

Introducción a la genómica en la nube con



Dr. Matthieu J. Miossec (@RealMattJM)

Bioinformatics analyst @ Wellcome Centre for human genetics



UNIVERSITY OF
OXFORD

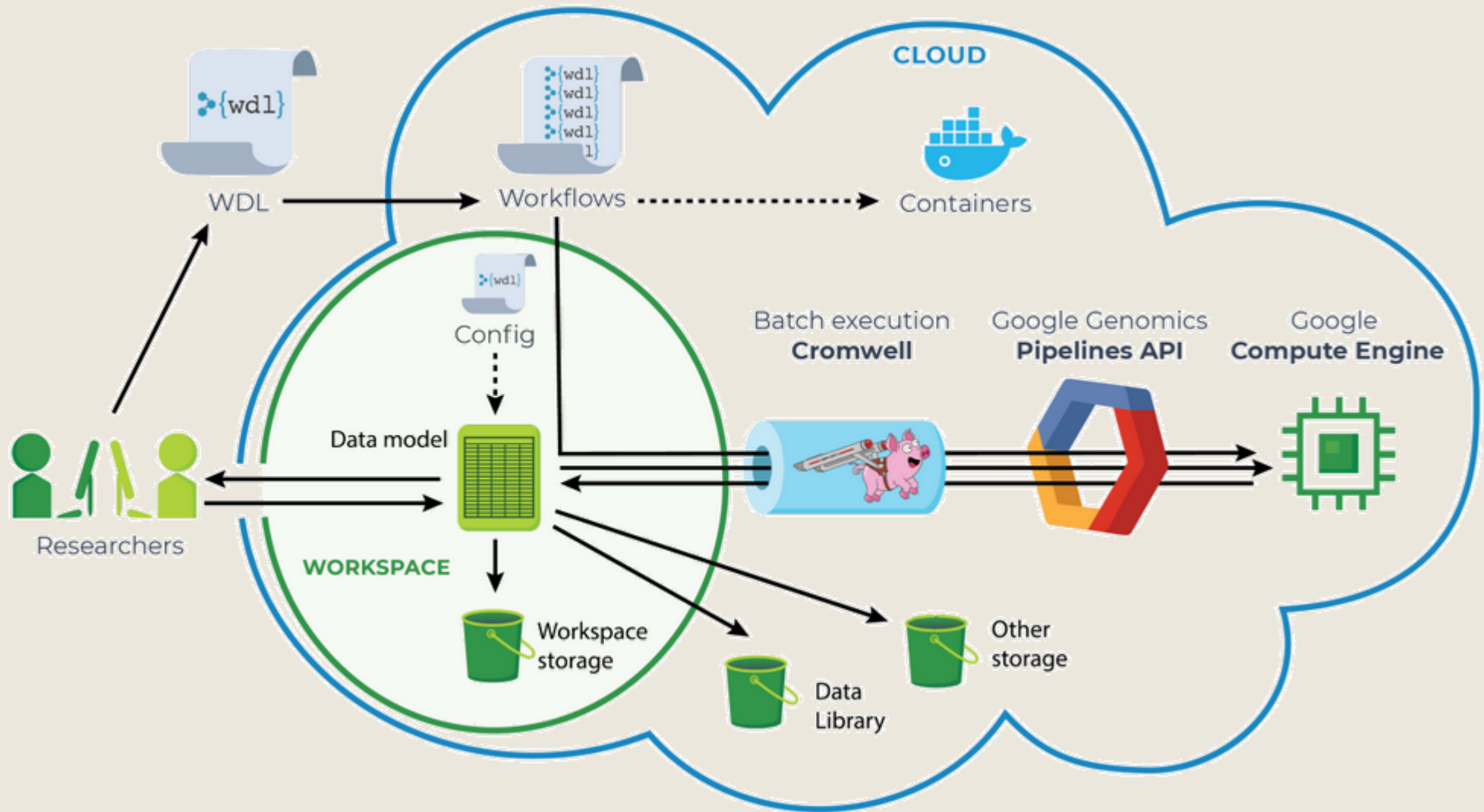
Programa Unidad 9

- 17 de mayo (lunes) – Introducción a la genómica en la nube
- 19 de mayo (miércoles) – Introducción a la plataforma Terra
- 24 de mayo (lunes) – Otras herramientas en Terra

¿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/1000G/GnomAD...(vcf)

- ...Están proporcionados por la plataforma Terra!

