

Anotación visualización del genoma del SARS-CoV-2

Prokka: rapid prokaryotic genome annotation

JBrowse

Adrián Camilo Rodríguez Ararat

Asistente de investigación grupo Natura, Universidad Icesi

email:acro83@gmail.com

Introducción

La anotación del genoma completo es el proceso de identificar características de interés en un conjunto de secuencias de ADN genómico y etiquetarlas con información útil. Prokka es una herramienta informática que permite anotar rápidamente genomas bacterianos, de arqueas, víricos. Produce archivos de salida conformes a las normas.

Prokka es una herramienta de software para anotar rápidamente genomas bacterianos, arqueológicos y virales, y producir archivos de salida que sólo requieren pequeños ajustes para enviarlos a GenBank/ENA/DBJ

JBrowse es un navegador del genoma rápido e integrable, construido completamente con JavaScript y HTML5. JBrowse-in-Galaxy (JiG) fue escrita para ayudar a construir instalaciones complejas de JBrowse directamente desde Galaxy, aprovechando las últimas características de Galaxy y con muchas funciones javascript para manejar la coloración de las características que serían casi imposibles de escribir sin la ayuda de esta herramienta además de ser integrable a Apollo.

Objetivos

Cargar datos en Galaxy

Anotar genoma con Prokka

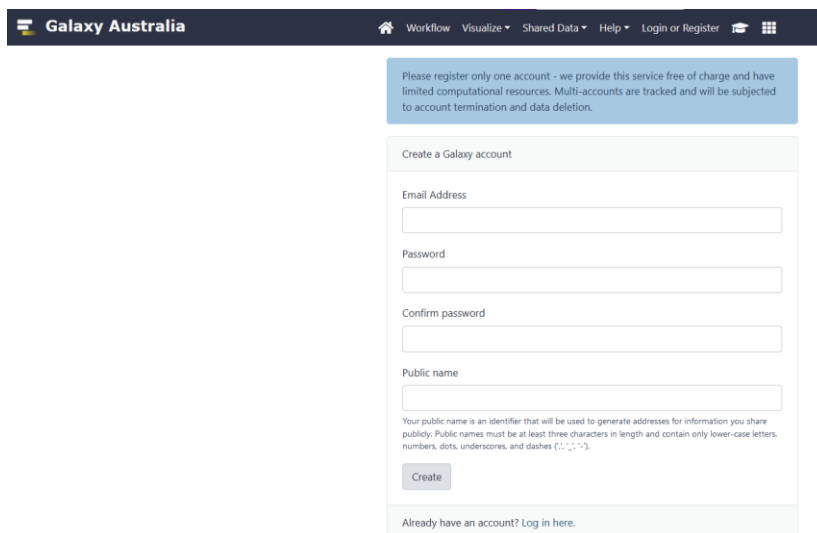
Visualizar anotaciones JBrowse

Requerimientos crear

Tiempo estimado 60 minutos

1. Crear cuenta en Galaxy (<https://usegalaxy.org.au/login>)

Introducir sus datos de usuario e iniciar la sección



The screenshot shows the 'Create a Galaxy account' form on the Galaxy Australia website. At the top, there is a navigation bar with the logo and links for Workflow, Visualize, Shared Data, Help, Login or Register, and a user menu. Below the navigation bar, a blue box contains a warning: 'Please register only one account - we provide this service free of charge and have limited computational resources. Multi-accounts are tracked and will be subjected to account termination and data deletion.' The form itself has a title 'Create a Galaxy account' and four input fields: 'Email Address', 'Password', 'Confirm password', and 'Public name'. Below the 'Public name' field, there is a small text block explaining that the public name is an identifier used for generating addresses for shared information and must be at least three characters long, containing only lower-case letters, numbers, dots, underscores, and dashes. A 'Create' button is located at the bottom of the form. At the very bottom, there is a link for users who already have an account: 'Already have an account? Log in here.'

