



Matthieu J. Miossec

twitter: @RealMattJM

Unidad 9: Análisis genómicos reproducibles en la nube



Programa Unidad 9

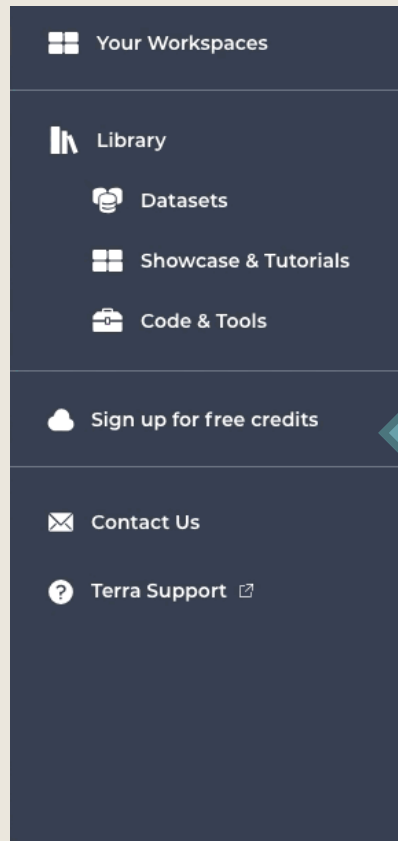
- 11 de mayo (hoy!) – Introducción a la genómica en la nube con Terra.
- 13 de mayo (miércoles) – GATK ‘Best Practices’ GVCF Workflow [en Terra]
- 18 de mayo (lunes) – Otras herramientas [en Terra]:
 - Mutect2 (variantes somáticas, cáncer)
 - GermlineCNVCaller (CNVs de la línea germinal) [en Terra]

Ultima oportunidad...

¿Han creado cuenta y pedido créditos?

■ Tener una cuenta en Terra **no cuesta nada**.


- Se puede definir espacios de trabajos, copiar y construir workflows (y más) sin costo.
- Lo que si genera costo son los servicios proporcionados por la plataforma Google Cloud.
 - Computación
 - Almacenamiento
 - Descarga de datos



Cada nuevo usuario puede aprovechar de **\$300 (US)** en créditos nube que pueden ser usados durante **60 días***.

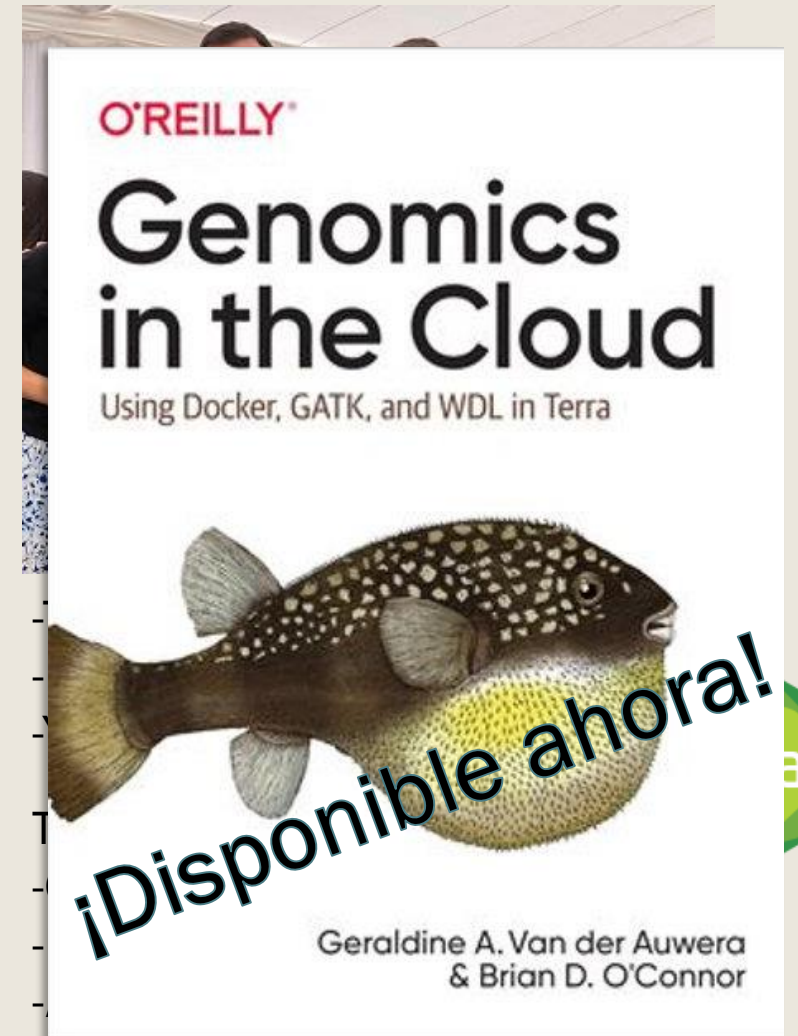
*Google Cloud también tiene una oferta idéntica, pero necesita ingresar una tarjeta de crédito. ☹

