Orthocore: herramienta para encontrar el core conservado en un linaje genómico.

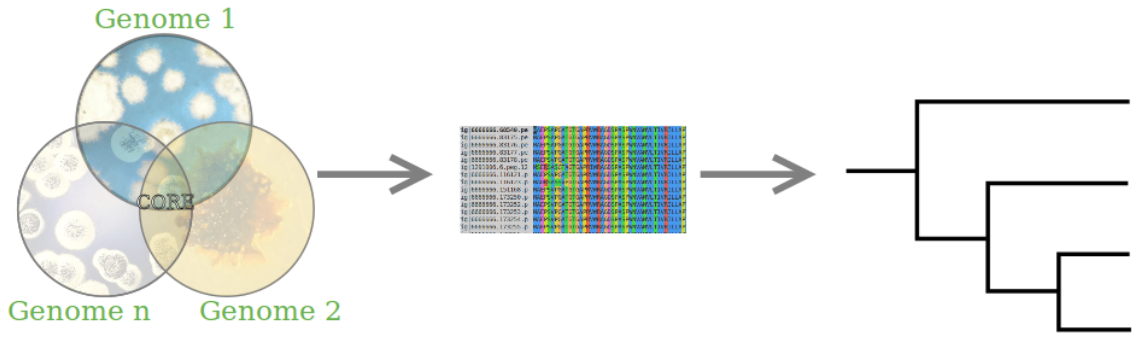


Figure 1: coreWiki

El pangenoma es el contenido génico total de un linaje taxonómico. Las familias génicas del pangenoma pueden clasificarse según sus patrones de presencia y ausencia en cada genoma del linaje. De acuerdo a esta clasificación los principales grupos de familias génicas en un pangenoma son el *core*, el *shell* y el *cloud (dispensable) genome*. El *core genome* es el conjunto de familias con presencia en todos los genomas del linaje. El *shell genome* es el grupo de familias presentes en la mayoría de los genomas, mientras que el *cloud genome* o *dispensable genome* es aquel grupo de familias que sólo ocurre en unos cuantos genomas.

Orthocore es un desarrollo bioinformático que realicé para calcular las familias génicas más conservadas del core genome. Orthocore obtiene el core conservado es decir familias de ortólogos presentes en todos los genomas del grupo, libres de parálogos e independientes de un genoma de referencia.

Los cambios en los perfiles de promiscuidad aparece en copias extras de familias de ortólogos y debido a la pérdida y ganancia de genes en organismos cercanos, se necesita un algoritmo para poder organizar filogenéticamente un organismo.

El core genome permite la obtención de filogenias complicadas

Dos secuencias son homólogas si poseen un ancestro común. Ortólogos y parálogos constituyen los dos principales tipos de homólogos. Los ortólogos provienen de eventos de especiación de un ancestro común mientras que los parálogos evolucionan por eventos de duplicación. La comparación de la variación molecular entre ortólogos ha sido utilizada para establecer relaciones filogenéticas entre organismos. Por ejemplo comparar la secuencia de 16s de RNA ribosomal condujo al descubrimiento del dominio Archaea en 1977 [@woese_phylogenetic_1977].

Los dos factores importantes para establecer las relaciones filogenéticas diferenciando entre Archaea, Bacteria y Eucarya son los siguientes: 1) la presencia conservada de la unidad de 16s en los tres dominios mencionados, y 2) la suficiente divergencia entre estas secuencias en organismos de estos dominios. Ahora bien, establecer relaciones filogenéticas entre Archaea y Bacteria es más sencillo que establecerlas entre organismos pertenecientes al mismo género o inclusive a la misma especie. En ocasiones como en el caso del género *Streptomyces* la secuencia de 16s por sí sola no posee la suficiente variación para resolver la filogenia [@labeda_phylogenetic_2017] ya que la variación entre estas secuencias suele ser menor al 1%.

