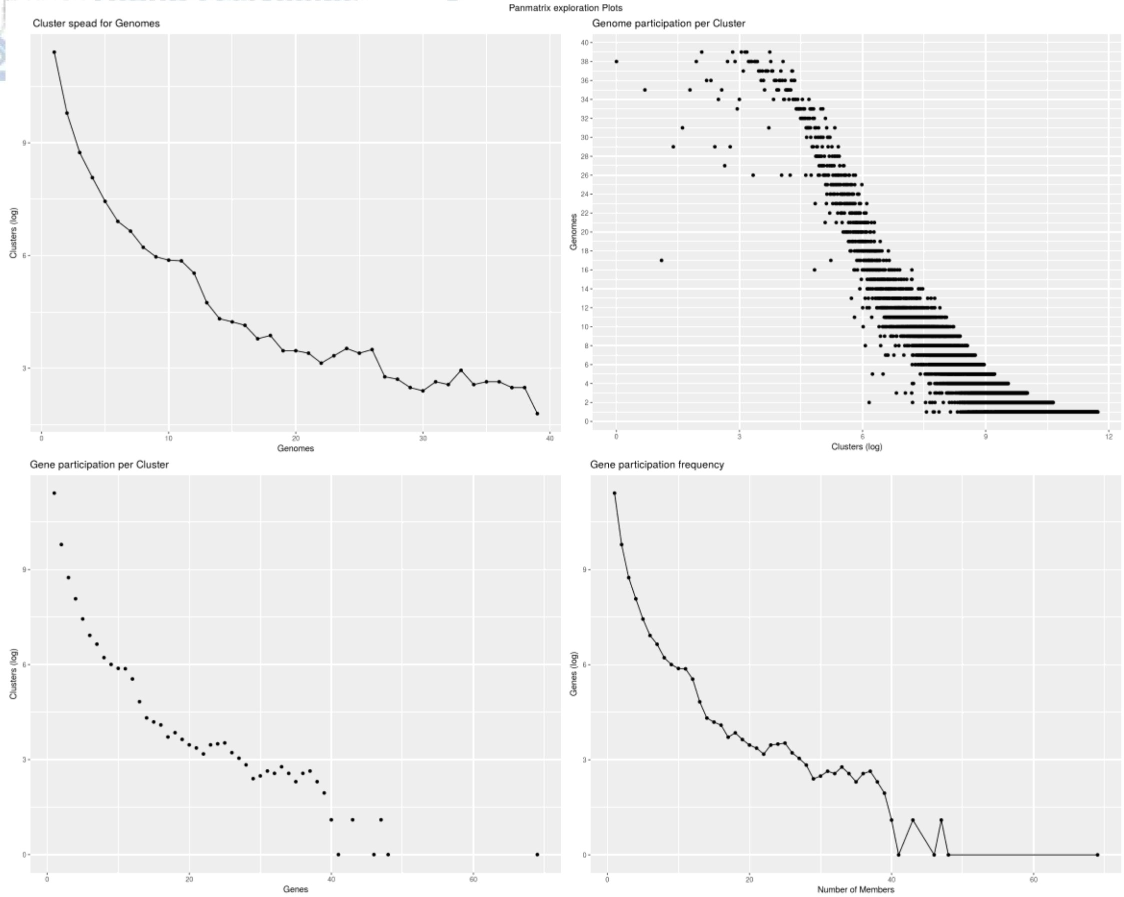


Figure 10 Summary plots and information for the Bacillus Cereus sample, produced by a standard sequence clustering pipeline with the default settings (Blast and MCL), but maintaining only the best bidirectional hits during the homology detection

3.5.2 Bacillus thuringiensis with best bi-directional hits

The dataset consists of seventeen strains (17) of Bacillus thuringiensis. The sample pangenome size of the genome collection is 60328 gene families, sample core genome was discovered to be 90 and the number of orfan genes is 44110. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.24956$, with an intercept of $k = 4837.87588$. Using the Chao estimator, a pangenome size of $n = 213312$ C.I. 95% = (63877, 6655804), with variance of $s^2 = 5958473.3486$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 9 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.004072	0.004787	0.1416
Mixing.prop	0.2987	0.6663	0.03249

Table continues below

	Comp_4	Comp_5	Comp_6
Detection.prob	0.4788	0.5058	0.8843
Mixing.prop	0.002048	0.000000001131	0.0002602

	Comp_7	Comp_8	Comp_9
Detection.prob	0.9893	0.9999	1
Mixing.prop	0.0001306	0.00003496	0.0000001394

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
9 components	0	575896	124386

Finally the sample fluidity is $\phi = 0.8577856$ with $s = 0.1419978$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.8646134$ with $\sigma = 0.13102$.

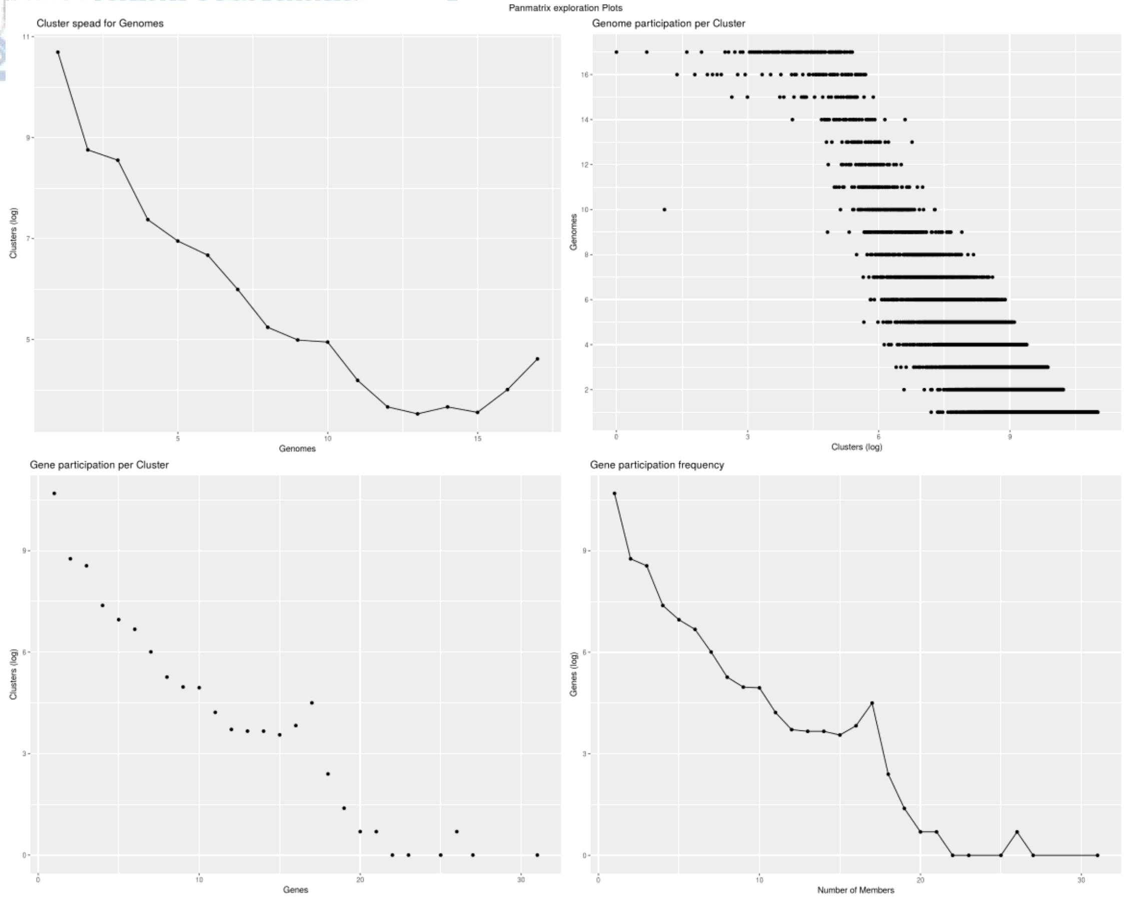


Figure 11 Summary plots and information for the Bacillus Thurigensis sample, produced by a standard sequence clustering pipeline with the default settings (Blast and MCL), but maintaining only the best bidirectional hits during the homology detection

3.5.3 Streptococcus pneumoniae with best bi-directional hits

The dataset consists of the twelve strains (12) of Streptococcus pneumoniae. The sample pangenome size of the genome collection is 11484 gene families, the sample core size is 119 gene families and the number of orfan genes is 6870. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.56344$, with an intercept of $k = 1995.30055$. Using the Chao estimator, a pangenome size of $n = 21947$ C.I. 95% = (11943, 250106), with variance of $s^2 = 122920.51518$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 5 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5
Detection.prob	0.05064	0.289	0.6013	0.8936	1
Mixing.prop	0.8919	0.06766	0.02562	0.01289	0.001902

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
5 components	42	22051	31656

Finally the sample fluidity is $\phi = 0.6854532$ with $s = 0.0996208$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.6848218$ with $\sigma = 0.0990319$.

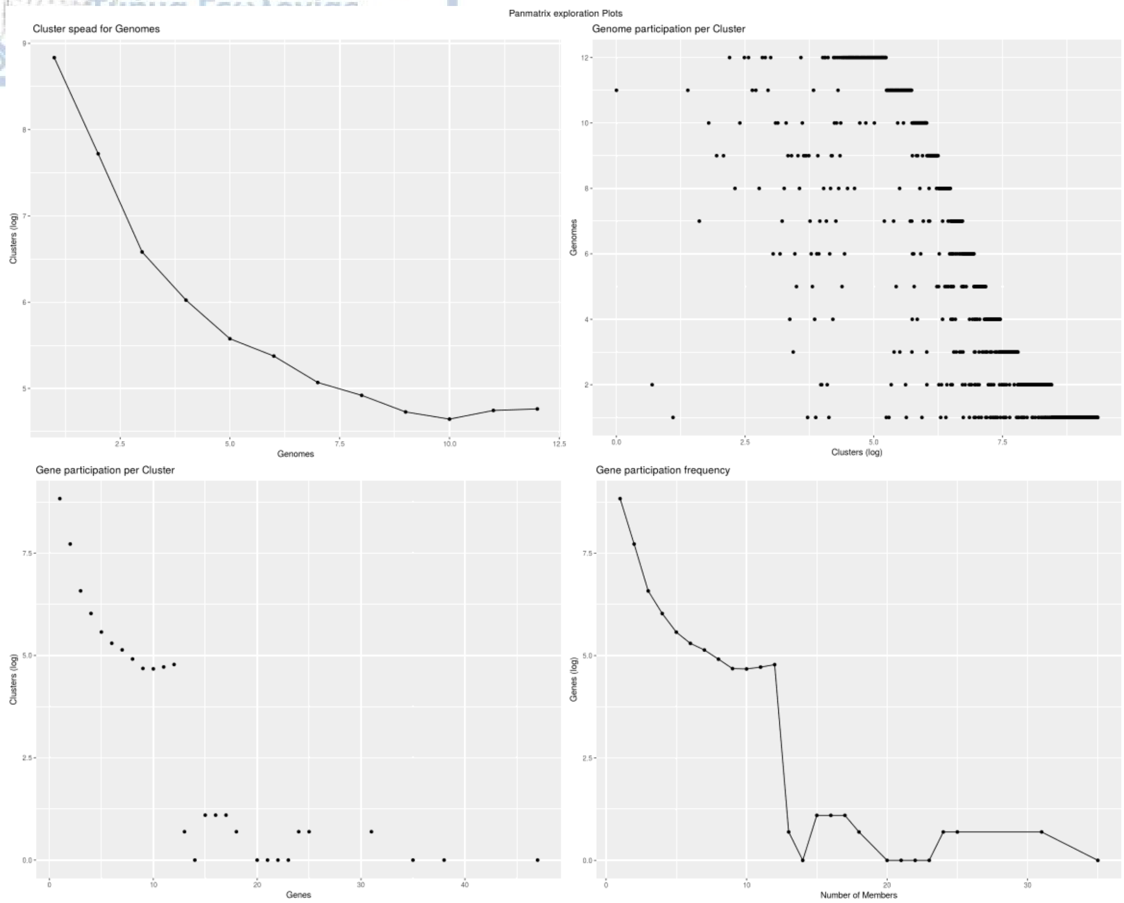


Figure 12 Summary plots and information for the Streptococcus Pneumoniae sample, produced by a standard sequence clustering pipeline with the default settings (Blast and MCL), but maintaining only the best bidirectional hits during the homology detection

3.5.4 Streptococcus Pyogenes with best bi-directional hits

The dataset consists of twelve strains (12) of Streptococcus pneumoniae. The sample pangenome size of the genome collection is 11028 gene families, 85 geme families comprise the sample core genome and the number of orfan genes is 6444. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.55862$, with an intercept of $k = 1839.18593$. Using the Chao estimator, a pangenome size of n= 19631 C.I. 95% = (11444, 189147), with variance of $s^2 = 85323.81203$ occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 4 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4
Detection.prob	0.056	0.3406	0.7729	1
Mixing.prop	0.9089	0.06371	0.02345	0.003986

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
4 components	77	19347	30488

Finally the sample fluidity is $\phi = 0.701132$ with $s = 0.1109669$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.6962727$ with $\sigma = 0.1204666$.

