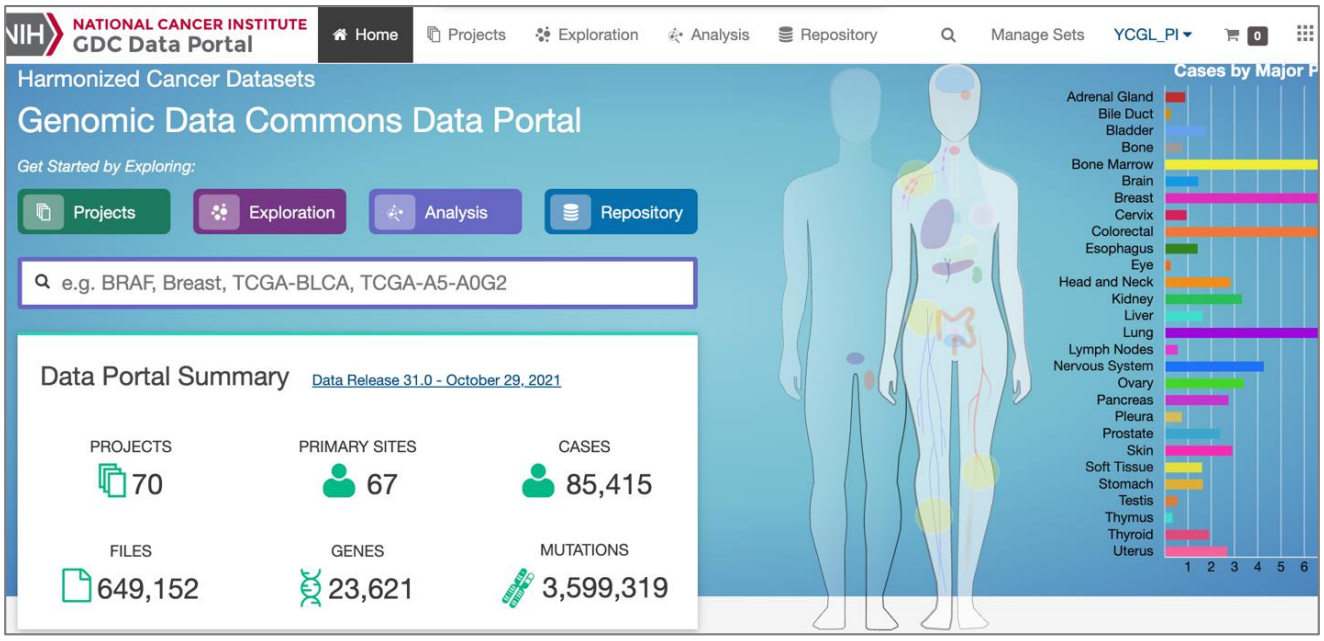
Cancer Genomics Cloud 使用經驗分享

李元綺

臺北醫學大學/醫學資訊研究所

自生物資訊領域開始發展以來，數據量就一直是個重要的課題，隨著電腦技術躍進，舉凡 DNA 定序、基因表現微陣列 (Gene expression microarrays) 等數據議題有更多人投入研究並逐一克服。伴隨著電腦與科技的發展，生物科技也從沒停止演進過，到了次世代定序 (Next generation sequencing) 的發明，大型合作研究、追蹤型研究、世代研究 (cohort studies)、縱貫型研究等，已經不僅是單純電腦容量的問題，因為還包括複雜的分析工具、視覺化呈現、重複性流程等工作，電腦的處理器 (CPU)、讀取記憶體 (RAM) 也都得跟著升級。而一般的實驗室其實很難具有這些設備及達成這些工具上的要求，除了少部分專業且經費充裕的核心實驗室，他們能夠擁有自身大型的伺服器、與經驗豐富的生物資訊分析專家可以協助。為了讓這些問題妥善地解決，所以癌症基因組雲 (The Cancer Genomics Cloud, CGC)便應運而生。

