CFT 2: TECNOLOGÍAS ÓMICAS: Asignación de Variantes usando Datos de Secuenciación Masiva (Variant Calling Using "NGS" Data)

Fecha: 19/10/2017 - 20/10/2017

Horario: 10:00-13:00 / 14:30-16:30

Metodología didáctica: curso teórico-práctico con ejercicios presenciales y no presenciales. Curso presencial en el aula de impartición

Coordinador: David Posada, Universidade de Vigo (dposada@uvigo.es)

Profesor: Fernando Cruz, Centro Nacional de Análisis Genómico-Centro de Regulación Genómica (CNAG-CRG, Barcelona)

DESCRIPCIÓN

Este curso pretende aportar una visión general de los datos de secuenciación masiva (NGS), centrándose fundamentalmente en aquellos producidos por los secuenciadores de Illumina y sus principales aplicaciones en la investigación Genómica. Además proporciona una formación básica en el uso de técnicas bioinformáticas para la gestión, manipulación y análisis de datos de secuenciación. Se ahondará en el método de asignación de variantes de una sola base (SNVs) basado en mapeo a un genoma referencia, siguiendo el protocolo recomendado por GATK (Broad Institute). Finalmente, se realizaran algunos ejercicios para manipular, filtrar ficheros de SNVs y calcular algunos estadísticos poblacionales.

Contenidos

1. Tipos de Variación Genómica.
2. Datos de Secuenciación Masiva. Ventajas y Limitaciones.
3. Formatos de archivo para datos de secuenciación, alineamiento y asignación de variantes.
4. Herramientas Bioinformáticas (FastQC, BWA, Picard, GATK, VCFTools)
5. Asignación de Variantes basada en Mapeo (GATK Best Practices)
6. Descripción general de otros métodos para asignar variantes

Programa docente

Este curso se centrará principalmente en las herramientas necesarias para el análisis de datos de secuenciación masiva y su aplicación a la detección de variantes genómicas (haciendo hincapié en las variantes de una sola base o SNVs). Se organizará en dos clases presenciales de 5 horas cada una y 15 horas de trabajo individual del alumno. Las clases presenciales incluirán pequeños ejercicios prácticos con la intención de asentar los conceptos teóricos expuestos y orientar al alumno cara al trabajo no presencial. La estructura del curso será la siguiente:

Parte I (no presencial): 2-6/10/2017

- 5 horas de ejercicios no presenciales (Tutorial de comandos Unix imprescindibles y *Gawk* scripts de una sola línea)

Parte II (presencial): 19-20/10/2017

- 10 horas presenciales consistentes en teoría (4h) y ejercicios prácticos (6h), distribuidas en dos sesiones de 5 horas.

Parte III (no presencial): 23/10/2015-06/11/2017

- 10 horas no presenciales. Aplicación del Protocolo a otros datos (5h) y análisis de SNVs (5h).

Metodología docente

Este curso pretende dotar al estudiante de una base teórica imprescindible para comenzar a analizar datos de secuenciación masiva y de habilidades bioinformáticas que se irán consolidando mientras aprenden el protocolo de asignación de variantes (del inglés *variant calling*) recomendado por GATK. Además de las clases en el presenciales, los alumnos tendrán que completar algunos trabajos prácticos y cuestiones relativas a la interpretación de los datos por cuenta propia (individualmente o en parejas). Se proporcionará una guía específica para todos los temas a tratar y se garantiza la tutoría a lo largo de todo el curso.

Aula y medios disponibles

Las clases presenciales se impartirán en un aula que permita conexión a internet para poder acceder a la plataforma online que contendrá los protocolos, tutoriales y ejercicios prácticos del curso. Dicha plataforma será una materia Moodle del Campus do Mar. Los alumnos deberán usar su ordenador portátil en el que si es necesario instalarán, con anterioridad a las clases presenciales, un programa para usar una terminal *ssh* (p.ej. PuTTY o MobaXterm en Windows[1]). Esta terminal permitirá conectarse a una cuenta en un servidor Unix con los programas y los datos necesarios para las prácticas.

[1]NOTA: En sistemas operativos Linux y MacOS no es necesario ya que viene instalada por defecto.