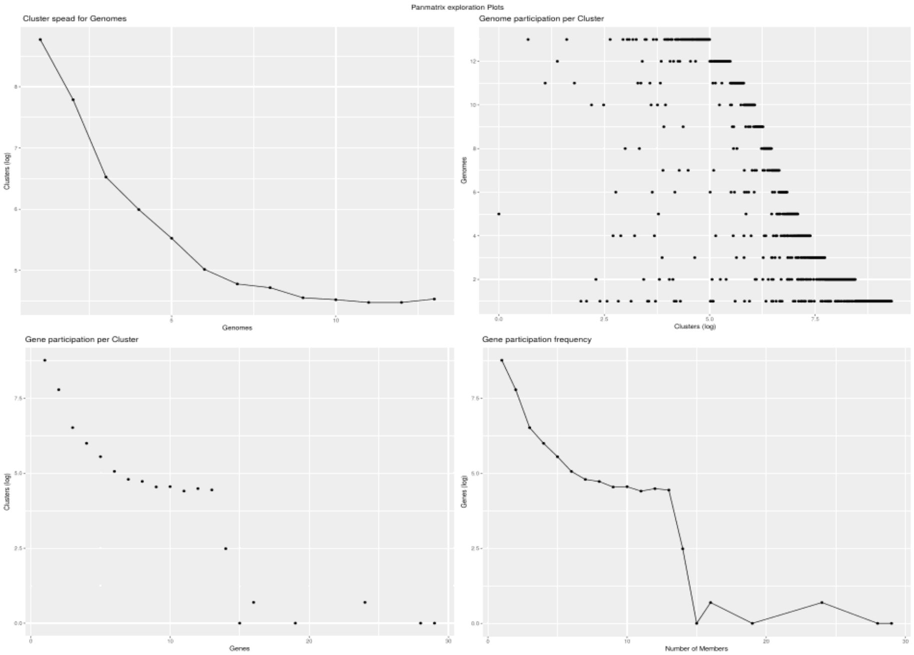


Figure 13 Summary plots and information for the Streptococcus Pyogenes sample, produced by a standard sequence clustering pipeline with the default settings (Blast and MCL), but maintaining only the best bidirectional hits during the homology detection

3.6 Comparison of DB and BBH dataset samples

As can be observed above, the datasets produced with default BLAST settings and by choosing the best bidirectional hits produce different datasets. The best bidirectional hits (BBH) scheme produces a pangenome of 19.5% bigger size, with 27.4% more orfan genes. No core genome was discovered and an open pangenome is predicted in both cases. Concerning the subsets, the following results occur:

- In the Bacillus Cereus Group, there is a very small core genome of six (6) in the BBH variation and ten (10) in the DB set with an open pangenome of approximately 120 and 102 thousand gene families respectively with more than 50% of which being orfan genes, something that is reflected by the high fluidity scores that are 0.877 for the DB and 0.908 for the BBH datasets.

- In the Bacillus Thurigiensis subset, results estimated are more similar: A pangenome of approximately fifty three (53) and sixty (60) thousand in the DB and BBH sets with a core genome of ninety one (91) and ninety (90) gene families with a fairly large fluidity score, φ= 0.8276 and φ= 0.8503.

- Streptococcus Pneumoniae strain datasets, are quite smaller in size compared to those of the Bacillus group: sample Pangenomes of ten (10) and eleven (11) thousand genome families with one hundred fourty-five (145) and one hundred nineteen (119) core gene families with a little more than 50% of the total gene families present being orfan genes in both cases.

- Streptococcus Pyogenes datasets produce pangenomes of similar size to the S. Pneumoniae sets: ten (10) and eleven (11) thousand genome families with one hundred and three (103) and eighty-five (85) core gene families with a little more than 50% of the total gene families present being orfan genes in both cases.

3.7 Three Species dataset

The sample pangenome size of the three groups genome collection is 20763 gene families, no sample core was found, and the number of orfan genes is 15371. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome a = 0.30404, with an intercept of k = 1802.32008. Using the Chao estimator, a pangenome size of $n =$ 65138 C.I. 95% = (21953,1676108), with variance of $s^2 =$ 1297926.97679 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 9 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6	Comp_7	Comp_8	Comp_9
Detection.prob	0.0184099	0.0201156	0.2161462	0.4447411	0.6058099	0.7397026	0.8717815	0.9705084	1
Mixing.prop	0.3793133	0.5717951	0.0356060	0.0131210	0.0000000	0.0000000	0.0001646	0.0000000	0

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
9 components	0	54335	42893.35

Finally the sample fluidity is $\phi = 0.8922678$ with $s = 0.1619261$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.8891019$ with $\sigma = 0.1690081$.

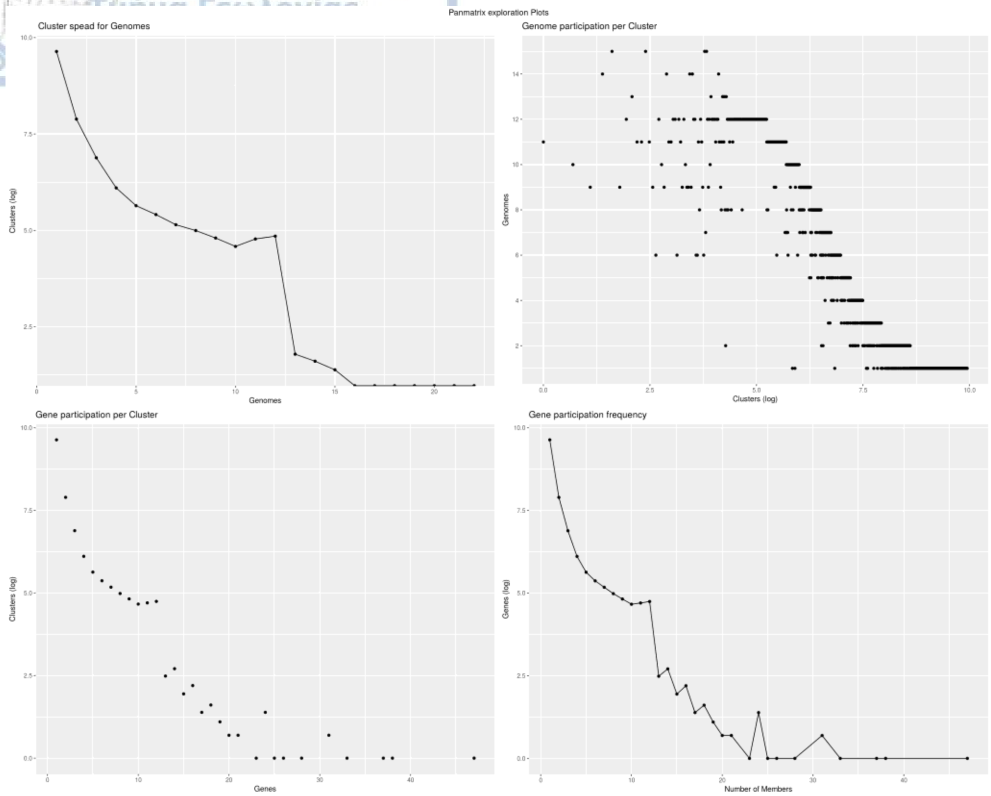


Figure 14 Summary plots and information for the 3 species sample pangenome

3.7.1 Buchnera Aphidicola

The dataset consists of six (6) Buchnera Aphidicola proteobacteria strains. The sample pangenome size of the genome collection is 2335 gene families, the sample core size is 2 gene families, and the number of orfan genes is 1855. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome a = 0.3449, with an intercept of k = 566.90363. Using the Chao estimator, a pangenome size of $n =$ 9828 C.I. 95% = (2488,368729), with variance of $s^2 =$ 377203.56526 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 3 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.0096415	0.3106778	1.0000000
Mixing.prop	0.9749189	0.0250712	0.0000098

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
3 components	0	30149	3330.829

Finally the sample fluidity is $\phi = 0.8682637$ with $s = 0.2204161$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.8537481$ with $\sigma = 0.2242996$.

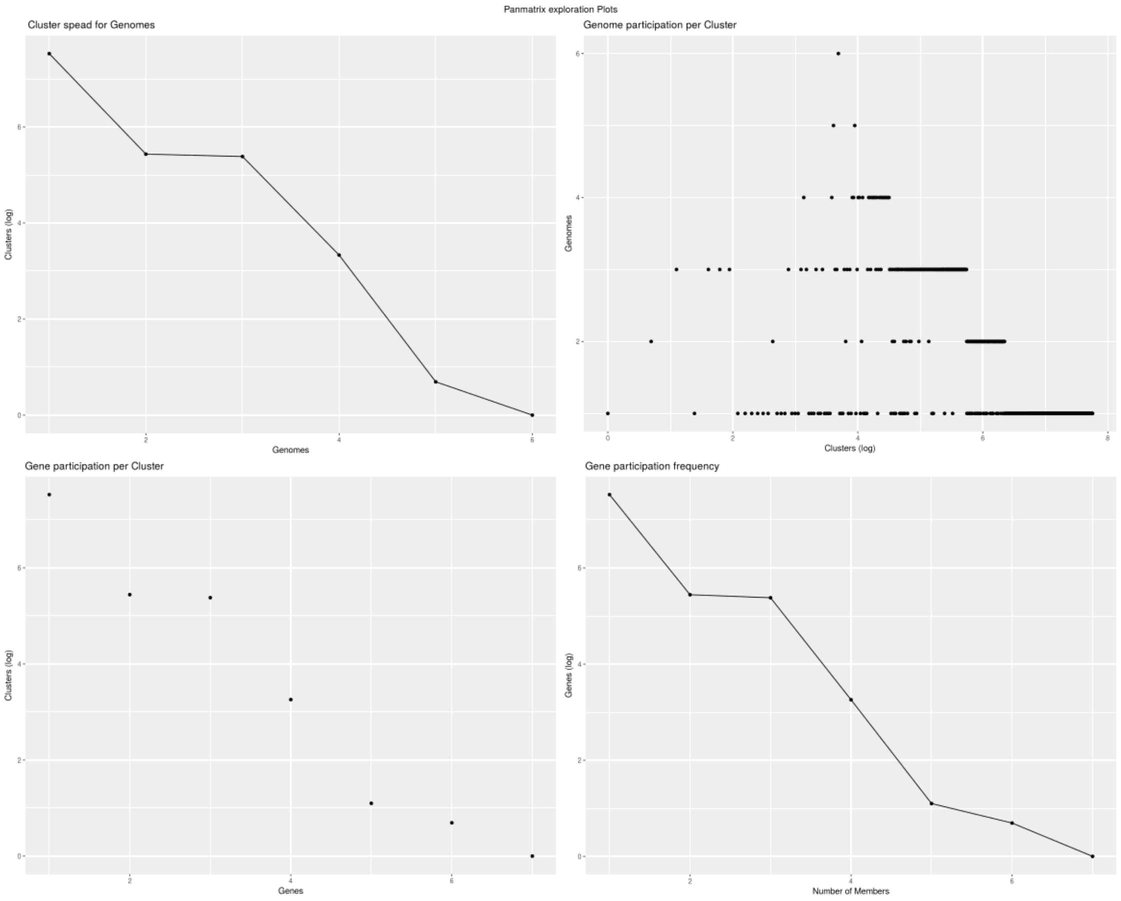


Figure 15 Summary plots and information for the Buchnera Aphidicola sample pangenome

3.7.2 Streptococcus pneumoniae

The dataset consists of twelve (12) of Streptococcus pneumoniae bacteria strains. The sample pangenome size of the genome collection is 10951 genes, the sample core size is 120 gene families, and the number of orfan genes is 6335. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome a = 0.50979, with an intercept of k = 1965.28548. Using the Chao estimator, a pangenome size of $n =$ 20177 C.I. 95% = (11370,214023), with variance of $s^2 =$ 101820.31186 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 6 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4	Comp_5	Comp_6
Detection.prob	0.0328943	0.0598360	0.2964512	0.6148279	0.9101975	1.0000000
Mixing.prop	0.2683650	0.6196189	0.0686302	0.0283327	0.0134505	0.0016027

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
6 components	34	20905	31295.56

Finally the sample fluidity is $\phi =$ 0.6712107 with $s =$ 0.0978154, while a population estimate through the use of permutations gives $\hat{\phi} =$ 0.6699617 with $\sigma =$ 0.10035 .

