EL PANGENOMA: UNA MANERA DE ENTENDER LA DIVERSIDAD GENÓMICA

PAN-GENOME: HOW TO UNDERSTAND GENOMIC DIVERSITY

Ram González Buenfil[1]

[1] Benemérita Universidad Autónoma de Puebla

Facultad de Biología

Licenciatura en Biotecnología

ram.glez@gmail.com

Resumen

Sólo hasta la llegada de la era genómica con herramientas como la secuenciación masiva y el análisis bioinformático, en los primeros genomas secuenciados de procariontes, se encontró la gran variedad genética que existe entre miembros de una misma especie. Al analizar esta variedad se evidencia que las familias de genes pueden agruparse de acuerdo con su nivel de conservación entre muestras. Se llama genoma *común* al conjunto de genes presentes en la mayoría de muestras y genoma *accesorio* al conjunto que varía entre las muestras, y pangenoma es el conjunto de estos dos grupos. Esta variación interespecífica puede explicarse por distintos fenómenos que involucran la interacción de diferentes organismos en su nicho ecológico, como la transferencia horizontal de genes y la deriva y las migraciones genéticas. En este artículo de divulgación se ofrecen algunas perspectivas

de los pangenomas, una introducción a su estudio y una fuente bioinformática para su análisis.

Palabras clave: Pangenoma, Genes *comunes*, Genes *accesorios,* Variación genética, Transferencia horizontal de genes, Deriva genética.

ABSTRACT

Not until the arrival of the genomic era with tools such as massive sequencing and bioinformatic analysis, in the first sequenced prokaryotic genomes, was a great genetic variety within the members of a same species found. By analyzing this variety, it is evident that gene families can be grouped according to their level of conservation between samples. The set of genes present in most samples is called *core* genome, and the set of genes that varies among samples is called *accessory* genome; pan-genome refers to the set that includes these two groups. This interspecific variation can be explained by various phenomena that involve the interaction of different organisms in their ecological niche, such as horizontal gene transfer and genetic drift and migrations. This informative paper presents some perspectives about pan-genomes, an introduction to their study and a bioinformatic pipeline for analysis.

Keywords: Pan-genome, *Core* genes, *Accessory* genes, Genetic variation, Horizontal gene transfer, Genetic drift.

Introducción

La existencia de una enorme variabilidad genética intraespecífica entre cepas de la misma especie se evidenció tan pronto como se analizaron los primeros genomas secuenciados de procariontes (Bowman *et al.*, 1991). Para describir esta variación, se han acuñado términos como genomas *común* y *accesorio* (Young *et al.*, 2006). El genoma *común* se refiere a las familias de genes "esenciales" que se encuentran en todos los miembros secuenciados hasta el momento, mientras el genoma *accesorio* se refiere a aquellos genes que son "dispensables" o "prescindibles" y no se encuentran en todos los genomas (Tettelin *et al.*, 2005). El pangenoma se refiere a todas las familias de genes encontradas en la especie como un todo (Ku *et al.*, 2015) (ver Figura 1; ver Tabla 1 para definiciones completas de estos y otros términos) (McInerney, McNally y O'Connell, 2017). Algunas especies de procariontes tienen pangenomas extensos (o "abiertos"), es decir, manifiestan una gran diferencia en el contenido de genes, mientras que otras

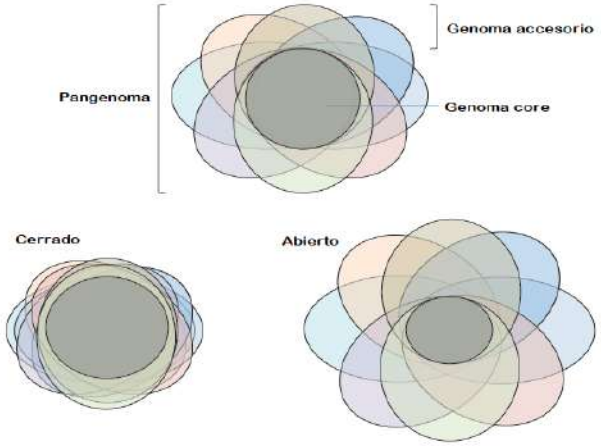

Figura 1. Representación esquemática de pangenomas como diagramas de Venn.
Las especies difieren en el tamaño de sus pangenomas. Los pangenomas más grandes y abiertos están correlacionados con con un tamaño de población más efectivo y la habilidad de migrar (McInerney *et al.*, 2017).

especies tienen pangenomas pequeños (o "cerrados"), donde la diversidad genética en la especie es muy pequeña.