Para resolver el problema de escasa variación en secuencias de 16s se puede concatenar la información de otros ortólogos, siempre que estos aparezcan en todos los organismos que se estén estudiando, es decir siempre que

sean parte del core genómico. Los ortólogos suelen buscarse por similitud de secuencia, pero estos métodos suelen también capturar parálogos, que pueden confundir la elucidación de eventos de especiación.

Tres aplicaciones de Orthocore serán presentadas en las siguientes secciones de este capítulo. En la primera aplicación el core conservado en *Actinomycetales* permitió organizar filogenéticamente este orden facilitando el entendimiento en cambios de promiscuidad en la enzima PriA mediante la distinción de patrones de pérdida y ganancia de genes en las rutas en las que esta enzima participa: histidina y triptofano. En la segunda aplicación y cepas cepas de *Salmonella Tifi*. Finalmente, en organismos del microbioma del tomate orthocore de utilizó para identificar genes marcadores que permitieran distinguir cepas de *Clavibacter Michiganensis* de otras especies. Para poder obtener una cantidad de secuencias conservadas con suficiente variación para una buena filogenia se necesita obtener el core.

Cómo el contexto genómico puede ser una marca para sugerir cambio funcional en una proteína. Se deseaba predecir si PriA estaba teniendo un cambio de promiscuidad debido a los patrones de pérdida y ganancia de genes en Actinomyces. Para poder obsrva patrones de pérdida y ganancia primero se necesitaba un árbol de especies. Este árbol de especies fue hecho con Orthocores.

Best Bidireccional Hits vs all vs all

Así pues, se necesitaba obtener los genes conservados del orden *Actinomycetales* para poder entender la promiscuidad enzimática de PriA. Existe el core, y el core conservado, uno es lo que hay en común en grupos de genomas, y otros lo que está listo para realizar la filogenia. En grupos muy grandes, de más de 100 genomas fragmentados puede quedar vacío. Para eliminar el sesgo de hacer Best Bidireccional Hits con un organismo, se diseñó en 2014 el método de las estrellas.

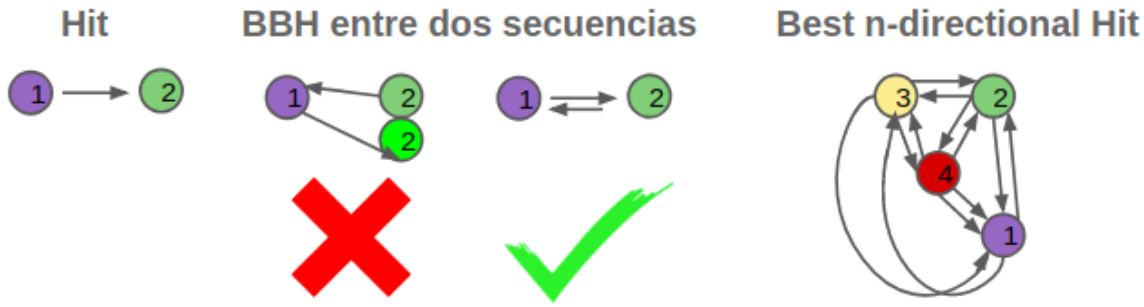


Figure 2: Best-n-directional

Otras aplicaciones de Orthocore

Orthocores resolvió la filogenia de Actinomyces, ayudando a encontrar perfiles de promiscuidad en PriA

Orthocores fue diseñado para resolver el problema de _____ La promiscuidad coocurre con variaciones en el contexto genómico. Orthocres ayudó a establecer la filogenia de actinomyces.