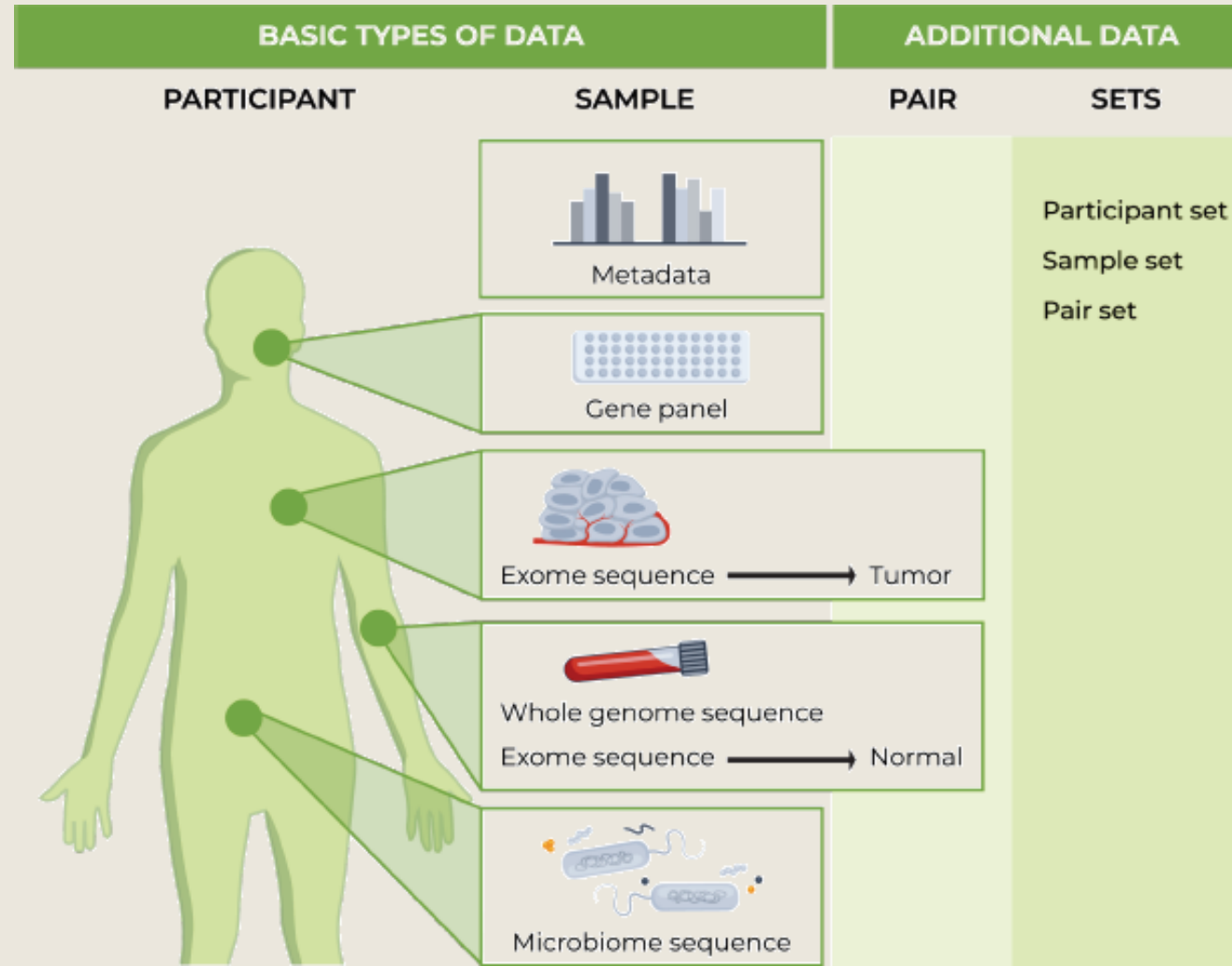


- No tiene ningún costo de almacenamiento para nosotros!
- Es crucial no gastar recursos subiendo lo que ya esta disponible.

- Esto vale también por algunos archivos test que existen para probar la plataforma.