起初是因為美國國家癌症研究院(NCI, National Cancer Institute)發展的大型 NGS (next generation sequencing) 數據：癌症基因組圖譜 (TCGA, The Cancer Genome Atlas)。TCGA 是一個大約 2.5 PB、包含 WXS (可轉譯區塊定序)、RNA-seq (RNA 定序)、WGS (全基因定序) 的大型數據庫。為此，他們發展了 NCI GDC (Genomic Data Commons) Data Portal (圖一)，單純地處理 TCGA 的搜尋與下載。但是進一步的分析處理，僅有前面所提及的核心實驗室能夠進行，對於沒有足夠計算資源或適當技術知識的研究人員來說，將會是個很大的挑戰。因此 NCI 資助了三個癌症基因組雲 CGC，作為試點項目計畫，包括 Broad Institute、Institute for Systems Biology 以及 Seven Bridges。表一為此三大雲端系統使用的平台及實驗室主持人資訊介紹等，旨在探索新方法，使得研究人員能夠更有效率地運用這些數據集。



【圖一：GDC data portal 的主要頁面，網址為 https://portal.gdc.cancer.gov/】

【表一：分析 TCGA 的三大雲端系統比較與介紹】

	Broad Institute	Institute of Systems Biology	Seven Bridges Genomics
PI	Gad Getz	Ilya Shmulevich	Deniz Kural
Collaborators	UC Berkeley, UC Santa Cruz	Google, SRA International	None
Cloud Platform	**Google**	**Google**	**Amazon Web services**
Unique Tech used	ADAM/Spark	Google Genomics Platforms	SBG platform
Tools incorporated	Firehose	Regulome explorer, Gene Spot (focus on interactive data visualisatoin , exploration, and analysis	>30 public pipelines https://igorsbgenomics.com/lab/public/pipelines/
Cloud Pilot website	http://firecloud.org/	http://cgc.systemsbiology.net	https://www.sbgenomics.com/cancer-genomics-cloud/

關於上述三個雲平台的使用，只要您是對生物數據集研究有興趣的人，都可以藉由註冊後使用其資源。以下內容將以自身實驗室最常使用的雲平台進行介紹： Seven Bridges Genomics CGC (SBG-CGC)。SBG-CGC 於 2016 年 2 月開始開放公眾使用，可以創建個人帳號，或經由 eRA Commons 的帳戶登錄 (如欲申請 eRA Commons 的帳號，可以向北醫大研發處聯繫，此個人帳號必須依附於一個機構中。例如我們是北醫大的研究職員，我們的個人帳號就必須依附於北醫大機構之下)。

SBG-CGC 會提供一筆試用費用，約 100 美金，因為 SBG-CGC 為商業平台，一開始雖然是透過 NCI 資助建立，但之後的營運交由平台自行管理、自負盈虧。

關於 SBG-CGC 的幾個特色介紹如下(1)：

一、可與他人共同協作，如遇執行面及操作上的問題，亦可以主動加入 SBG 的內部人員一起共同研究解決問題。

二、經由一個直觀的界面，可快速造訪並使用大型公共基因組數據集。在安全的資管控制之下，平台裡有許多內建並可使用的生物資訊工具和工作流 (workflow)。當需要更多的計算資源時，可隨時地進行擴充。其中也包括軟體開發工具包：一種應用程序編程介面 (API)，可自動化、視覺化、及查詢工具，對協作、重複性研究都可以完善支援。