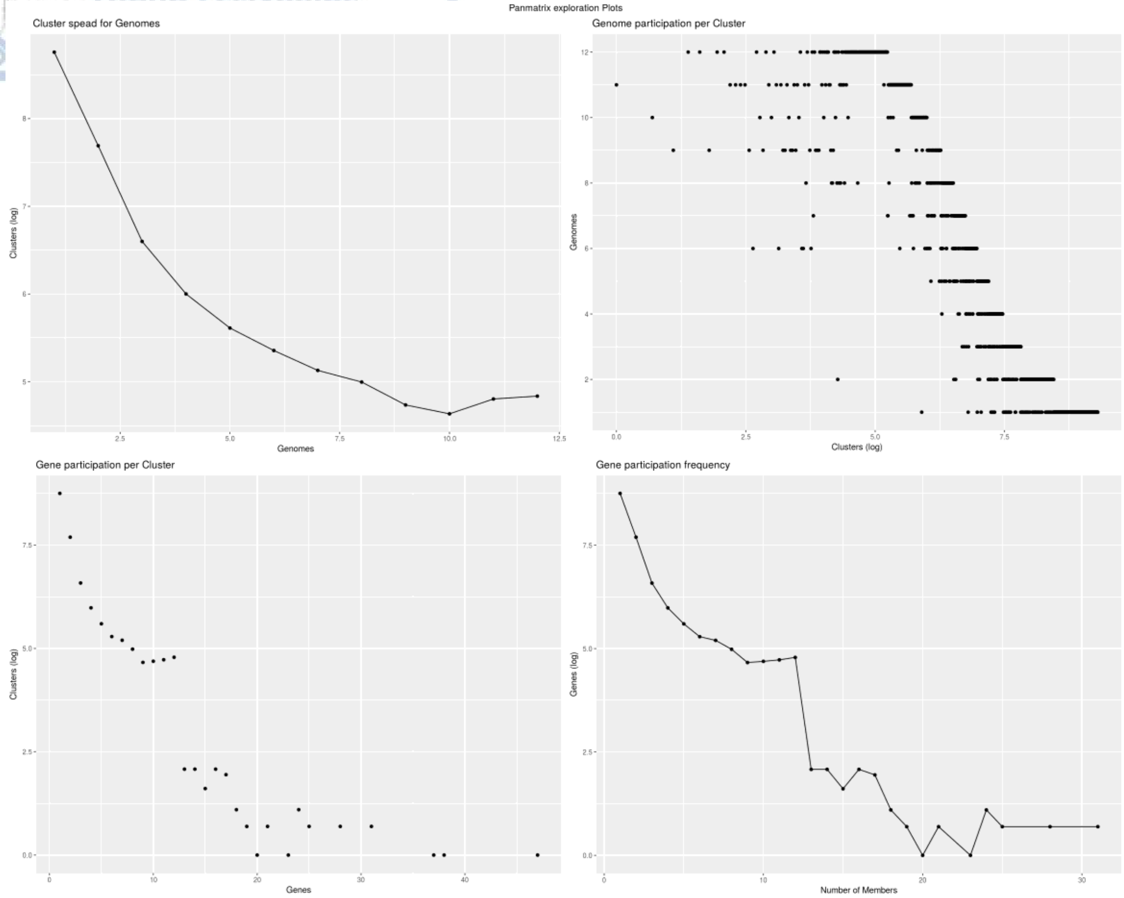


Figure 16 Summary plots and information for the Streptococcus Pneumoniae sample pangenome

3.7.3 Pyrococcus

The dataset consists of four (4) of Pyrococcus genomus. The sample pangenome size of the genome collection is 7780 gene families, the sample core size is 0 gene families, and the number of orfan genes is 7482. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome a = 0.02231, with an intercept of k = 1958.01206. Using the Chao estimator, a pangenome size of $n =$ 112598 C.I. 95% = (8601,13389474), with variance of $s^2 =$ 47650910.08926 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 4 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3	Comp_4
Detection.prob	0.0035738	0.0074906	0.1655041	1.0000000
Mixing.prop	0.8307720	0.1645480	0.0046761	0.0000039

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
4 components	2	407331	2805.628

Finally the sample fluidity is $\phi = 0.9693334$ with $s = 0.0049403$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.9693222$ with $\sigma = 0.0045424$.

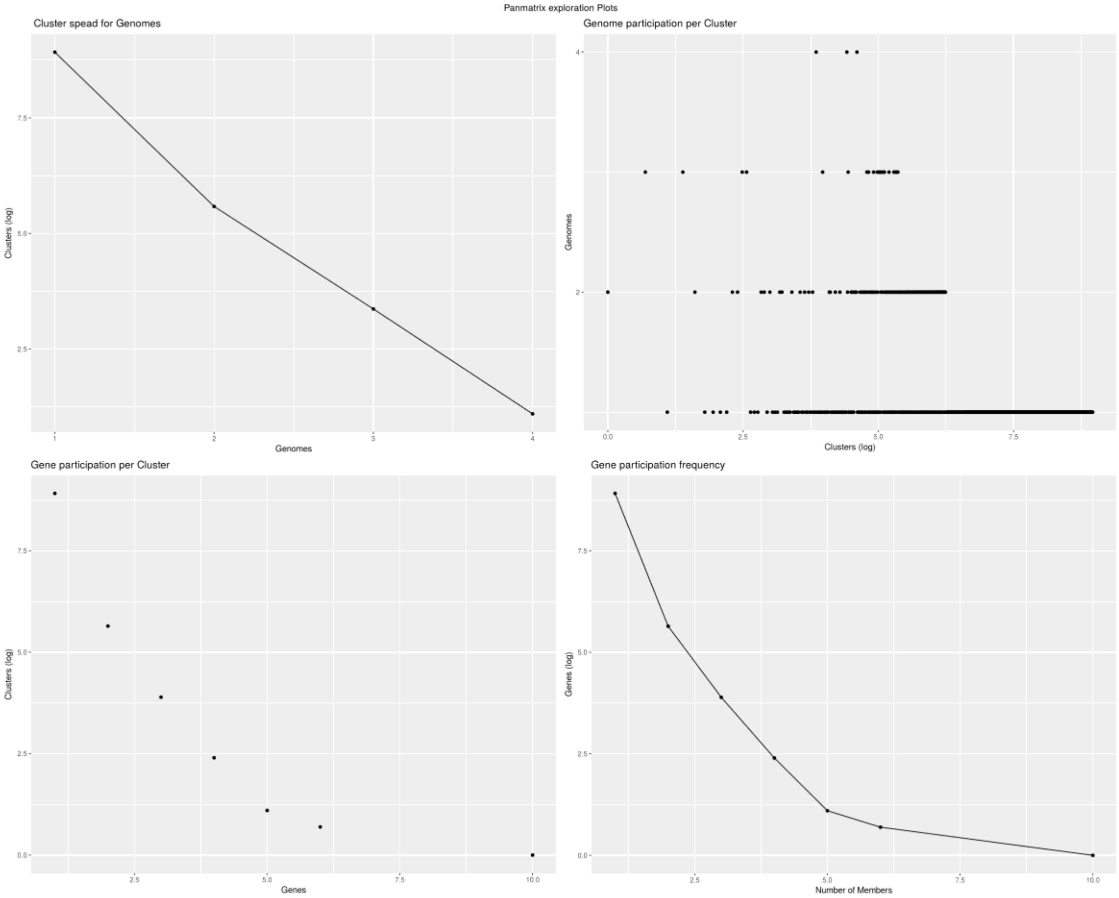


Figure 17 Summary plots and information for the Pyrococcus sample pangenome

3.8 Photosynthetic Species

The sample pangenome size of the photosynthetic species genome collection is 190759 gene families, the sample core size is 102 gene families, and the number of orfan genes is 150326. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.58993$, with an intercept of $k = 3846.06608$. Using the Chao estimator, a pangenome size of $n = 1094452$ C.I. 95% = (204331,60362862), with variance of s^2= 87971756.31804 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 9 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.001151	0.03289	0.1455
Mixing.prop	0.9764	0.01449	0.002419

	Comp_4	Comp_5	Comp_6
Detection.prob	0.4147	0.5436	0.7868
Mixing.prop	0.003842	0.002333	0.0001513

	Comp_7	Comp_8	Comp_9
Detection.prob	0.909	0.9807	1
Mixing.prop	0.0001621	0.0002213	0.00003119

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
9 components	48	1535780	441631

Finally the sample fluidity is $\phi = 0.6618692$ with $s = 0.2408981$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.6532063$ with $\sigma = 0.2413633$.

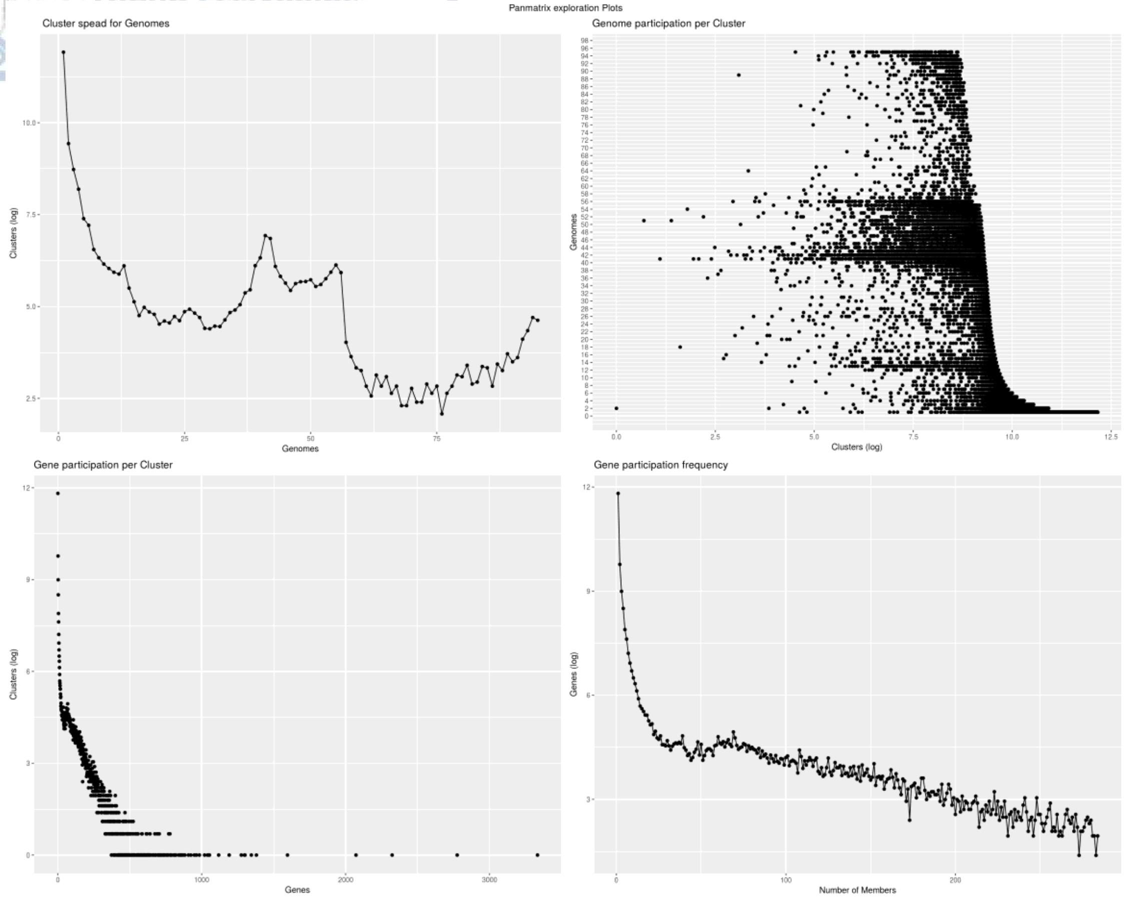


Figure 18 Summary plots and information for the photosynthetic species dataset pangenome

3.8.1 Viridiplantae species

This subset contains only the 56 Viridiplantae genomes. The sample pangenome size of the genome collection is 167290 gene families, the sample core size is 74 gene families, and the number of orfan genes is 120585. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.2017215$, with an intercept of $k = 5257.3362573$. Using the Chao estimator, a pangenome size of $n = 1031333$ C.I. 95% = (179589,60866893), with variance of s^2= 94744793.7 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 8 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.001419	0.04645	0.1928
Mixing.prop	0.9813	0.01206	0.001347

	Comp_4	Comp_5	Comp_6
Detection.prob	0.4565	0.7432	0.8966
Mixing.prop	0.0004251	0.002794	0.001063

	Comp_7	Comp_8
Detection.prob	0.9759	1
Mixing.prop	0.0009456	0.00005931

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
8 components	107	1801753	350749

Finally the sample fluidity is $\phi = 0.4618706$ with $s = 0.175542$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.4656785$ with $\sigma = 0.1794218$.