2. Buscar en Tools Prokka o Se puede encontrar en "Annotation" en Galaxy.

Galaxy Australia

Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario

Tools

prokka

Upload Data

Show Sections

Prokka Prokaryotic genome annotation

barrnap Locate ribosomal RNA's in a fasta file. (GFF output)

Roary the pangenome pipeline - Quickly generate a core gene alignment from gff3 files

FLUJOS DE TRABAJO

All workflows

Prokka Prokaryotic genome annotation (Galaxy Version 1.14.6+galaxy0)

Favorite Versions Options

Contigs to annotate

9: Prokka on data 1: fsa

FASTA format

Locus tag prefix

(--locustag)

Locus tag counter increment

1

(--increment)

GFF version

3

(--gffver)

Force GenBank/ENA/DDJB compliance

No

Equivalent to --addgenes --mincontiglen 200 --centre Prokka (or other centre specified below) (--compliant)

Add 'gene' features for each 'CDS' feature (--addgenes)

No

Minimum contig size (--mincontiglen)

200

NCBI needs 200

Sequencing centre ID

(--centre)

Se anotará el genoma del SARS-CoV-2, identificaremos los genes putativos y luego los compararemos con una base de datos de genes conocidos para encontrar cuáles se alinean mejor. La caracterización estructural de los genes putativos puede hacerse de varias maneras. Una forma de encontrar estas regiones requiere la observación de que las regiones de ADN que se traducen en proteínas comienzan con un codón de inicio (ATG) y terminan con un codón de parada (TAG, TAA o TGA). Si un genoma fuera aleatorio, entonces después de encontrar un codón de inicio, esperaríamos ver un codón de parada después de unos $64/3 \sim 21$ codones. Pero los genes reales suelen ser mucho, mucho más largos que 21 codones. En consecuencia, en organismos simples como los virus y las bacterias, una buena forma de encontrar genes es buscar largos tramos de codones que conecten un codón de inicio con un codón de parada, sin codones de parada intermedios. Una vez se determinan los genes putativos a lo largo del genoma, podemos compararlos con una base de datos de genes utilizando un algoritmo llamado BLAST con el fin de validar/anotar características estructurales y funcionales.

Estos dos pasos los realiza nuestra siguiente herramienta, llamada Prokka, que se utiliza para anotar los genomas de virus, bacterias y arqueas.

3. En "Contigs to annotate" -> "Browse Dataset" -> "Upload", cargue los archivos `pretrim.sars-cov2.4.fa` y `all_consensus.fasta`. Ajuste el "Reino" a "Viruses" y deje todos los demás parámetros por defecto.

Galaxy Australia Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario Using 0%

Tools ☆

prokka

Upload Data

Show Sections

Prokka Prokaryotic genome annotation

barrnap Locate ribosomal RNA's in a fasta file. (GFF output)

Roary the pangenome pipeline - Quickly generate a core gene alignment from gff3 files

FLUJOS DE TRABAJO

All workflows

Prokka Prokaryotic genome annotation (Galaxy Version 1.14.6+galaxy0)

☆ Favorite Versions Options

Contigs to annotate

9: Prokka on data 1: fsa

Browse Datasets

FASTA format

Locus tag prefix

(--locustag)

Locus tag counter increment

1

(--increment)

GFF version

3

(--gffver)

Force GenBank/ENA/DDJB compliance

No

Equivalent to --addgenes --mincontiglen 200 --centre Prokka (or other centre specified below) (--compliant)

Add 'gene' features for each 'CDS' feature (--addgenes)

No

Minimum contig size (--mincontiglen)

200

NCBI needs 200

Sequencing centre ID

(--centre)

History + - ⚙

buscar conjuntos de datos ?

Anotación genoma SARS-CoV-2

14 shown

1.68 MB

g

13: Prokka on data 1: tx t

12: Prokka on data 1: err

11: Prokka on data 1: ts v

10: Prokka on data 1: tbl

9: Prokka on data 1: fsa

8: Prokka on data 1: sqn

7: Prokka on data 1: ffn

6: Prokka on data 1: faa

5: Prokka on data 1: fna

4: Prokka on data 1: gbk

3: Prokka on data 1: gff

2: pretrim.sars-cov2.4.fa

1: all_consensus.fasta

Tenga en cuenta que Prokka sólo anota contigs de una longitud mínima de 200. Seleccione si desea recibir un correo electrónico al finalizar y haga clic en "Ejecutar". EL proceso de anotación toma segundos.