Identificación de genes marcadores de *Clavibacter michiganensis*

Micrococcales es un orden de Actinobacteria que contiene a *Clavibacter*, *Micrococcus* y *Microbacterium*, entre otros. Clavibacter es un género que puede causar enfermedades en plantas. En particular la especie

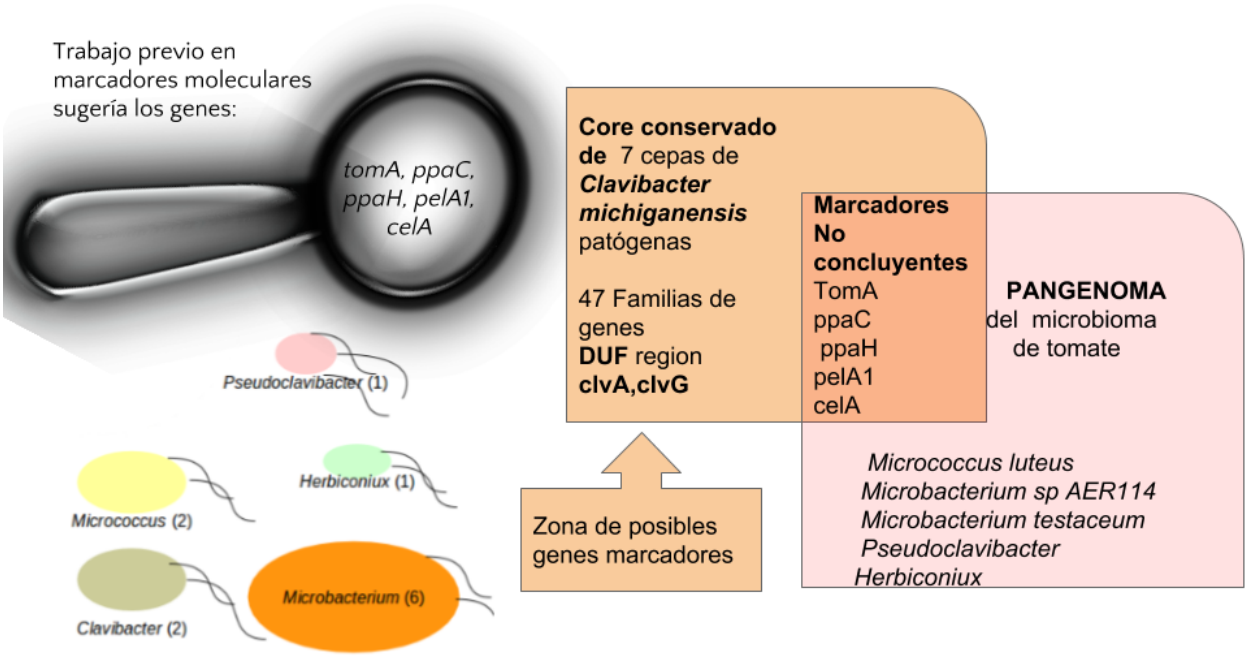


Figure 3: Marcadores-Clavibacter

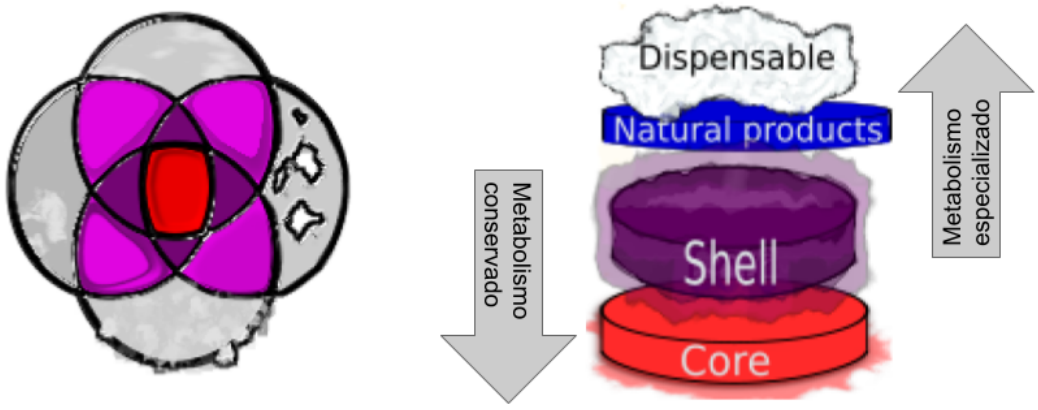


Figure 4: Metabolismo-Pangenoma

Clavibacter michiganensis es una bacteria causante de la enfermedad del cancer del tomate. Clavibacter ha sido frecuentemente aislada en compañía de otros micrococales morfológicamente parecidos, por lo que una prueba de diagnóstico se hacía necesaria. de y TomA clvF [@yasuhara-bell_genes_2014]