Agradecimientos

- Data Sciences Platform  **BROAD**
INSTITUTE
Broad Institute of Harvard and MIT
<https://gatk.broadinstitute.org/>

Por los materiales (Terra, workflows, docs...), gráficos y el apoyo otorgado durante la preparación de la clase de 2019.

Agradecimiento especiales,
al equipo de Viña del Mar (Nov 2018)



-Kate Noblett

¿Por qué trabajar en la nube?

- **Trabajar y compartir muchos datos en línea sin deber almacenarlos localmente.**
- Facilita colaboraciones internacionales, se comparte más fácilmente.
- Alternativa a un HPC local difícil de acceso o sobreocupado o inadecuado.
- Permite probar nuevas herramientas sin preocuparse tanto de la instalación (no necesita reinventar la rueda).
- Hace que todo un estudio sea reproducible (y accesible).

Genómica en la nube



<https://aws.amazon.com/health/genomics/>



<https://azure.microsoft.com/en-us/services/genomics/>



Google Cloud

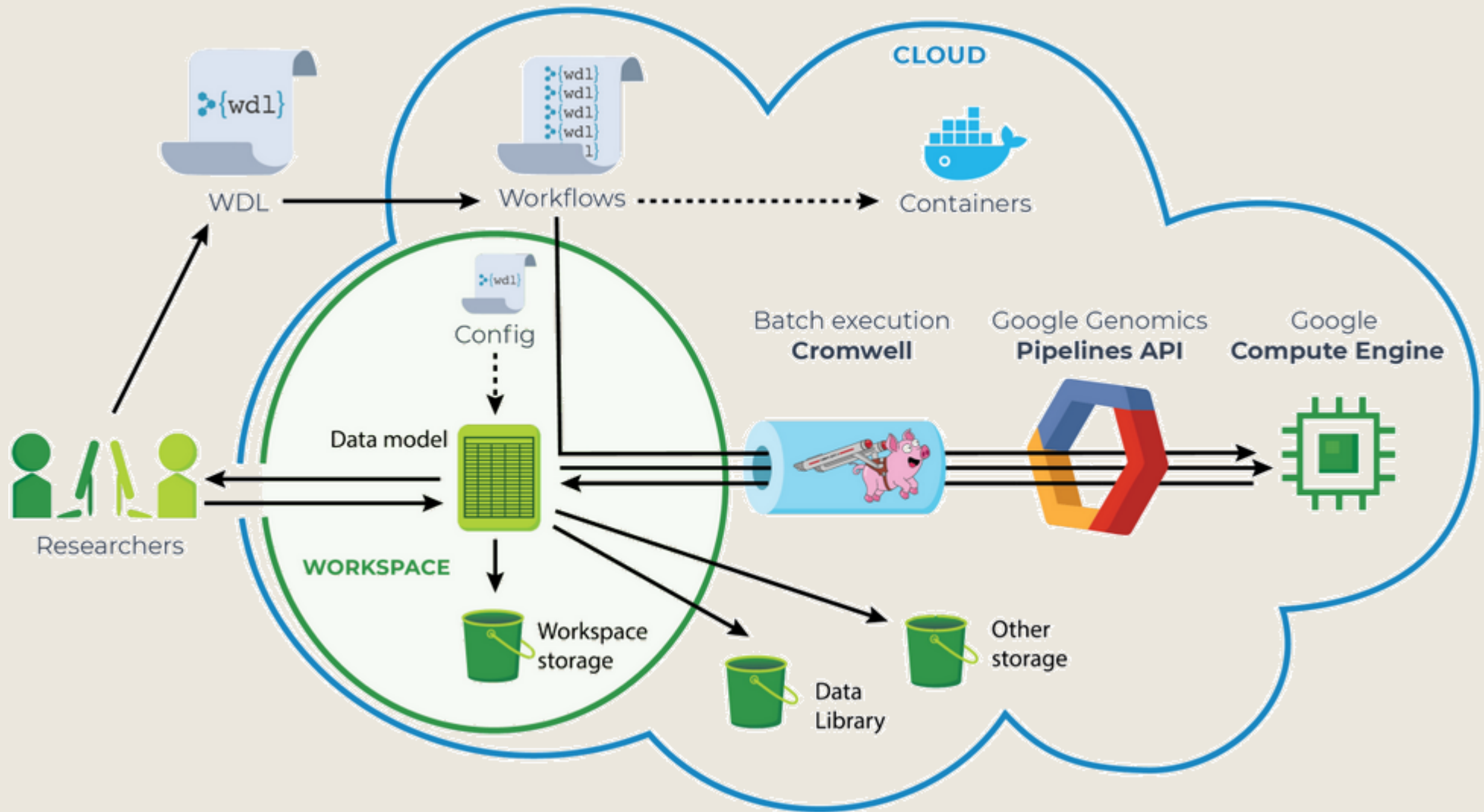
<https://cloud.google.com/life-sciences>