El entendimiento que se tenga del pangenoma de una especie dependerá de si la extensa diversidad de la especie se haya muestreado y de la cantidad de genomas de esta diversidad que se hayan secuenciado.

Para los procariontes, la fuente principal de variabilidad en el genoma es la transferencia horizontal de genes (THG), junto con la pérdida y las duplicaciones de genes, aunque estos últimos desempeñan un papel menos importante (Treangen y Rocha, 2011).

Evolución de genomas distinta a una estructura de árbol

Hace casi tres décadas, se observó una incongruencia entre secuencias de rARN 16s casi idénticas al género *Aeromonas* y bajos niveles de hibridación ADN-ADN (Martinez-Murcia, Benlloch y Collins, 1992). Aunque poco común, esta disparidad se atribuyó a la idea de un pangenoma ya que en aquel entonces no se conocían las secuencias genómicas. Sin embargo, pronto se hizo evidente que los genomas procarióticos son sustancialmente influenciados por la THG (Creevey *et al.*, 2004; Doolittle, 1999), cuestionando la hipótesis del árbol de la vida, aunque algunos piensan que la THG no afecta las filogenias (Daubin, 2003). Actualmente, las miles de secuencias genómicas disponibles revelan la penetrante influencia de introgresiones (movimiento de genes de una especie a otra) de distintos tipos (Bapteste *et al.*, 2012). Hasta la fecha, el análisis del pangenoma más grande llevado a cabo

para una sola especie incluyó 2,085 genomas de *Escherichia coli* (Land *et al.*, 2015) y se calcularon 3,188 familias de genes *comunes,* las cuales se definieron como presentes en 95% de los genomas analizados, y aproximadamente 90,000 familias únicas de genes. Por contraste, el patógeno intracelular *Chlamydia trachomatis* tiene un tamaño de pangenoma sólo un poco más grande que su genoma *común* con 974 genes de pangenoma, 821 genes pertenecientes al genoma *común* y 67 genomas secuenciados (Ding, Baumdicker y Neher, 2017). Esto presenta un rango de tamaño de genoma *común* de 3 a 84% para genomas muestreados correctamente. Al secuenciar más genomas, el genoma *común* tiende a hacerse pequeño y el *accesorio* a hacerse más grande (Lukjancenko, Wassenaar y Ussery, 2010). Resulta interesante que al explorar los patrones de presencia y ausencia de genes en una muestra de 573 genomas y al extrapolar un número más grande de genomas, se ha calculado que el pangenoma bacteriano entero tiene un tamaño infinito (Lapierre y Gogarten, 2009). Esto se ha relacionado con una "lluvia constante de material genético sobre los genomas" (Lapierre y Gogarten, 2009) e implica que los genomas tienen una fuente casi ilimitada de genes que se pueden muestrear.

Los pangenomas también pueden encontrarse en eucariontes (Ding *et al.*, 2017).Se piensa que el genoma humano tiene entre 15 y 40 Mb de ADN *accesorio* (Li *et al.*, 2010), mientras que 14 genomas del cocolitóforo *Emiliania huxley* tienen sólo 69.5% de genes en común entre todos sus genomas. Sin embargo, en los eucariontes

la herencia genética es algo diferente con niveles de THG más bajos que en los procariontes (Ku *et al.*, 2015) y niveles más altos de duplicación (Lynch, 2003). Este artículo se enfoca en los procariontes, principalmente porque no se han hecho muchos estudios de variación intraespecífica a nivel genómico a lo largo de varias especies de eucariontes como se han hecho con procariontes.

El proceso que lleva a la generación de pangenomas aún requiere una explicación teórica exhaustiva que incorpore la existencia de una distribución de tamaños de pangenomas, desde tamaños mínimos a muy extensos (McInerney *et al.*, 2017). La THG es una forma de mutación y puede manejarse como tal en modelos evolutivos de pangenomas. Estos modelos también deben considerar la variación en tamaños de población efectivos (definida como el número de individuos que contribuyen con descendencia a la siguiente generación), las tazas de mutación, los coeficientes de selección, la influencia de la deriva aleatoria, los tipos de especiación y la existencia de una variación en la tendencia de una especie procariótica particular para formar extensos pangenomas (Ding, Baumdicker y Neher, 2017; McInerney, McNally y O'Connell, 2017). No es suficiente ingresar nuevos alelos o genes a una célula (introgresión) para asegurar la retención de los mismos (Shapiro, 2016). Se puede asumir que el ingreso de genes es bastante frecuente por la plenitud de elementos móviles genéticos y el ADN exógeno, pero las interrogantes son qué promueve la retención y por qué no existe un genoma "típico" para cada especie procariótica.