Había que estandarizar la anotación genpomica, para ello se utilizó la plataforma myRAST
https://github.com/nselem/myrast
Clavisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados

La idea de que orthocore puede ser usado para obtener los genes marcadores de un grupo taxonómico frente a otro fue generalizada en el backend del software Clavisual. Ya se ha explicado previamente que el core puede salir vacío por diversas razones, entre ellas baja calidad de los genomas, genomas provenientes de organismos muy divergentes o bien razones biológicas un core no convergente. Así pues, es posible que si sólo se utiliza el core no se obtengan marcadores. Pero el core puede relajarse de varias maneras una de ellas es el Pseudocore, como el core pero con un genoma de referencia, y a otra es establecer un porcentaje de presencia /ausencia de interés. El pseudocore consiste en _____ y la metodología está depositada en github en el repositorio_____. Los porcentajes de genomas son diferentes porque al no bastar los best bidireccional hits conservados, todo el pangenoma es decir todos los genes contenidos en los genomas del grupo de interés necesitan ser clasificados por familias, para de ahi obtener las familias que tienen presencia en un porcentaje %p y ausencia en un porcentaje a% del grupo externo. Estos perfiles fueron desarrollados para Clavisual utilizando FasthOrtho para clasificar las familias y de ahi obtener los grupos. Con ellos se consiguieron marcadores para Kutobacterium. Porque qué pasa

El blast del orthocore fue optimizado cambiando hacer un blast todos contra todos por archivos genómicos individuales genomai_vs_genomaj.blast y luego concatenando estos blasts

Salmonella

Orthocores fue usado aqui para reconstruir filogenias de genomas de *Salmonella* y como parte del CORASON el algoritmo que sirve para organizar filogenéticamente variantes de clusters ya sea biosintéticos, islas de patogenicidad, operones o cualquier región paricalmente sinténica de un genoma bacteriano centrada en un gen.

Relación entre genes marcadores, Orthocores y la promiscuidad enzimática.

Finalmente, al aplicar Orthocore para detectar genes marcadores se vuelve indirectamente reclutamientos al metabolismo especializado, cómo, pues porque dentro de los marcadores hay productos naturales como la clavidicina. Hay vemos que clvABCDEF participan en metabolismo secundario, las enzimas de metabolismo central de este cluster pueden presentar cierta promiscuidad.

```
# Example 1
# make a very simple plot
x <- c(2,3,4,5,6,7,8,9,10)
y <- c(2684,2504,2161,1867,1613,1217,745,432,320)
 plot(x,y, xlab="Número de genomas ", ylab="Número de Genes en el Core genome", main="Core Genome de or
```