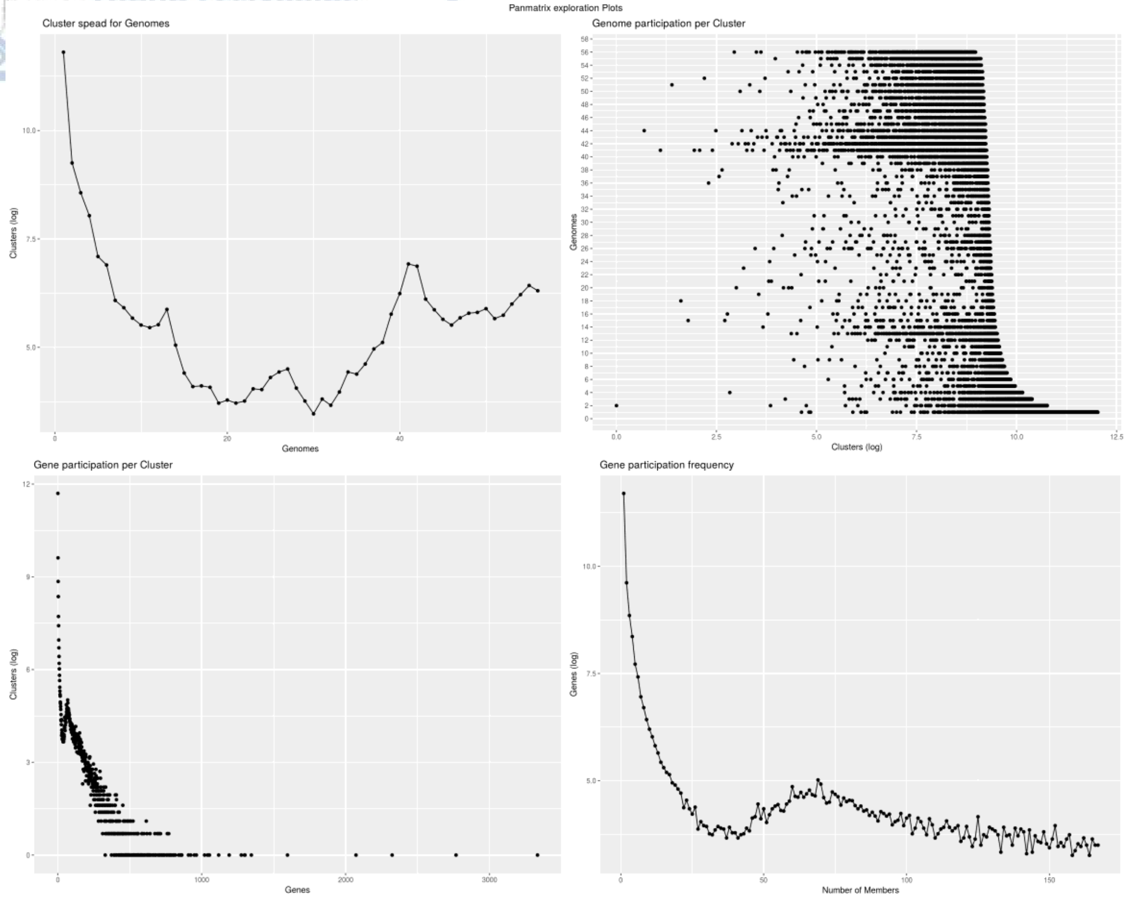


Figure 19 Summary plots and information for the Viridiplantae species dataset pangenome

3.8.2 Cyanobacteria species

This subset contains only the 39 Cyanobacteria genomes. The sample pangenome size of the genome collection is 26686 gene families, the sample core size is 367 gene families, and the number of orfan genes is 16039. Fitting the Genomes according to Heap's Law, results to the estimation of an open pangenome $a = 0.3918113$, with an intercept of $k = 1753.6772477$. Using the Chao estimator, a pangenome size of $n = 86730$ C.I. 95% = (28078,2616541), with variance of s^2= 2340842.69 occurs.

Using Binomial mixture model, it is estimated that the optimal fit for the model comprises by 9 with the following mixing probabilities:

	Comp_1	Comp_2	Comp_3
Detection.prob	0.00499	0.005542	0.07828
Mixing.prop	0.5887	0.3352	0.03509

	Comp_4	Comp_5	Comp_6
Detection.prob	0.2188	0.4277	0.6563
Mixing.prop	0.01477	0.007077	0.004267

	Comp_7	Comp_8	Comp_9
Detection.prob	0.8368	0.9499	1
Mixing.prop	0.003386	0.004971	0.006474

while the pangenome characteristics are estimated to be:

	BIC.table.Core.size	BIC.table.Pan.size	BIC.table.BIC
9 components	707	109261	88662

Finally the sample fluidity is $\phi = 0.4303886$ with $s = 0.0693669$, while a population estimate through the use of permutations gives $\hat{\phi} = 0.4282775$ with $\sigma = 0.0661636$.

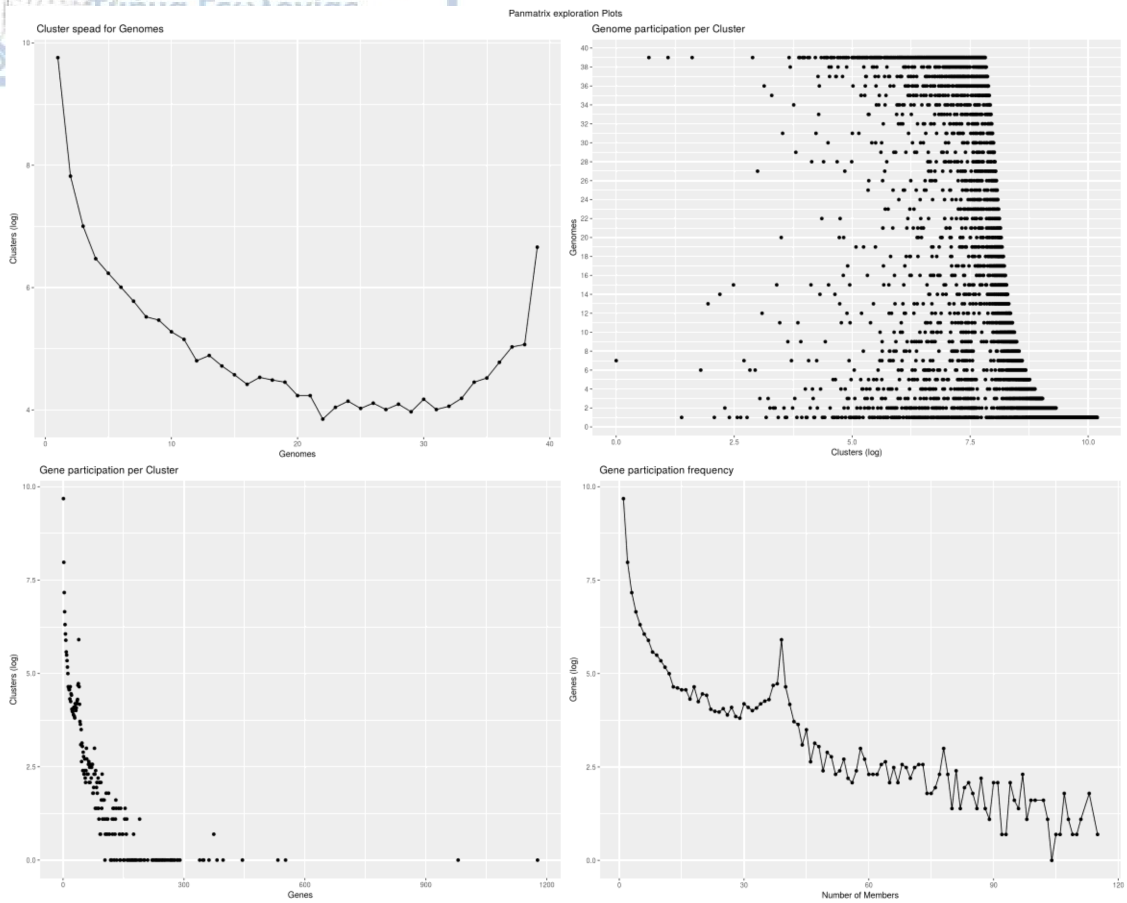


Figure 20 Summary plots and information for the Cyanobacteria species dataset pangenome

3.10 Clustering with fluidity

In the following section genome clustering with fluidity is showcased on the datasets already examined.

3.10.1 Four bacterial species

This dataset consists of genomes of Streptococcus pneumoniae,Streptococcus Pyogenes,Bacillus Cereus and Bacillus Thuringiensi. Both versions of the dataset will be examined, i.e. with all homologs accepted as hits and bidirectional best hits (BBH). Based on Gap statistic the dataset, in both versions, splits optimally in two (2) clusters that separate the Bacillus and the

Streptococcus genomes while the Dunn statistic results suggest three (3) clusters in the "default" dataset and two (2) clusters in the "BBH" dataset separation. The three (3) cluster scheme results in a clear separation of the Bacillus genomes, the Streptococcus pneumoniae and the Streptococcus Pyogenes.