¿Qué es Terra?

- A primera vista, una **plataforma** diseñada para **investigación biomédica** que permite trabajar **en la nube**.
- En realidad, ofrece mucho más:
 - **Recursos:** Una librería completa de datos y métodos.
(inc. todos los workflows GATK, automatizado de punto a punto!)
 - **Compartimiento:** Todo (datos, nuevo métodos [Docker], espacios de trabajo Terra) es compartible (con algunos colaboradores o con toda la comunidad Terra).
 - **Análisis en tiempo real:** Los resultados de un análisis pueden ser organizados y manipulados desde Terra usando el **Jupyter notebook**.


La Arquitecta Terra



Workspace

- Corresponde a un espacio de trabajo bien delineado.
 - Se adjunta a un proyecto de facturación cuando se crea.
 - Se puede compartir con otros investigadores, como dueño tengo la posibilidad de restringir el acceso otorgado:
 - (Project) Owner → Dueño: Todos los derechos sobre un 'workspace'.
 - Writer → Escritor: Puede crear/modificar metadata, configuración de métodos..
 - Reader → Lector: Puede ver el contenido de un 'workspace' pero no modificarlo
 - 2 opciones: Permiso para **ejecutar** y permiso para **compartir**

Google Bucket

- Cada 'workspace' tiene su 'Google bucket'  (Cubo Google) dedicado en el cual...
 - Subimos nuestros datos iniciales (ej. FASTQ, BAM/BAI).
 - Los datos generados a través de la ejecución de herramientas en el workspace están almacenados.
 - Podemos descargar datos del Cubo Google...por un precio (típicamente pequeño).

(El costo de almacenamiento esta cubierto por el proyecto de facturación destacado al 'workspace')

Datos de Referencia

(sin costo de almacenamiento!)

- Los **datos de referencia** que se usan comúnmente durante el análisis de secuencias genómicas...

- Genoma humano de referencia (hg19/b37 o hg38, .fasta)
- Las variantes de las base de datos dbSNP/HapMap/1000G...(vcf)

- ...Están proporcionados por la plataforma Terra!



- No tiene ningún costo de almacenamiento para nosotros!
- Es crucial no gastar recursos subiendo lo que ya esta disponible.

- Esto vale también por algunos archivos test que existen para probar la plataforma (ejemplos más tarde).

Datos y Metadatos: Estructuración de los datos