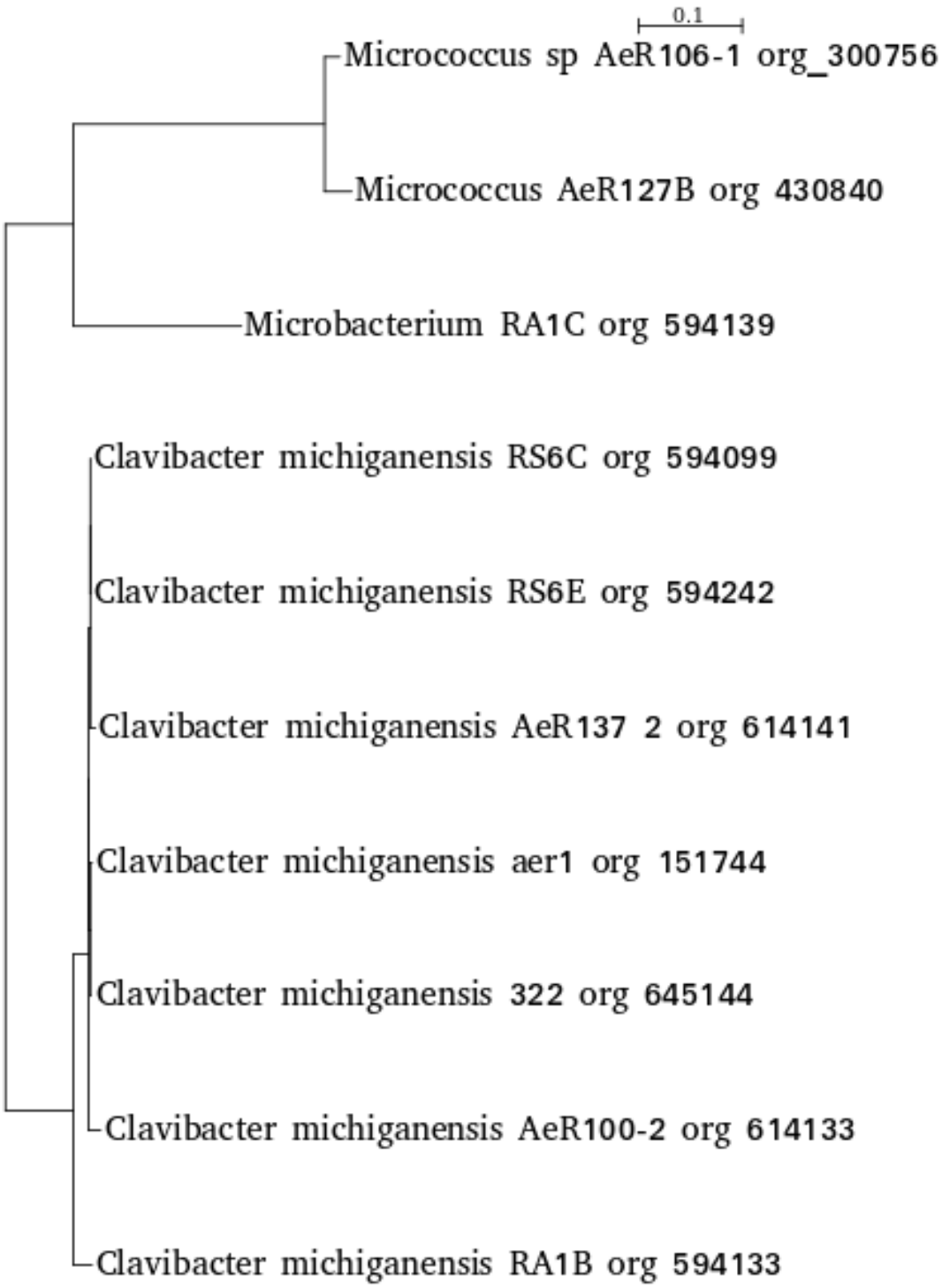


Figure 5: ArbolOrthocore

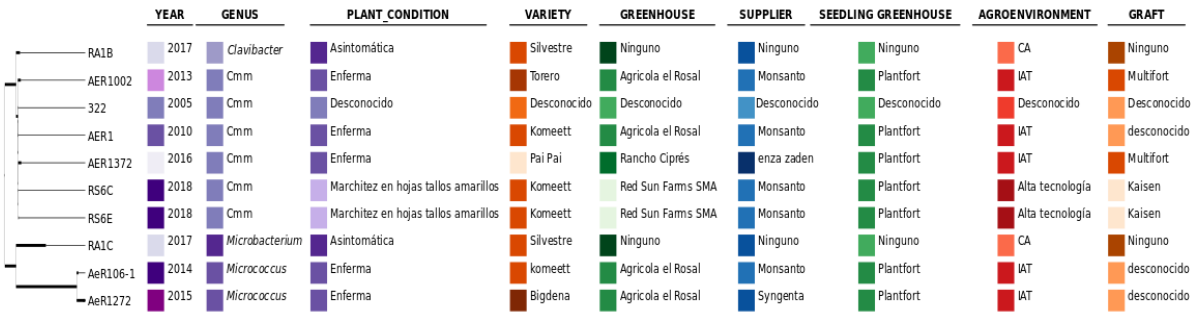


Figure 6: ArbolClavisual

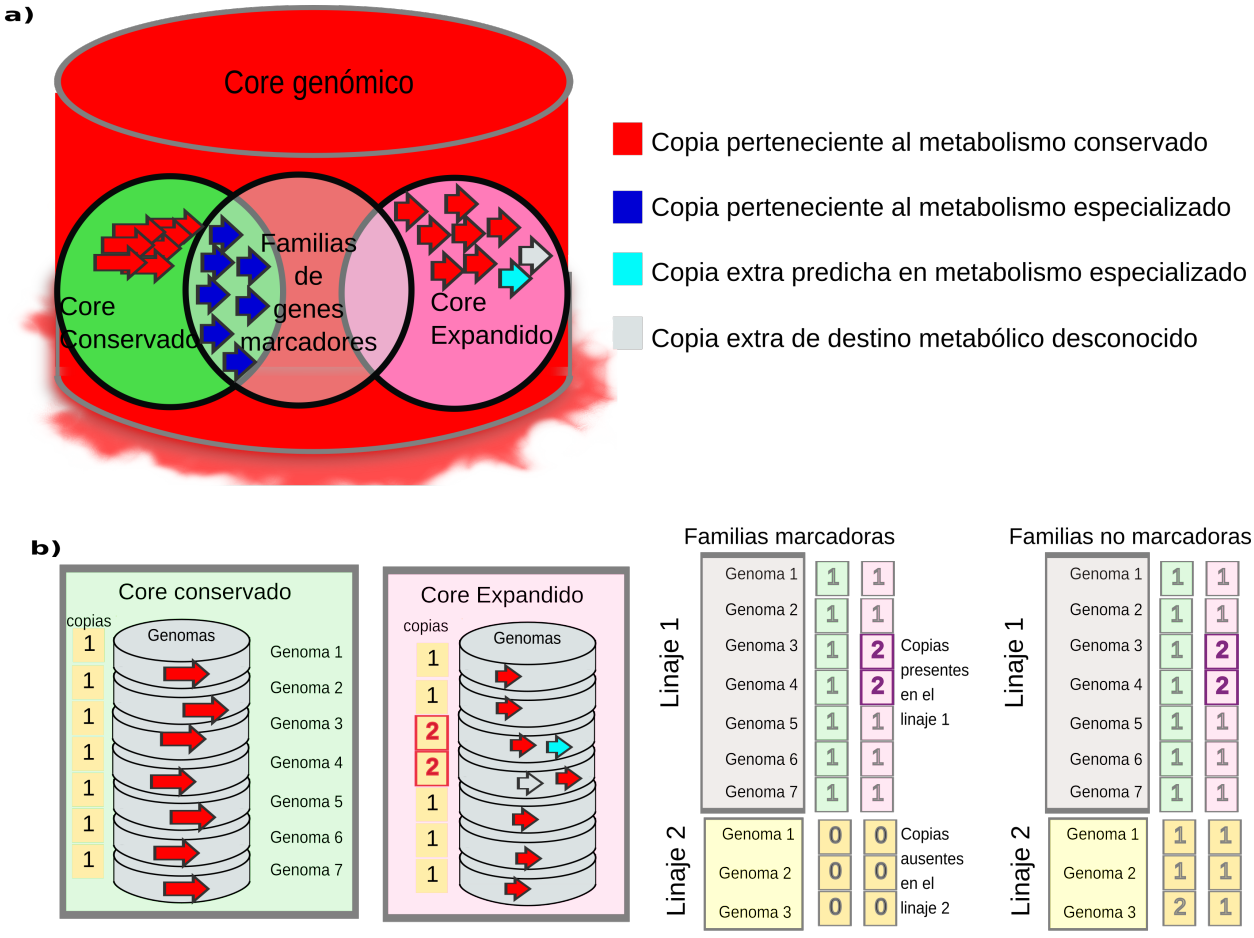