Table 3 Proposed number of clusters in DB

Clusters	Index	Value
10	Average Silhuette Width	0.24944
2	Gap Statistic	0.96782
3	Dunn	1.00219
7	Entropy	0.66971

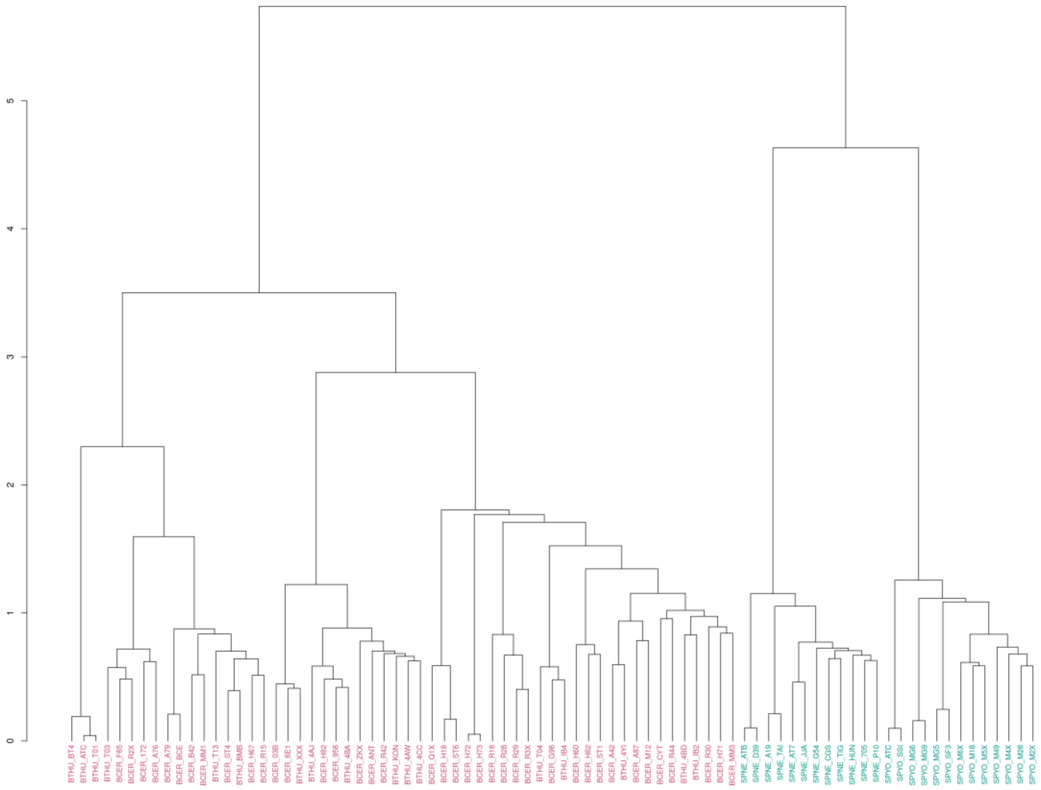


Figure 21 Two clusters separation of four species dataset produced

Table 4 Proposed number of clusters in BBH

Clusters	Index	Value
10	Average Silhuette Width	0.23506
2	Gap Statistic	0.98569
2	Dunn	1.00215
7	Entropy	0.69664

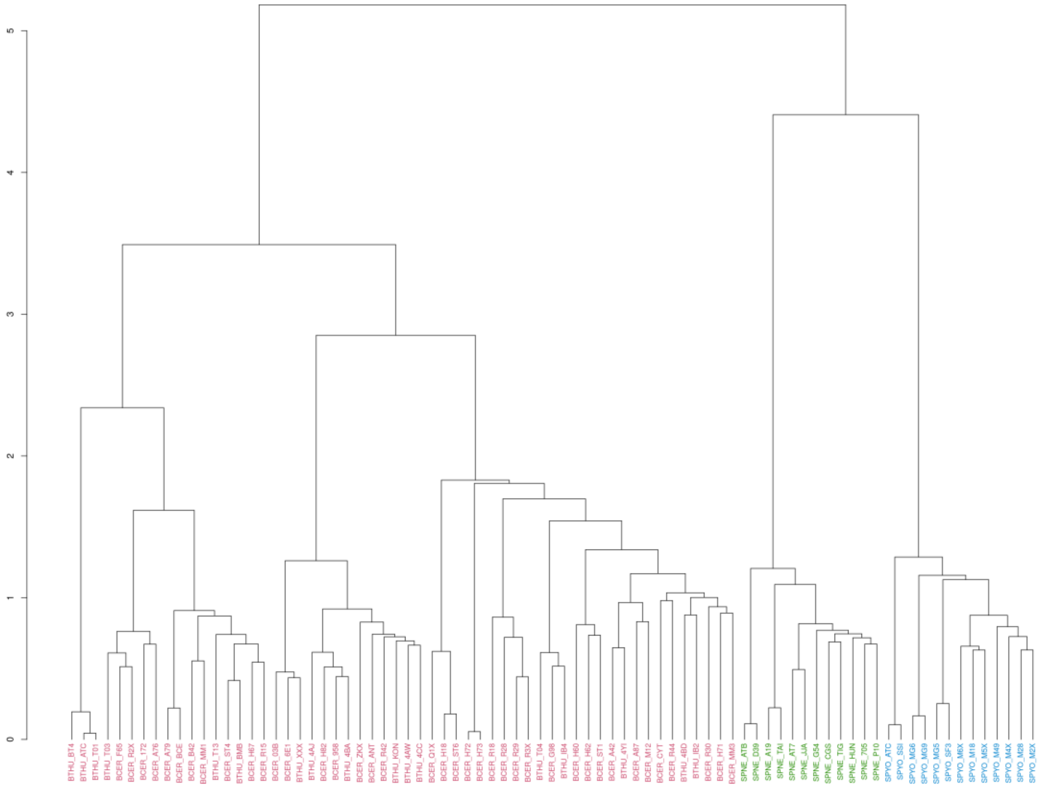


Figure 22 Three clusters separation of four species dataset produced by standard pipeline (BLAST and MCL)

3.10.2 Three Species dataset

This dataset consists of Streptococcus Pneumoniae, Buchnera Aphidicla and Pyrococcus genomes. The Gap statistic results suggest three (3) clusters: One (1) cluster contains the

Streptococcus Pneumoniae genomes, one (1) consists of four (4) Buchnera Aphidicla genomes and the last contains a mix of the four (4) Pyrococcus genomes and the remaining (3) Buchnera genomes. The Dunn Index results suggest two (2) clusters, with one (1) containing the Streptococcus Pneumoniae genomes and the other the Buchnera and Pyrococcus genomes.

Table 5 Proposed number of clusters in the Three species dataset

Clusters	Index	Value
3	Average Silhuette Width	0.2501
3	Gap Statistic	0.99353
2	Dunn	0.98473
5	Entropy	0.64091

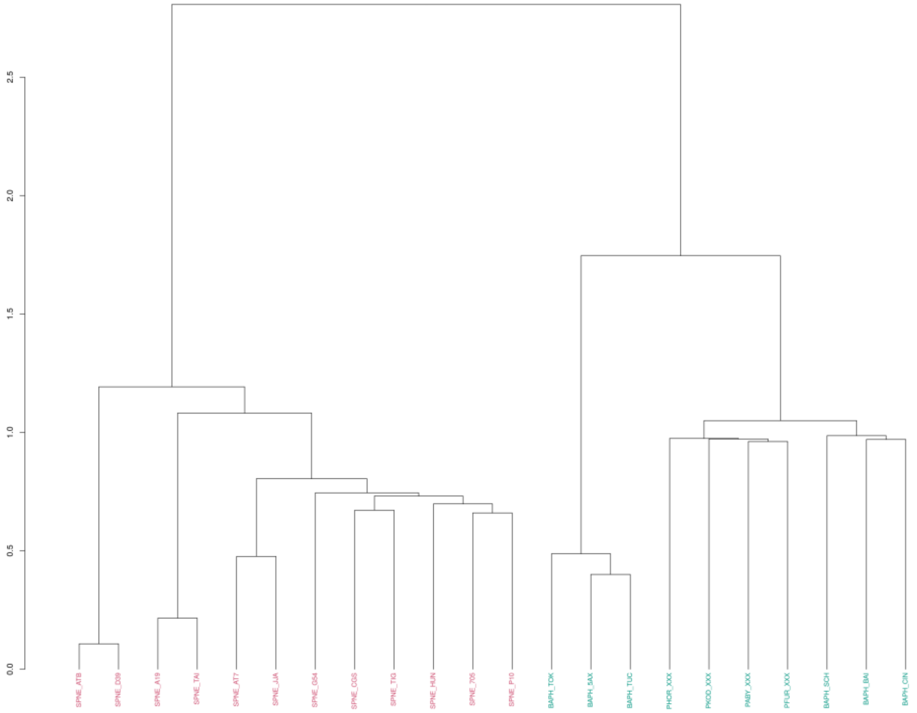


Figure 23 Three species dataset split into two clusters

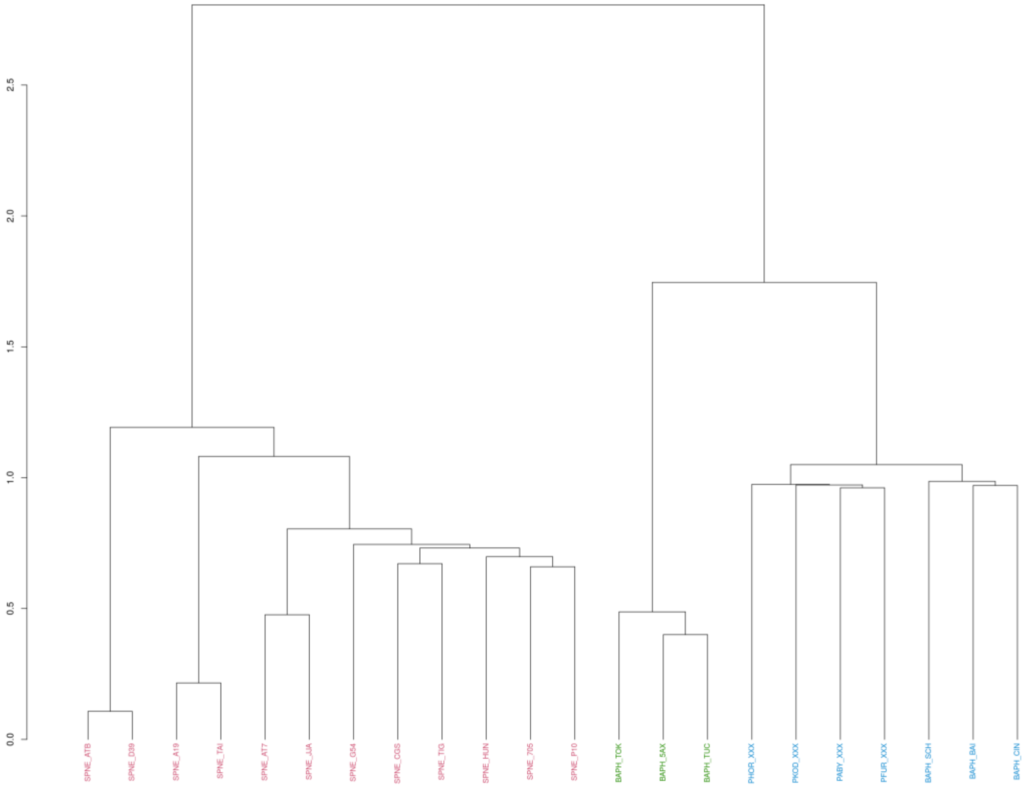


Figure 24 Three species dataset split into three clusters

3.10.3 95 genomes dataset

This dataset consists of photosynthetic species, of the Viridiplanate and the Cyanobacteria Phylum. The Dunn index results to 2 (two) clusters clearly spliting the dataset between the Viridiplanate and Cyanobacteria genomes while the Gap statistic results into 8 clusters.

Table 6 Photosynthetic species proposed number of clusters

Clusters	Index	Value
2	Average Silhuette Width	0.48942
8	Gap Statistic	0.58324
2	Dunn	0.94983
4	Entropy	0.22123

Clusters according to Dunn Statistic:

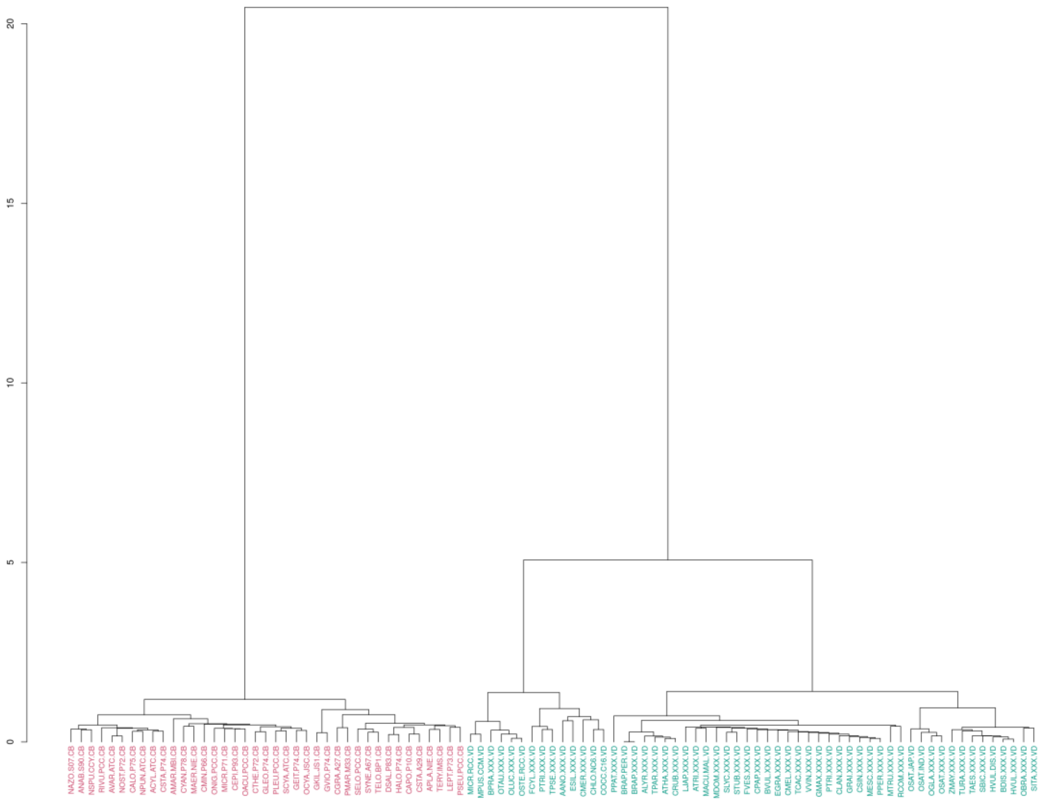


Figure 25 Two clusters separation of 95 photosynthetic species

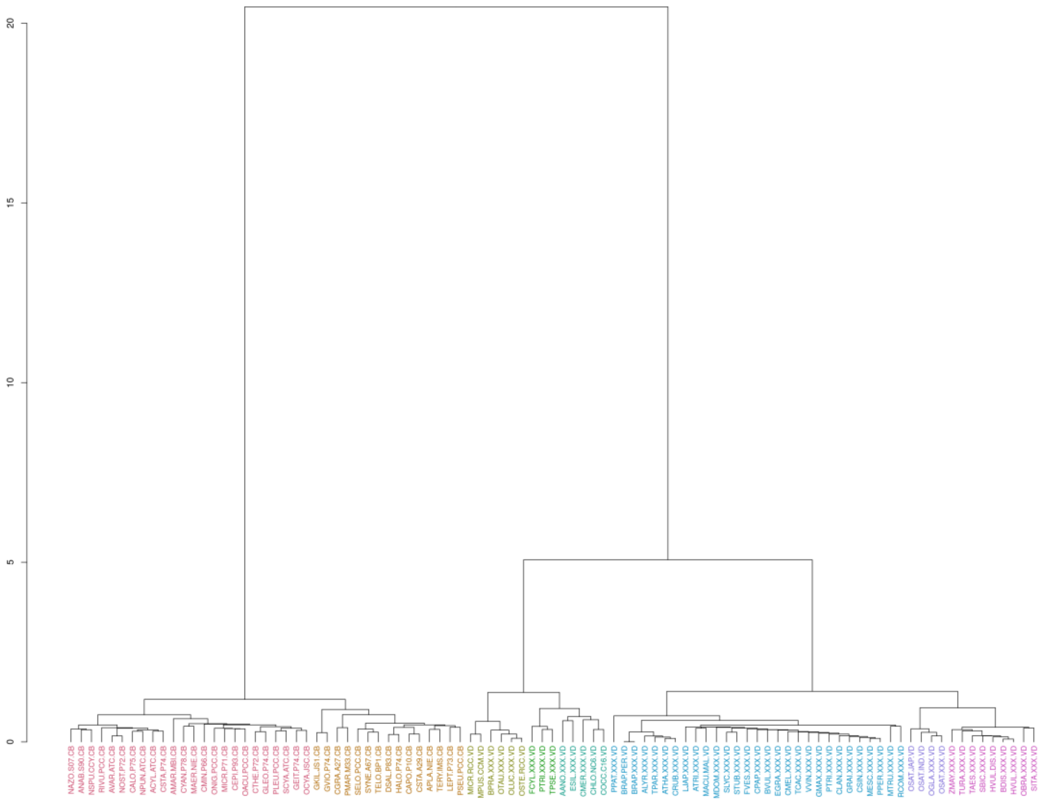


Figure 26 Eight clusters separation of 95 photosynthetic species

4 Discussion

In this thesis, the statistical properties that occur during an analysis known as pangenomic are examined. After outlining the existing knowledge surrounding this process through the available literature, the mathematical tools that allow the creation of a software package that enables this analysis are presented. The use of this package is demonstrated on many publicly available data.

First, a dataset that consists solely of Mycoplasma Pneuomoniae bacteria strains is examined: a closed pangenome is estimated, and the results of both the Chao estimator and the binomial mixture produce very close results for the pangenome size with the core genome comprising most it. This is also evident in the fluidity score which predicts 1.96% unique genes per strain. Then, a dataset consisting of 81 strains of Streptococcus Pneumoniae, Streptococcus Pyogenes, Bacillus cereus and Bacillus Thurigiensis where examined in two different versions: One with produced with the default settings used in the BLAST process (DB) and one using the best bidirectional hits between genomes. In both datasets, the entirety and each species collection of strains where examined. A big genomic difference between the bacteria examined is discernible through the fluidity scores, where 91.66% (DB set) and 93.28% (BBH set) different genes per genome are anticipated. A binomial mixture model predicts the absence of a core genome even if the organisms examined are of the same Phylum, Firmicutes. The diversity is reflected in the pangenome size estimation varying from a minimum of 319715 genes families (DB set, binomial mixture) to 495538 gene families (BBH set - Chao est.). In respect to the four (4) subsets:

- Both the Bacillus Cereus dataset variants show large diversity, an open pangenome with a core genome prediction of one (1) gene family and pangenome size estimation of approximately three hundred and twenty thousand (320k) to three hundred and fifty thousand (350k) in the BBH set and two hundred and forty-two thousand (242k) to two hundred and fifty-seven thousand (257k) gene family using binomial mixtures and the Chao estimator respectively.

- Concerning the Bacillus Thurigensis and comparing the DB and BBH sets, the DB subset produces an estimation of a smaller pangenome of approximately one hundred thousand (100k) less gene families with a prediction a much larger core of one hundred and four (104) to a predicted core of seventeen (17) genes families in the BBH subset. In both set there is a big discrepancy between results of the estimation of pangenome size computed with the mixture models and the Chao estimator, with the latter producing much smaller outcomes: n= 148494 C.I. 95% = (55963,3388574), with variance of s^2=2464843.37952 for the DB subset and n= 213312 C.I. 95% = (63877, 6655804), with variance of s^2= 5958473.3486 for the BBH subset.

- The Streptococcus Pneumoniae pangenome sample sizes are considerably smaller, therefore resulting to smaller size estimations 17854/17991 (Binomial Mixture / Chao Estimator) for the DB dataset and 21947/24859 for the BBH dataset. However, while binomial mixture models predict a mixture of five (5) components in both cases, a larger core genome of seventy-five (75) gene families is estimated for the BBH set in contrast of a core of forty-seven (47) gene families in the DB set even though the first set has a fluidity of ϕ = 0.6854 as compared to the lowest ϕ = 0.6473 of the second set.

- A similar pattern is also evident in the Streptococcus Pyogenesis, a smaller pangenomes sizes than Bacillus: 16288/16366 for the DB set and 19339/19631 gene families (Binomial Mixture / Chao Estimator) for the BBH set. However, a core genome of 66 gene families as compared to one of 77 gene families is observed in the DB as opossed in the BBH set.

The third big grouping consists of Pyrococcus species genomes, Streptococcus Pneumoniae strains and Buchnera Aphidicola strains. These genomes are quite diverse between them and considering that the dataset was synthesized to test the fluidity clustering scheme, a prediction of an open pangenome with no core families and pangenome size quite larger, 54331/65138 for Binomial mixtures and Chao est., that the pangenome size of the individual species datasets, is not surprising.

The final dataset comprises of 95 Viridiplantae and Cyanobacteria genomes and was chosen to examine the aspects of a more complex pangenome, i.e. that of species capable of photosynthesis. Both subsets exhibit core genomes and the sample core genome is 102 gene families while the core genome size predicted to be 48 gene families. However, this outcome should not be interpreted as definitive and a larger dataset containing more strains of the same organisms will provide more complete results.

As pertaining to clustering results using fluidity, it is observed that it can be used in successfully distinguishing between genomes of different genera of the same phylum but not between species, as observed in the case of the dataset containing four bacterial species. It can be also used to distinguish between genomes of different kingdoms as is evident in the case of Viridiplantae and Cyanobacteria in the dataset containing the photosynthetic species.

Some general remarks can be made concerning the techniques used and possible directions of research. Our first point concerns the Heap's law model which is not as effective, in terms of information derived, when applied to datasets consisting of diverse genomes as these datasets are expected to always have open pangenomes. Moreover, even though Heap's law models in the pangenomic context where originally applied on microbial data, they are usually presented as a golden standard (Golicz, Batley, and Edwards 2016, Carlos Guimaraes et al. (2015)) without any further mathematical scrutiny.

Secondly, the basis of a reliant pangenomic analysis is the stage of the genome alignment and clustering. In the cases examined, it is shown that different techniques can produce pangenomes of different size and cohesiveness leading to false conclusions.

The relevance of both points can be examined using the paradigm of the Buchnera genomes: our results from an examination of six genomes of the species suggest an open pangenome with no core genome. However, the literature findings impart a closed pangenome with about 20% to 26% of the gene families comprising the core genome (Mira et al. (2010), Manzano-Marín et al. (2012)).

Feature development aims include: a) application of automated testing to code in order to further quality control, b) optimization of code and documentation up to CRAN, a formally regulated R package repository, publication standards, c) creation of an interactive web application, with the R functionality Shiny based on the workflow presented and finally d) Integration of process in an existing pipeline to offer a complete analysis (Kintsakis, Psomopoulos, and Mitkas 2016; Psomopoulos, Vrousgou, and Mitkas 2015).

Appendix

Genome summary - Dataset 1 (M. Pneumoniae)

Genomes	1	2	3	4	5	6	7
Clusters	33	10	13	16	22	16	1100

Genome summary - Dataset 2 (4 species)

Genomes	1	2	3	4	5	6
Clusters	93296	23071	12247	5615	3386	2398

Genomes	7	8	9	10	11	12	13	14
Clusters	1681	1380	1029	855	728	693	511	360

Genomes	15	16	17	18	19	20	21	22	23
Clusters	284	249	239	210	189	142	123	106	81

Genomes	24	25	26	27	28	29	30	31	32	33	34	35
Clusters	61	71	48	46	43	34	40	36	36	34	18	21

Genomes	36	37	38	39	40	41	42	43	44	45	46	47
Clusters	21	22	22	22	25	21	18	15	25	16	11	7

Genomes	48	49	50	51	52	53	54	55	56	57	58	59
Clusters	16	17	10	13	21	16	11	9	7	1	1	0

Genomes	60	61	62	63	64	65	66	67	68	69	70	71
Clusters	1	1	3	0	2	1	0	2	3	0	0	0

Genomes	72	73	74	75	76	77	78	79	80	81
Clusters	0	0	0	0	0	0	0	0	0	0

Dataset 2 - subset 2 (B. Cereus)

Genomes	1	2	3	4	5	6
Clusters	67714	16410	6847	3763	2235	1385

Genomes	7	8	9	10	11	12	13	14
Clusters	1021	682	506	479	403	316	168	120

Genomes	15	16	17	18	19	20	21	22	23	24	25
Clusters	110	91	70	78	59	42	45	40	44	38	32

Genomes	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Clusters	39	27	25	19	15	12	18	22	16	13	20	16	16	8

Dataset 2 - subset 2 (B. Thurigiensis)

Genomes	1	2	3	4	5	6	7
Clusters	35740	6704	5220	1862	1204	921	491

Genomes	8	9	10	11	12	13	14	15	16	17
Clusters	251	184	176	99	61	54	49	44	68	115

Dataset 2 - subset 3 (S. Pneumoniae)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12
Clusters	5659	2127	766	432	279	225	168	163	120	113	133	146

Dataset 2 - subset 4 (S. Pyogenes)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12	13
Clusters	5333	2243	749	420	278	165	125	127	103	91	106	94	118

Genome summary - Dataset 3 (4 species BBH)

Genomes	1	2	3	4	5	6

Clusters	128528	26613	12787	4976	2750	1903

Genomes	7	8	9	10	11	12	13	14
Clusters	1340	1083	838	732	644	574	415	254

Genomes	15	16	17	18	19	20	21	22	23	24
Clusters	220	186	166	161	132	103	92	75	48	38