Estrategias para determinar pangenomas en procariontes: Fuente de *Roary*

Actualmente, la implementación de herramientas bioinformáticas especializadas en la anotación de genes y la comparación entre ellos permiten diseñar *fuentes* (serie de metodologías a seguir) para encontrar el pangenoma de una especie o interespecie. Una fuente muy utilizada para estos fines es la fuente de *Roary*, disponible en https://sanger-pathogens.github.io/Roary/ para Linux, el cual toma genomas ensamblados y anotados en formato GFF3 (producido por *PROKKA* (Seemann, 2014)) para calcular el pangenoma. Este formato puede analizar conjuntos de datos con miles de muestras sin comprometer la calidad de los resultados, lo cual es computacionalmente imposible con los métodos existentes, y puede analizar 128 muestras en menos de una hora usando 1 GB de RAM y un solo procesador (Page *et al.*, 2015). Llevar a cabo este análisis usando otros métodos existentes tomaría semanas y cientos de GB de memoria RAM. Sin embargo, *Roary* no está hecho para análisis

Figura 2. Página principal de Roary, una fuente para el análisis de pangenomas.

Recuperado el 20 de noviembre de 2017 de https://sanger-pathogens.github.io/Roary/

metagenómicos o para comparar conjuntos de genomas extremadamente diversos.

Roary trabaja con archivos en formato GFF3, los cuales pueden obtenerse por la anotación de genes en *PROKKA* o desde el sitio FTP del NCBI, ftp://ftp.ncbi.nlm.nih.gov/. Una vez instalado el programa, se puede ingresar el comando roary -h en la terminal y se desplegarán las instrucciones de uso de *Roary* (Figura 3). Entre los archivos que el programa arroja como resultado está un archivo csv que contiene una lista de cada uno de los genes del pangeoma aislados (cada genoma de cepa o especie ingresada), así como un archivo FASTA que contiene los genes accesorio y otro que contiene los genes del pangenoma en un formato FASTA y un archivo con las estadísticas del pangenoma: número de

```
Usage:   roary [options] *.gff

Options: -p INT     number of threads [1]
         -o STR     clusters output filename [clustered_proteins]
         -f STR     output directory [.]
         -e         create a multiFASTA alignment of core genes using PRANK
         -n         fast core gene alignment with MAFFT, use with -e
         -i         minimum percentage identity for blastp [95]
         -cd FLOAT  percentage of isolates a gene must be in to be core [99]
         -qc        generate QC report with Kraken
         -k STR     path to Kraken database for QC, use with -qc
         -a         check dependancies and exit
         -b STR     blastp executable [blastp]
         -c STR     mcl executable [mcl]
         -d STR     mcxdeblast executable [mcxdeblast]
         -g INT     maximum number of clusters [50000]
         -m STR     makeblastdb executable [makeblastdb]
         -r         create R plots, requires R and ggplot2
         -s         don't split paralogs
         -t INT     translation table [11]
         -z         don't delete intermediate files
         -v         verbose output to STDOUT
         -w         print version and exit
         -y         add gene inference information to spreadsheet, doesn't work with -e
         -iv STR    Change the MCL inflation value [1.5]
         -h         this help message
```

Figura 3 .Menú de opciones de Roary.

Recuperado el 20 de noviembre de 2017 de https://sanger-pathogens.github.io/Roary/

genes *comunes* y *accesorios* y el total de genes.

Para visualizar los resultados de manera gráfica, Marco Galardini contribuyó al proyecto con la creación de una secuencia de comandos, roary_plots.py, la cual provee tres figuras: la primera muestra el árbol filogenético comparado con una matriz que contiene la presencia y ausencia de los genes *comunes* y *accesorios* (Figura 4), la segunda es una gráfica de pastel del resumen de genes y el número de aislados en los que están presentes (Figura 5), y la tercera es una gráfica con la frecuencia de genes contra el número de genomas ingresados (Figura 6). Esta secuencia de comandos está disponible para descarga en el repositorio https://github.com/sanger-pathogens/Roary/tree/master/contrib/roary_plots.