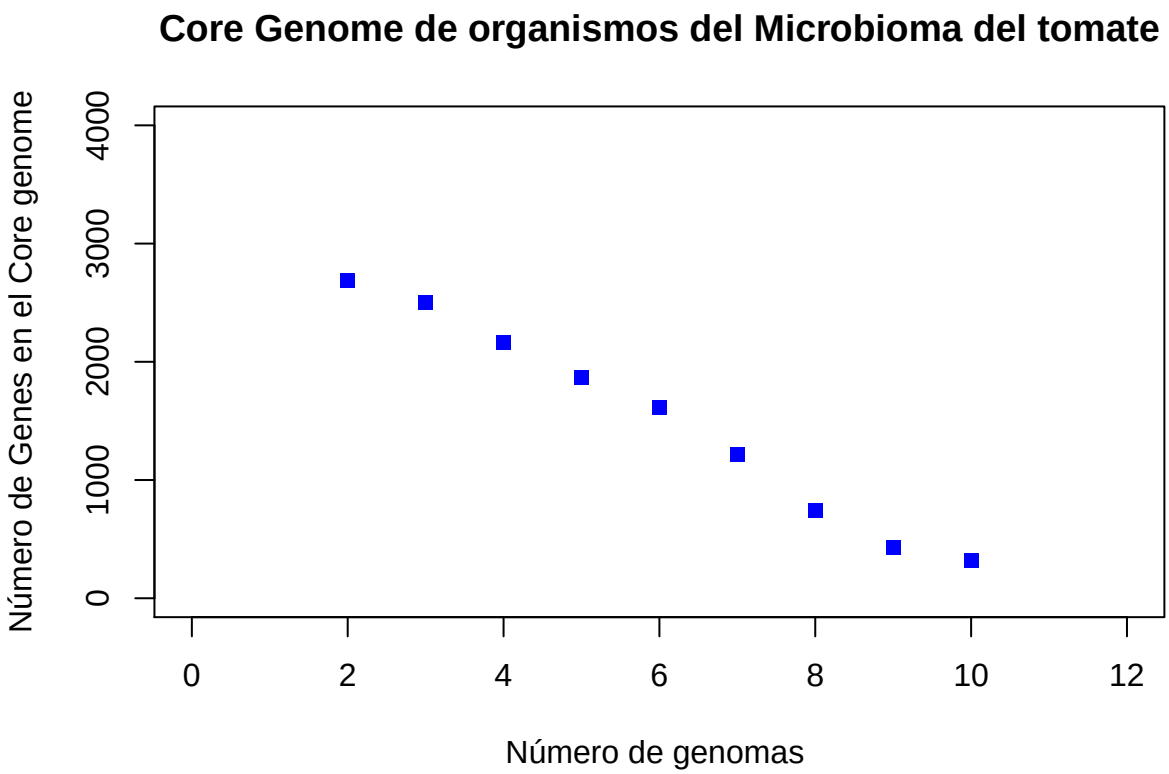


Figure 7: CoreMarcadores



Lo relevante para la promiscuidad son los procesos de divergencia, por ello en el siguiente capítulo analizamos encontrar el origin y destino metabolico, de copias extras de familias enzimáticas. Orthocore está disponible https://github.com/nselem/orthocore/wiki