Genomes	25	26	27	28	29	30	31	32	33	34	35	36
Clusters	46	27	23	27	26	20	25	22	20	9	19	15

Genomes	37	38	39	40	41	42	43	44	45	46	47	48
Clusters	13	18	14	19	14	16	12	16	13	7	4	12

Genomes	49	50	51	52	53	54	55	56	57	58	59	60
Clusters	17	8	9	20	10	12	10	4	1	0	0	0

Genomes	61	62	63	64	65	66	67	68	69	70	71	72
Clusters	0	1	0	0	0	0	0	1	0	0	0	0
Genomes	73	74	75	76	77	78	79	80	81			
Clusters	0	0	0	0	0	0	0	0	0			

Dataset 3 - subset 1 (B. Cereus)

Genomes	1	2	3	4	5	6	7
Clusters	90692	17896	6262	3209	1707	1002	772

Genomes	8	9	10	11	12	13	14	15	16	17
Clusters	502	390	357	350	252	115	75	69	63	44

Genomes	18	19	20	21	22	23	24	25	26	27	28	29
Clusters	48	32	32	30	23	28	34	30	33	16	15	12
Genomes	30	31	32	33	34	35	36	37	38	39		
Clusters	11	14	13	19	13	14	14	12	12	6		

Dataset 3 - subset 2 (B. Thurigiensis)

Genomes	1	2	3	4	5	6	7

Clusters	44110	6358	5183	1597	1045	789	400			
Genomes	8	9	10	11	12	13	14	15	16	17
Clusters	189	147	141	66	39	34	39	35	55	101

Dataset 3 - subset 3 (S. Pneumoniae)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12
Clusters	6870	2254	722	413	264	216	159	137	113	104	115	117

Dataset 3 - subset 4 (S. Pyogenes)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12	13
Clusters	6444	2412	682	401	251	151	119	112	95	92	88	88	93

Dataset 4 (3 Species)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Clusters	15371	2661	976	447	282	224	172	148	122	98	119	128	6	5	4	0	0	0	0	0	0	0

Dataset 4 - subset 1 (B. Aphidicola)

Genomes	1	2	3	4	5	6
Clusters	1857	229	218	28	2	1

Dataset 4 - subset 2 (S. Pneumoniae)

Genomes	1	2	3	4	5	6	7	8	9	10	11	12
Clusters	6356	2188	735	404	274	212	169	148	114	103	122	126

Dataset 4 - subset 3 (Pyrococcus genomes)

Genomes	1	2	3	4
Clusters	7482	266	29	3

Dataset 5 (Photosynthetic species)

Genomes	1	2	3	4	5	6
Clusters	150326	12502	6173	3613	1621	1353

Genomes	7	8	9	10	11	12	13	14	15
Clusters	700	559	471	418	377	359	449	245	169

Genomes	16	17	18	19	20	21	22	23	24
Clusters	116	145	128	120	92	100	95	113	101

Genomes	25	26	27	28	29	30	31	32	33	34
Clusters	129	138	124	110	82	81	87	86	103	126

Genomes	35	36	37	38	39	40	41	42
Clusters	135	156	215	234	451	558	1022	945

Genomes	43	44	45	46	47	48	49	50	51
Clusters	442	337	281	230	277	291	293	307	256

Genomes	52	53	54	55	56	57	58	59	60	61
Clusters	270	316	378	459	374	56	38	28	26	17

Genomes	62	63	64	65	66	67	68	69	70	71	72	73
Clusters	13	23	17	22	14	17	10	10	16	11	11	18

Genomes	74	75	76	77	78	79	80	81	82	83	84	85
Clusters	14	17	8	14	17	23	22	30	18	19	29	28
Genomes	86	87	88	89	90	91	92	93	94	95		
Clusters	17	31	26	41	33	37	61	77	110	102		

Viridiplantae species

Genomes	1	2	3	4	5	6	7
Clusters	134228	10425	5245	3090	1205	991	438

Genomes	8	9	10	11	12	13	14	15	16
Clusters	370	291	248	234	250	356	156	82	60

Genomes	17	18	19	20	21	22	23	24	25	26	27	28
Clusters	61	59	41	44	41	43	57	56	74	84	90	58

Genomes	29	30	31	32	33	34	35	36	37	38
Clusters	43	32	45	39	53	84	80	101	143	166

Genomes	39	40	41	42	43	44	45	46
Clusters	319	513	1015	965	451	353	283	248

Genomes	47	48	49	50	51	52	53	54	55	56
Clusters	293	326	332	362	288	311	403	500	618	547

Cyanobacteria species

Genomes	1	2	3	4	5	6	7	8
Clusters	17310	2494	1101	647	510	406	323	250

Genomes	9	10	11	12	13	14	15	16	17	18
Clusters	237	196	173	122	133	112	97	83	93	89

Genomes	19	20	21	22	23	24	25	26	27	28	29	30
Clusters	86	69	69	47	57	63	56	61	55	60	53	65

Genomes	31	32	33	34	35	36	37	38	39
Clusters	55	58	66	86	92	119	153	159	781

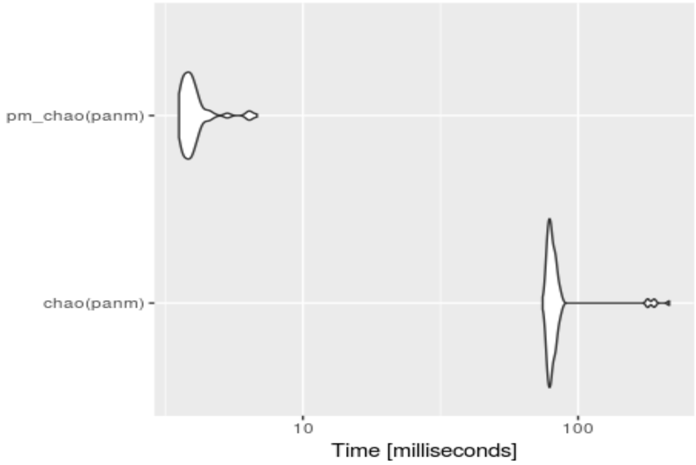
Benchmarks

Following benchmarks where run at the mpneumoniae dataset from package micropan. All commands where evaluated 100 times. Function names with a “pm” suffix belong to package pasaR. Package pasaR shows a clear advantage over micropan in terms of speed, with the only exception being the binomial mixture models. This happens due to different parameter choices in the optimization of the log-likelihood function: Micropan calls for a maximum of 300 iterations with a relative tolerance of 10^-6 while in pasaR the number of maximum iterations is 200 times the number of components examined with a relative tolerance of 10^-8.

Chao Estimator runtime comparison

```
## Unit: milliseconds
```

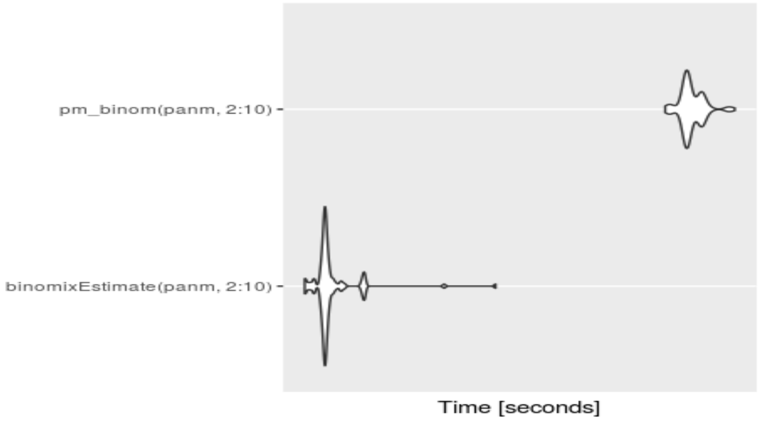
expr	min	lq	mean	median	uq	max	neval
chao(panm)	76.31	78.86	85.62	80.53	82.1	190.6	100
pm_chao(panm)	3.644	3.881	4.146	3.991	4.132	6.603	100



Binomial models runtime comparison

```
Unit: seconds
```

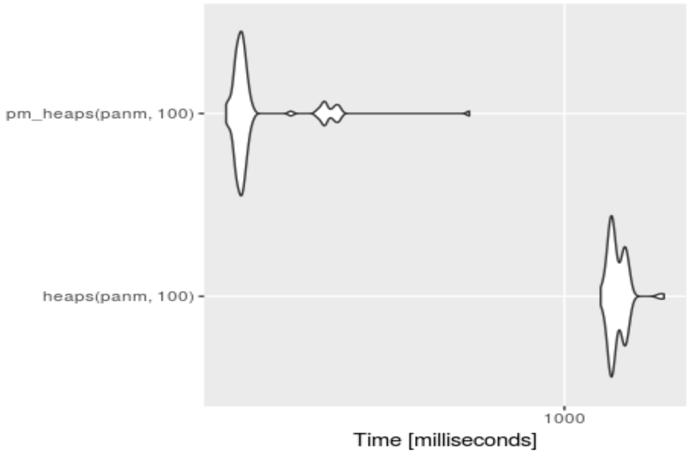
expr	min	lq	mean	median	uq	max	neval
binomixEstimate(panm, 2:10)	1.053	1.077	1.12	1.121	1.127	1.259	100
pm_binom(panm, 2:10)	2.656	2.752	2.799	2.811	2.83	2.983	100



Heaps model runtime comparison

Unit: milliseconds

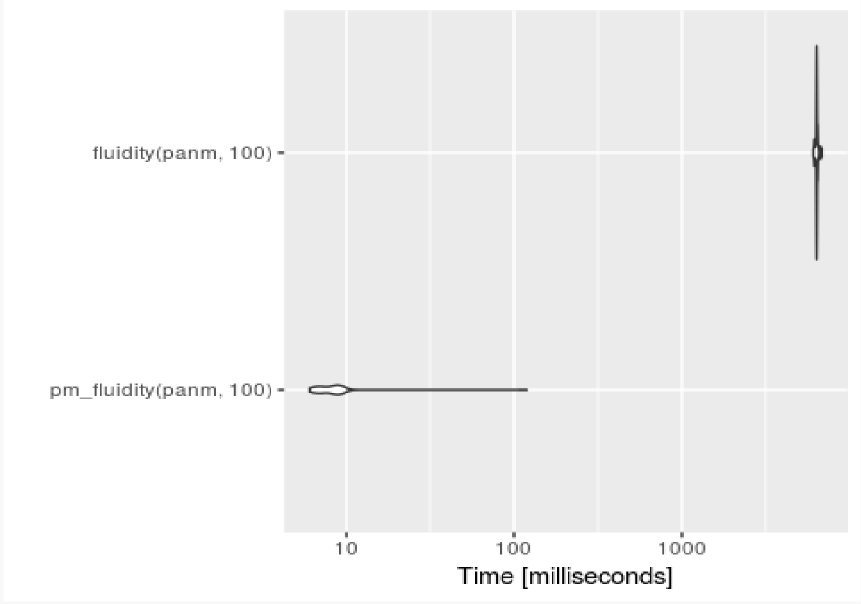
expr	min	lq	mean	median	uq	max	neval
heaps(panm, 100)	1240	1294	1347	1341	1406	1500	100
pm_heaps(panm, 100)	158	166	177.3	169.5	173.8	293.3	100



Fluidity runtime comparison

Unit: milliseconds

expr	min	lq	mean	median	uq	max	neval
pm_fluidity(panm, 100)	6.012	6.85	7.703	7.839	8.464	10.45	100
fluidity(panm, 100)	6068	6273	6309	6332	6357	6528	100


fluidity(panm, 100)
pm_fluidity(panm, 100)
10
100
1000
Time [milliseconds]

References

Aken, Bronwen L, Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, et al. 2016. “Database update The Ensembl gene annotation system,” 1–19. doi:10.1093/database/baw093.

Carlos Guimaraes, Luis, Leandro Benevides de Jesus, Marcus Vinicius Canario Viana, Artur Silva, Rommel Thiago Juca Ramos, Siomar de Castro Soares, and Vasco Azevedo. 2015. “Inside the Pan-genome - Methods and Software Overview.” Current Genomics 16 (4): 245–52. doi:10.2174/1389202916666150423002311.