<https://cloud.google.com/life-sciences/docs/resources/public-datasets>

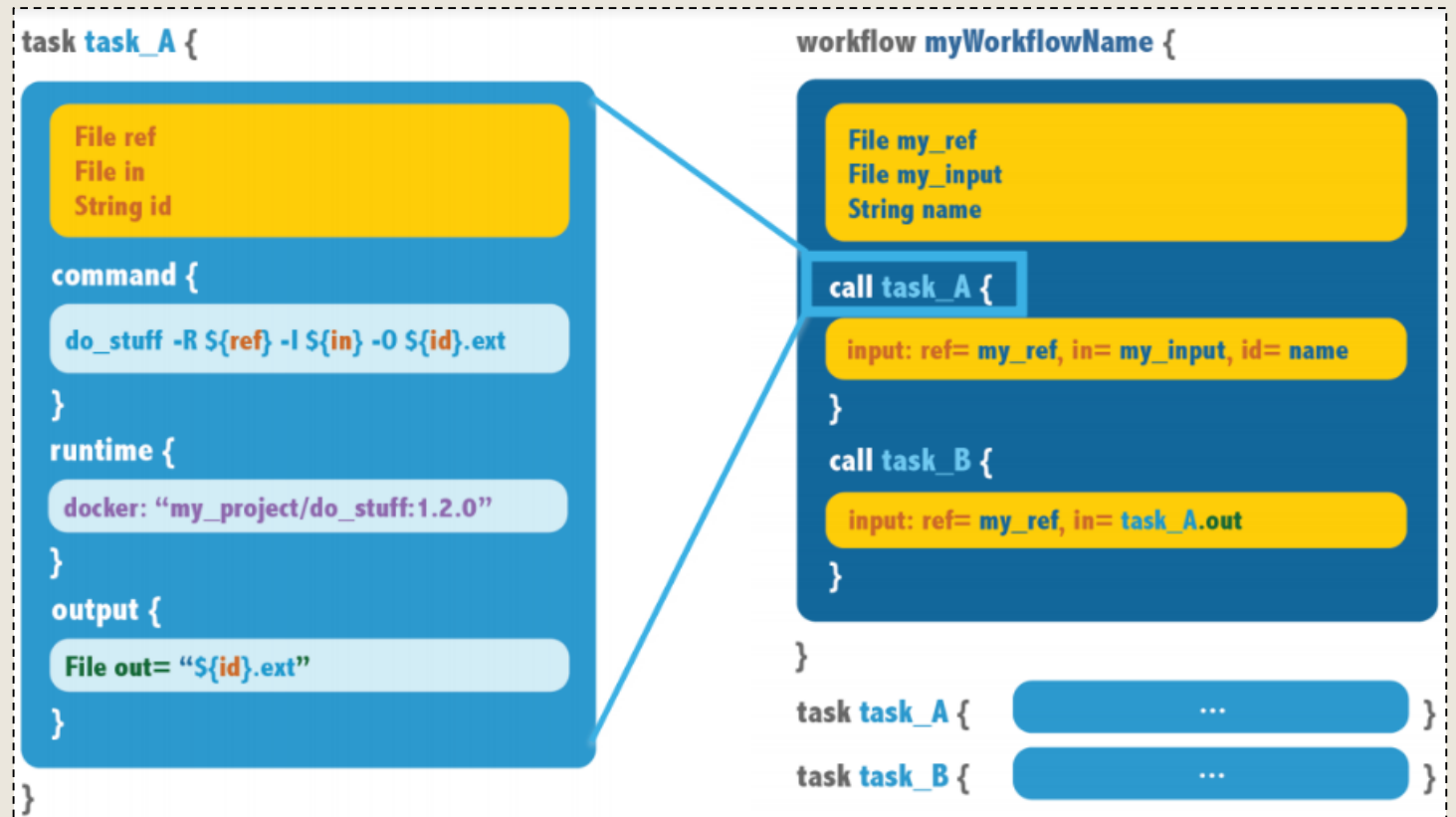
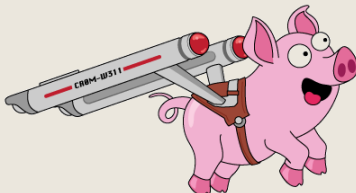
Datos y Metadatos: Estructuración de los datos



Workflow Description Language



- Un lenguaje simple para describir ‘workflows’.
- Reúne datos de entrada/salida, herramientas y comandos
- Interpretado y ejecutado por **Cromwell**.

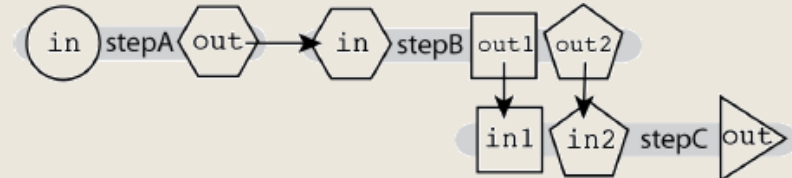


Organización de Tareas en WDL

- Existen tres maneras de organizar nuestras tareas.
 - Cadena lineal o con input/output múltiples.

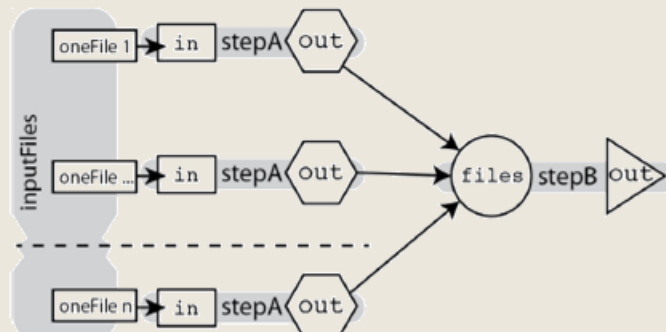


```
call stepA
call stepB { input: in=stepA.out }
call stepC { input: in=stepB.out }
```



```
call stepC { input :
    in1=stepB.out1,
    in2=stepB.out2 }
```

- Scatter-gather (Dispersar-Reunir)



```
Array[File] inputFiles

scatter(oneFile in inputFiles) {
    call stepA { input: in=oneFile }
}

call stepB { input: files=stepA.out }
```

Referirse a un 'container' en WDL

- Simple! Una vez el 'container' Docker listo, lo ponemos en línea a través del **Google Container Repository** (acceso privado/público) o **Docker Hub** (público).
- En el WDL, nos referimos al 'container' Docker en el cuerpo de **runtime** con una línea.
 - docker:"broadinstitute/gatk:4.1.2.0" ➔

```
runtime {  
  docker: "my_project/do_stuff:1.2.0"  
}
```