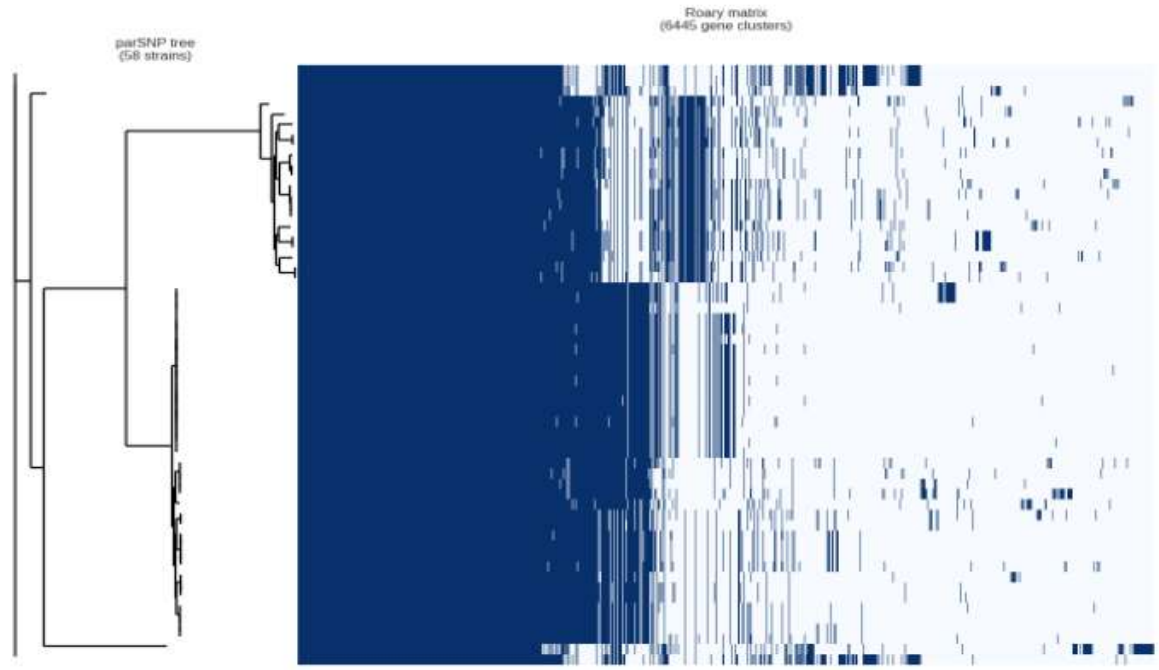

Figura 4. Diagrama de árbol-matriz de Roary.

Cada nodo es un genoma ingresado. Cada gen tiene una coloración azul y las zonas donde se encuentran muy concentrados indican que el mismo gen se comparte entre muestras.