Sistema de Evaluación

Se evaluarán los siguientes aspectos:

- Asistencia y participación: 20%
- Ejercicios presenciales: 30%
- Ejercicios no presenciales: 50%

Profesores del curso

Dr. Fernando Cruz: Desde 2014 trabaja como Bioinformático post-doctoral en el Centro Nacional de Análisis Genómico (CNAG, Barcelona) dentro del grupo de "Genome Assembly and Annotation" dirigido por Tyler Alioto. Tras realizar varios trabajos de genética de poblaciones y molecular, obtuvo su doctorado en la Universidade de Vigo en 2005.A continuación, se trasladó al

Smurfit Institute of Genetics del Trinity College Dublin (Irlanda) dónde realizó un estudio de Genómica comparada en el que empleó herramientas bioinformáticas. Posteriormente, trabajó durante un año en la Universidad de Uppsala (Suecia) y dos años y medio en el Laboratorio de Bioinformática Evolutiva de la Universidad de Lausana (Suiza). En 2011, es contratado como investigador post-doctoral en la Estación Biológica de Doñana (EBD-CSIC, Sevilla) a través del proyecto Europeo EcoGenes EU FP7. En dicho centro se integra en el proyecto del Genoma del Lince Ibérico donde, empleando datos de re-secuenciación de 11 linces, se encargó de identificar SNPs, realizar análisis de Genómica Poblacional y reconstruir la historia Demográfica de este félido.

Su experiencia docente incluye: Profesor del curso CFT2 ”Variant Calling Using NGS Data” en el Campus Do Mar durante la edición 2014-2015. Co-organizador y monitor de prácticas del Curso “Introduction to the Analysis of Next Generation Sequencing Data” coordinado por Matthew T. Webster, Uppsala University (19h) - 5-8 Marzo, 2013, EBD-CSIC, Sevilla; Profesor de apoyo en el segundo año del curso “Introduction to Bioinformatics” del Master Program of the Biological School (12h) - University of Lausanne, 2009; Profesor del curso “Problem Based Learning” del Master Program of the Biological School (42h) - University of Lausanne, 2008; Profesor de prácticas de Genética General (25h) y Técnicas en Genética (45h) - Universidade de Vigo, 2003-2005.

Sus publicaciones son:

- F. Abascal*, A. Corvelo*, F.Cruz* et al. (2016) Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biology, 17(1): 251-268
- F. Cruz*, I. Julca*, J. Gómez-Garrido, D. Loska, M. Marcet-Houben, E. Cano, B. Galán, L. Frias, P. Ribeca, M. Gut, M. Sánchez-Fernández, J.L. García, I.G. Gut, P. Vargas, T.S. Alioto y T. Gabaldón (2016) Genomics data from the Mediterranean olive tree. GigaScience GigaScience, 5(29): 1-12
- F. Cruz, Adrian C. Brennan, Alejandro Gonzalez-Voyer, Violeta Muñoz-Fuentes, Muthukrishnan Eaaswarkhanth, Séverine Roques and F. Xavier Picó (2012) Genetics and Genomics in Wildlife Studies: Implications for Ecology, Evolution and Conservation Biology. BioEssays, 34(3): 245-246.
- F. Cruz, J. Roux and M. Robinson-Rechavi (2009). The expansion of amino-acid repeats is not associated to adaptive evolution in mammalian genes. BMC Genomics, 10(1), 619.
- F. Cruz, C. Vilà and M. T. Webster (2008) The legacy of domestication: Accumulation of deleterious mutations in the dog genome. Molecular Biology and Evolution, 25 (11): 2331-2336
- F. Cruz, D. G. Bradley and D. J. Lynn (2007) Evidence of Positive Selection on the Atlantic Salmon CD3gamma-delta Gene. Immunogenetics 59(3): 247-251
- F. Cruz, Pérez M. and P. Presa (2005) Distribution and abundance of microsatellites in the genome of bivalves. Gene 346: 241-247
- M. Pérez, F. Cruz and P. Presa (2005) Distribution Properties of Polymononucleotide Repeats in molluscan genomes. J. Heredity 96(1): 40-51
- A. Pérez-Figueroa, F. Cruz, A. Carvajal-Rodríguez, E. Rolán-Alvarez and A. Caballero (2005) The evolutionary forces maintaining a wild polymorphism of *Littorina saxatilis*: model selection by computer simulations. J. Evol. Biol. 18(1): 191-202

*Los autores con este superíndice han contribuido igualmente a este trabajo.