Chao, Anne. 1987. “Estimating the Population Size for Capture-Recapture Data with Unequal Catchability.” Biometrics 43 (4): 783. doi:10.2307/2531532.

Dalquen, Daniel A., and Christophe Dessimoz. 2013. “Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals.” Genome Biology and Evolution 5 (10): 1800–1806. doi:10.1093/gbe/evt132.

Deza, Michel Marie, and Elena Deza. 2009. Encyclopedia of distances. Springer. doi:10.1007/978-3-642-00234-2.

Dunn, J. C. 1973. “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.” Journal of Cybernetics 3 (3): 32–57. doi:10.1080/01969727308546046.

Free Software Foundation 2007. “GNU GENERAL PUBLIC LICENSE Version 3”. url:https://www.gnu.org/licenses/gpl-3.0.en.html

Golicz, Agnieszka A., Jacqueline Batley, and David Edwards. 2016. “Towards plant pangenomics.” Plant Biotechnology Journal 14 (4): 1099–1105. doi:10.1111/pbi.12499.

Heaps, H S. 1978. Information retrieval: computational and theoretical aspects. Orlando, FL, USA: Academic Press, Inc. http://search.proquest.com/docview/57244815?accountid=142596.

Hirsch, C. N., J. M. Foerster, J. M. Johnson, R. S. Sekhon, G. Muttoni, B. Vaillancourt, F. Penagaricano, et al. 2014. “Insights into the Maize Pan-Genome and Pan-Transcriptome.” The Plant Cell 26 (1): 121–35. doi:10.1105/tpc.113.119982.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2007. An Introduction to Statistical Learning. Vol. 64. 9-12. doi:10.1016/j.peva.2007.06.006.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, et al. 2017. “Four Simple Recommendations to Encourage Best Practices in Research Software.” F1000Research 6: 876. doi:10.12688/f1000research.11407.1.

Jiménez, Rafael C., Mateusz Kuzak, Monther Alhamdoosh, Michelle Barker, Bérénice Batut, Mikael Borg, Salvador Capella-Gutierrez, et al. 2017. “Four Simple Recommendations to Encourage Best Practices in Research Software.” F1000Research 6: 876. doi:10.12688/f1000research.11407.1.

Kintsakis, Athanassios M., Fotis E. Psomopoulos, and Pericles A. Mitkas. 2016. “Data-Aware Optimization of Bioinformatics Workflows in Hybrid Clouds.” Journal of Big Data 3 (1). Springer International Publishing: 20. doi:10.1186/s40537-016-0055-2.

Kintsakis, Athanassios M., Fotis E. Psomopoulos, and Pericles A. Mitkas. 2016. “Data-Aware Optimization of Bioinformatics Workflows in Hybrid Clouds.” Journal of Big Data 3 (1). Springer International Publishing: 20. doi:10.1186/s40537-016-0055-2.

Kislyuk, Andrey O, Bart Haegeman, Nicholas H Bergman, and Joshua S Weitz. 2011. “Genomic fluidity: an integrative view of gene diversity within microbial populations.” BMC Genomics 12 (1): 32. doi:10.1186/1471-2164-12-32.

Kovács, Ferenc, Csaba Legány, and Attila Babos. 2005. “Cluster Validity Measurement Techniques.” Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence 2006: 1–11. doi:10.7547/87507315-91-9-465.

Lapierre, Pascal, and J. Peter Gogarten. 2009. “Estimating the size of the bacterial pan-genome.” Trends in Genetics 25 (3): 107–10. doi:10.1016/j.tig.2008.12.004.

Li, Ruiqiang, Yingrui Li, Hancheng Zheng, Ruibang Luo, Hongmei Zhu, Qibin Li, Wubin Qian, et al. 2010. “Building the sequence map of the human pan-genome.” Nature Biotechnology 28 (1). Nature Publishing Group: 57–63. doi:10.1038/nbt.1596.

Li, Ying-hui, Guangyu Zhou, Jianxin Ma, Wenkai Jiang, Long-guo Jin, Zhouhao Zhang, Yong Guo, et al. 2014. “De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits.” Nature Biotechnology 32 (10). Nature Publishing Group: 1045–52. doi:10.1038/nbt.2979.

Manzano-Marín, Alejandro, Araceli Lamelas, Andrés Moya, and Amparo Latorre. 2012. “Comparative Genomics of Serratia spp.: Two Paths towards Endosymbiotic Life.” PLoS ONE 7 (10). doi:10.1371/journal.pone.0047274.

McInerney, James O., Alan McNally, and Mary J. O’Connell. 2017. “Why prokaryotes have pangenomes.” Nature Microbiology 2 (4). Macmillan Publishers Limited: 17040. doi:10.1038/nmicrobiol.2017.40.

McKiernan, Erin C., Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, et al. 2016. “How Open Science Helps Researchers Succeed.” eLife 5 (JULY): 1–19. doi:10.7554/eLife.16800.

Meila, Marina. 2007. “Comparing clusterings-an information based distance.” Journal of Multivariate Analysis 98 (5): 873–95. doi:10.1016/j.jmva.2006.11.013.

Mira, Alex, Ana B. Martín-Cuadrado, Giuseppe D&apos;Auria, and Francisco Rodríguez-Valera. 2010. “The bacterial pan-genome: A new paradigm in microbiology.” International Microbiology 13 (2): 45–57. doi:10.2436/20.1501.01.110.

Miyamoto, Sadaaki, Ryosuke Abe, Yasunori Endo, and Jun Ichi Takeshita. 2016. “Ward method of hierarchical clustering for non-Euclidean similarity measures.” In Proceedings of the 2015 7th International Conference of Soft Computing and Pattern Recognition, Socpar 2015, 60–63. IEEE. doi:10.1109/SOCPAR.2015.7492784.

Murtagh, Fionn, and Pierre Legendre. 2014. “Ward???s Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward???s Criterion?” Journal of Classification 31 (3): 274–95. doi:10.1007/s00357-014-9161-z.

Muzzi, Alessandro, Vega Masignani, and Rino Rappuoli. 2007. “The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials.” Drug Discovery Today 12 (11-12): 429–39. doi:10.1016/j.drudis.2007.04.008.

Newman, M. E. J. 2004. “Power laws, Pareto distributions and Zipf’s law.” Contemporary Physics 46 (5): 323–51. doi:10.1016/j.cities.2012.03.001.

Pearson, William R. 2013. “An Introduction to Sequence Similarity (‘ Homology ’) Searching.” Current Protocols in Bioinformatics 43 (3): 1–8. doi:10.1002/0471250953.bi0301s42.

Perez-Riverol, Yasset, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe da Veiga Leprevost, Christian Fufezan, et al. 2016. “Ten Simple Rules for Taking Advantage of Git and GitHub.” PLoS Computational Biology 12 (7): 1–11. doi:10.1371/journal.pcbi.1004947.

Proost, Sebastian, Michiel Van Bel, Dries Vaneechoutte, Yves Van De Peer, Bernd Mueller-roeber, and Klaas Vandepoele. 2015. “PLAZA 3 . 0 : an access point for plant comparative genomics Dirk Inz e” 43 (October 2014): 974–81. doi:10.1093/nar/gku986.

Psomopoulos, Fotis E, Olga T Vrousgou, and Pericles A Mitkas. 2015. “Large-Scale Modular Comparative Genomics : The Grid Approach [v1; Not Peer Reviewed].” F1000Research 2015 4(ISCB Com (377): 1. doi:10.7490/f1000research.1110127.1.

Psomopoulos, Fotis E., Athanassios M. Kintsakis, and Pericles A. Mitkas. 2016. “A Pan-Genome Approach and Application to Species with Photosynthetic Capabilities.” In 5th European Conference on Computational Biology (ECCB 2016), 5:2132. The Hague: F1000Research 2016. doi:10.7490/f1000research.1112964.1.

R Core Development Team. 2016. “R: a language and environment for statistical computing.” Vienna, Austria: R Foundation for Statistical Computing. doi:10.1017/CBO9781107415324.004.

Read, Betsy A., Jessica Kegel, Mary J. Klute, Alan Kuo, Stephane C. Lefebvre, Florian Maumus, Christoph Mayer, et al. 2013. “Pan genome of the phytoplankton Emiliania underpins its global distribution.” Nature 499 (7457): 209–13. doi:10.1038/nature12221.

Rhee, Seung Yon, and Marek Mutwil. 2014. “Towards revealing the functions of all genes in plants.” Trends in Plant Science 19 (4). Elsevier Ltd: 212–21. doi:10.1016/j.tplants.2013.10.006.

Rouli, L, V Merhej, P Fournier, and D Raoult. 2015. “The bacterial pangenome as a new tool for analysing pathogenic bacteria.” New Microbes and New Infections 7. Elsevier Ltd: 72–85. doi:10.1016/j.nmni.2015.06.005.

Rousseeuw, Peter J. 1987. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.” Journal of Computational and Applied Mathematics 20 (C): 53–65. doi:10.1016/0377-0427(87)90125-7.

Snipen, Lars, and Kristian Hovde Liland. 2017. “micropan: Microbial Pan-Genome Analysis.” https://cran.r-project.org/package=micropan.

Snipen, Lars, Trygve Almøy, and David W Ussery. 2009. “Microbial comparative pan-genomics using binomial mixture models.” BMC Genomics 10 (1): 385. doi:10.1186/1471-2164-10-385.

Sun, Chen, Zhiqiang Hu, Tianqing Zheng, Kuangchen Lu, Yue Zhao, Wensheng Wang, Jianxin Shi, et al. 2017. “RPAN: Rice pan-genome browser for ~3000 rice genomes.” Nucleic Acids Research 45 (2): 597–605. doi:10.1093/nar/gkw958.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, et al. 2005. “Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial ‘pan-genome’.” Proceedings of the National Academy of Sciences 102 (39): 13950–5. doi:10.1073/pnas.0506758102.

Tettelin, Herve, David Riley, Ciro Cattuto, and Duccio Medini. 2008. “Comparative genomics: the bacterial pan-genome.” Current Opinion in Microbiology 11 (5): 472–77. doi:10.1016/j.mib.2008.09.006.

The Computational Pan-genomics Consortium, 2016. “Computational Pan-Genomics: Status, Promises and Challenges.” May. doi:10.1101/043430.

The Uniprot Consortium. 2017. “UniProt : the universal protein knowledgebase.” Nucleic Acids Research 45 (November 2016): 158–69. doi:10.1093/nar/gkw1099.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. “Estimating the number of clusters in a data set via the gap statistic.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2): 411–23. doi:10.1111/1467-9868.00293.

Vandepoele, Klaas, Michiel Van Bel, Guilhem Richard, Sofie Van Landeghem, Bram Verhelst, Hervé Moreau, Yves Van De Peer, Nigel Grimsley, and Gwenael Piganeau. 2013. “Genomics update pico-PLAZA, a genome database of microbial photosynthetic eukaryotes” 15: 2147–53. doi:10.1111/1462-2920.12174.

Vernikos, George, Duccio Medini, David R. Riley, and Herve Tettelin. 2015. “Ten years of pan-genome analyses.” Current Opinion in Microbiology 23 (February): 148–54. doi:10.1016/j.mib.2014.11.016.

Wickham, Hadley. 2015. Advanced R. Boca Raton: CRC press. doi:10.1201/b17487.

Software References

Auguie, Baptiste. 2016. “gridExtra: Miscellaneous Functions for ‘Grid’ Graphics.” https://cran.r-project.org/package=gridExtra.

Gagolewski, Marek. 2017. “R Package Stringi: Character String Processing Facilities.” http://www.gagolewski.com/software/stringi/.

Hennig, Christian. 2015. “Fpc: Flexible Procedures for Clustering.” https://cran.r-project.org/package=fpc.

Snipen, Lars, and Kristian Hovde Liland. 2017. “Micropan: Microbial Pan-Genome Analysis.” https://cran.r-project.org/package=micropan.

Wickham Hadley. 2017a. “Stringr: Simple, Consistent Wrappers for Common String Operations.” https://cran.r-project.org/package=stringr.

Wickham Hadley. 2017b. “Tidyverse: Easily Install and Load ‘Tidyverse’ Packages.” https://cran.r-project.org/package=tidyverse.

Wickham, Hadley, Jim Hester, and Romain Francois. 2017. “Readr: Read Rectangular Text Data.” https://cran.r-project.org/package=readr.