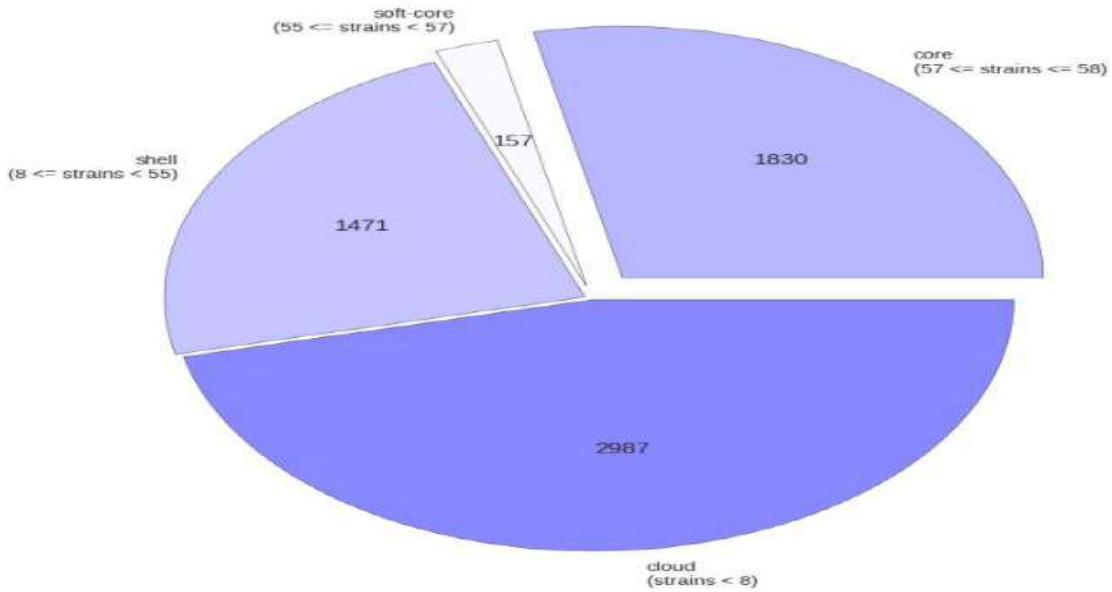

Figura 5. Gráfica de pastel de Roary.

Cada sección indica el número de genes comunes, genes accesorios, genes presentes en dos o más cepas (shell genes) y genes únicos específicos de cepas particulares (cloud genes).

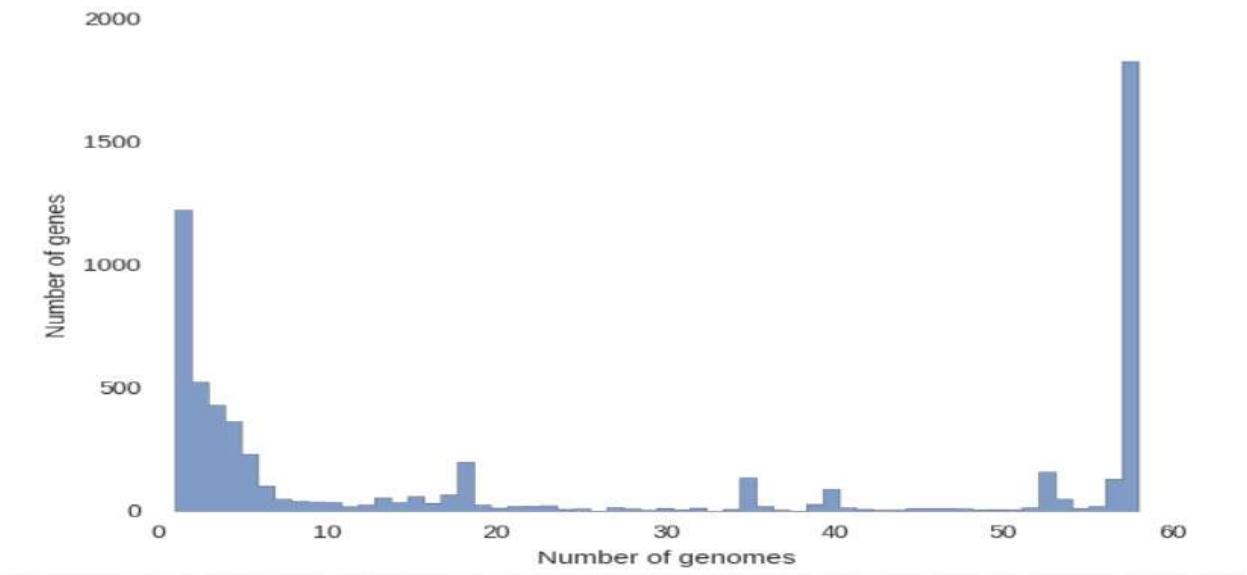

Figura 6. Gráfica de presencia y ausencia de genes por genoma.

Se observa que de los genomas analizados, 60 se comparten más de 1500 genes, mientras que 10 de los genomas analizados se comparten menos de 100 genes.

Conclusiones

El crecimiento de los pangenomas es enriquecido por eventos como la THG, la deriva genética, la duplicación, las introgresiones y los genomas secuenciados disponibles. Los factores más influyentes para determinar el tamaño de un pangenoma son el tamaño efectivo de población y la habilidad de migrar a nuevos nichos ecológicos (McInerney *et al.*, 2017). La mayoría de genes en la biósfera no se encuentran fuertemente unidos a un grupo específico de organismos. Es necesario hacer más investigaciones para comprender la relación existente entre los pangenomas y la THG, la selección, la deriva genética, la migración y el tamaño de la población.

Agradecimientos

Me gustaría agradecer al profesor Enrique González Vergara por introducirnos a la redacción científica y a

la Licenciatura en Biotecnología por brindarme las herramientas que hicieron posible la ejecución de este artículo.