En primer lugar, vea el archivo .gff que contiene las regiones identificadas por Prokka como genes putativos. ¿Cuántas hay? ¿Cuál es la más larga y la más corta?

4. Para visualizar nuestra anotación contenida en el archivo .gff, utilizaremos una herramienta de navegación del genoma llamada "JBrowse" que se encuentra en la sección "Tools" en la parte izquierda de la página.

Galaxy Australia Flujo de Trabajo Visualizar Datos Compartidos Ayuda Usuario Using 0%

Tools ☆

jbrowse

Upload Data

Descargar desde URL o descargar desde disco

JBrowse - Data Directory to Standalone upgrades the bare data directory to a full JBrowse instance

JBrowse genome browser

FLUJOS DE TRABAJO

All workflows

JBrowse genome browser (Galaxy Version 1.16.11+galaxy1)

☆ Favorite Versions Options

Reference genome to display

Use a genome from history

Built-in references

Select the reference genome

9: Prokka on data 1: fsa

Output JBrowse

Minimal for viewing (Documentation removed)

Genetic Code

1. The Standard Code

JBrowse-in-Galaxy Action

New JBrowse Instance

Track Group

+ Insert Track Group

General JBrowse Options [Advanced]

Plugins

Email notification

No

Send an email notification when the job completes.

Execute

History + - ⚙

buscar conjuntos de datos ?

Anotación genoma SARS-CoV-2

14 shown

1.68 MB

[tbl2asn-forever] Correcting dates in outdir/prokka.gbf

[tbl2asn-forever] Correcting dates in outdir/prokka.sqn

[tbl2asn-forever] Dates changed from 01-JAN-2019 to 02-SEP-2021

organism: Genus species strain

contigs: 6

bases: 179376

CDS: 90

12: Prokka on data 1: err

11: Prokka on data 1: ts v

10: Prokka on data 1: tbl

9: Prokka on data 1: fsa

8: Prokka on data 1: sqn

7: Prokka on data 1: ffn

6: Prokka on data 1: faa

Al ejecutar JBrowse, siga los siguientes pasos.

Genoma de referencia a mostrar. Seleccione "Use a genome from history" y elija el archivo .fna.
Haga clic en "Insert Track Group".
Haga clic en "Insert Annotation Track".
Seleccione su archivo .gff de la salida de Prokka.
En "Email notification", elija "Sí" si lo desea.
Luego haga clic en "Execute".

JBrowse debería ser muy rápido. Cuando termine, haga clic en ver archivo, que es un archivo HTML que podemos ver en el navegador. (Puede tardar un momento en cargarse).

haga clic en "Prokka on data XXX: gff" para mostrar nuestra hermosa anotación del genoma del SARS-CoV-2. Amplía la imagen para verla en todo su esplendor. Todas las flechas apuntan en la misma dirección para indicar que los genes se encuentran todos en la misma cadena del genoma. Esto tiene sentido porque el SARS-CoV-2 es un virus de ARN, lo que significa que su genoma sólo tiene una hebra. (Habría sido una muy mala señal que algunos genes apuntaran en la dirección contraria).

Referencias

- Cuccuru, G., Orsini, M., Pinna, A., Sbardellati, A., Soranzo, N., Travaglione, A., ... Fotia, G. (2014). Orione, a web-based framework for NGS analysis in microbiology. *Bioinformatics*, 30(13), 1928–1929. <https://doi.org/10.1093/bioinformatics/btu135>
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
- Seemann T. Prokka: rapid prokaryotic genome annotation *Bioinformatics* 2014 Jul 15;30(14):2068-9. PMID:24642063. <https://github.com/tseemann/prokka>
- prokka (Version 1.14.6)
- Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., & Holmes, I. H. (2009). JBrowse: A next-generation genome browser. *Genome Research*, 19(9), 1630–1638. <https://doi.org/10.1101/gr.094607.109>
- bio.tools: jbrowse url: <https://bio.tools/jbrowse>
- Phillip Compeau, SARS-CoV-2 Software Assignment: Genome Assembly and Annotation url: <http://compeau.cbd.cmu.edu/online-education/sars-cov-2-software-assignments/covid-19-genome-assembly-assignment/>