Acknowledgements

I would like to thank Professor Enrique González Vergara for introducing us to scientific writing and to the biotechnology undergraduate program for providing me with the tools needed to carry out this paper.

Referencias

Bapteste, E., Lopez, P., Bouchard, F., Baquero, F., McInerney, J. O. y Burian, R. M. (2012). Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proceedings of the National Academy of Sciences*, *109*(45), 18266-18272. https://doi.org/10.1073/pnas.1206541109

Bowman, 2, Smyth, J. L., Meyerowitz, D. R., Mizukami, E. M. 3, Ma, Y., Krizek, H., Meyerowitz, L. L. (1991). Letters To Nature. *Nature Development Cell Development Nature Nature Genes Dev*, *353*(8), 31-37. https://doi.org/10.1038/35054089

Creevey, C. J., Fitzpatrick, D. A., Philip, G. K., Kinsella, R. J., O'Connell, M. J., Pentony, M. M., McInerney, J. O. (2004). Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society B: Biological Sciences*, *271*(1557), 2551-2558. https://doi.org/10.1098/rspb.2004.2864

Daubin, V. (2003). Phylogenetics and the Cohesion of Bacterial Genomes. *Science*, *301*(5634), 829-832. https://doi.org/10.1126/science.1086568

Ding, W., Baumdicker, F. y Neher, R. A. (2017). pan-X: pan-genome analysis and exploration. *Nucleic Acids Research*, 1-12. https://doi.org/10.1093/nar/gkx977

Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. *Science*, *284*(5423), 2124-2128. https://doi.org/10.1126/science.284.5423.2124

Ku, C., Nelson-Sathi, S., Roettger, M., Garg, S., Hazkani-Covo, E. y Martin, W. F. (2015). Endosymbiotic gene transfer from prokaryotic pan-genomes: Inherited chimerism in eukaryotes. *Proceedings of the National Academy of Sciences*, *112*(33), 10139-10146. https://doi.org/10.1073/pnas.1421385112

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*(2), 141-161. https://doi.org/10.1007/s10142-015-0433-4

Lapierre, P. y Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends in Genetics*, *25*(3), 107-110. https://doi.org/10.1016/j.tig.2008.12.004

Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature Biotechnology*, *28*(1), 57-63. https://doi.org/10.1038/nbt.1596

Lukjancenko, O., Wassenaar, T. M. y Ussery, D. W. (2010). Comparison of 61 Sequenced Escherichia coli Genomes. *Microbial Ecology*, *60*(4), 708-720. https://doi.org/10.1007/s00248-010-9717-3

Lynch, M. (2003). The Origins of Genome Complexity. *Science*, *302*(5649), 1401-1404.

https://doi.org/10.1126/science.1089370

Martinez-Murcia, A. J., Benlloch, S. y Collins, M. D. (1992). Phylogenetic Interrelationships of Members of the Genera Aeromonas and Plesiomonas as Determined by 16S Ribosomal DNA Sequencing: Lack of Congruence with Results of DNA-DNA Hybridizations. *International Journal of Systematic Bacteriology*, *42*(3), 412-421. https://doi.org/10.1099/00207713-42-3-412

McInerney, J. O., McNally, A. y O'Connell, M. J. (2017). Why prokaryotes have pan-genomes? *Nature Microbiology*, *2*(4), 17040. https://doi.org/10.1038/nmicrobiol.2017.40

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan-genome analysis. *Bioinformatics*, *31*(22), 3691-3693. https://doi.org/10.1093/bioinformatics/btv421

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068-2069. https://doi.org/10.1093/bioinformatics/btu153

Shapiro, B. J. (2016). How clonal are bacteria over time? *Current Opinion in Microbiology*, *31*, 116-123. https://doi.org/10.1016/j.mib.2016.03.013

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Fraser, C. M. (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), 13950-13955. https://doi.org/10.1073/pnas.0506758102

Treangen, T. J. y Rocha, E. P. C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genetics*, *7*(1). https://doi.org/10.1371/journal.pgen.1001284

Young, J. P., Crossman, L.C, Johnston, A.W., Thomson, N.R., Ghazoui, Z.F., Hull, K.H., Parkhill, J. (2006). The genome of Rhizobium leguminosarum has recognizable core and accesory components. *Genome Biology. 7*, R34.