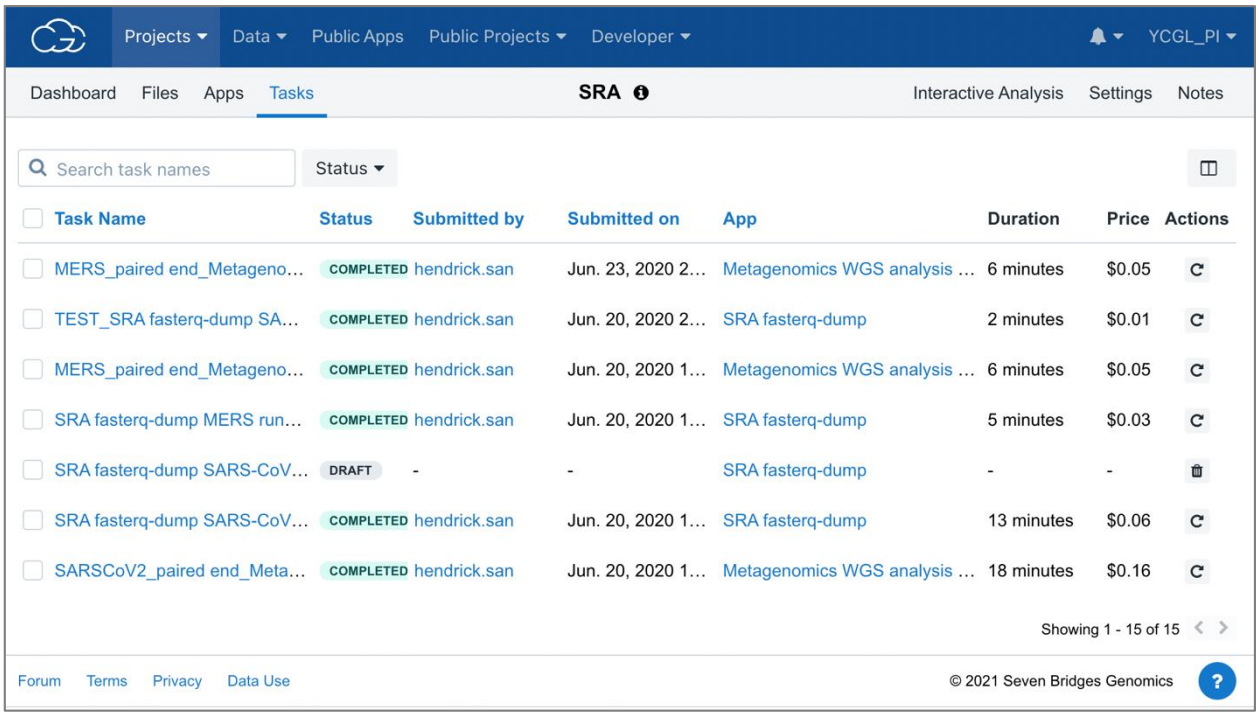
三、SBG-CGC 建置於 Amazon Web Services 的企業級雲服務上，為研究人員提供了公共基因組數據集，包括 TCGA 和 CCLE (the Cancer Cell Line Encyclopedia)。如果遇到一些公共基因組數據尚未建立在 CGC 裡面，也可以請求內部人員幫忙建立快速撈取程式的服務。以我們的例子來說，我們對於 dbGaP 裡面的資料集 (datasets) 有興趣，但是不想花費額外的時間下載再上傳至 CGC，便可以透過此項服務，讓內部人員幫忙撰寫相關程式，提供給使用人員勾選需要的 datasets，然後迅速地傳送至 CGC 裡使用。

四、配合 datasets，CGC 裡的數據分析工具和工作流有將近 200 多個，工具包括分析全基因組（Whole genome）和外顯子組 (exome) 的序列 variant calling、RNA 測序數據的差異表達分析 (RNA sequencing) 和複合數據視覺化等。更重要的是，這些工具及工作流將會根據用戶的需求，不斷地更新及修訂，用以確保使用者能夠更便利的操作平台上的所有功能。

五、CGC 上的工具打包在 Docker container 容器中，這是一種輕量級的軟體虛擬化技術（www.docker.com）。執行指令是使用通用工作流語言 (Common Workflow Language, CWL; www.commonwl.org) 進行描述，這是一項開放的資源、社區開發的規範，用於跨軟體和硬體環境可攜帶、移植、可擴充的方式來描述分析工作流和工具。

六、CGC 還為研究人員提供了一個強大的軟體開發工具包 (software development kit)，使用者能夠在 CWL 中輕鬆描述自己的工具和自行定義腳本，以便在平台上使用。可視覺化工作流編輯器 (visual workflow editor)，允許用戶從各個工具直觀地建構可重現的工作流。

七、數據分析必須是可擴展的，以便使用者能充分利用可用的數據集。CGC 基礎設施的彈性意味著隨著數據分析的擴大，會分配額外的計算資源以實現批次作業的並行化和處理。至於費用方面，根據 SBG-CGC 提出的說法，他們在過去的 10 年裡，雲計算成本降低了 80% 以上，使其成為分析大型基因組數據集中最具成本效益的方法。例如，一位研究人員能夠在大約 3 小時內對 11,000 名 TCGA 參與者進行有針對性的變異位點偵測 (target variant calling)，所需費用將不到 15 美元。而以我們自身做 RNA-seq, 20 cases v.s. 20 control 樣本為例，也是能以 10 美元以內的價格並且在數小時內完成。至於如何選擇合適的計算資源仍是需要多方嘗試及累積經驗，圖二為我們在此平台上所使用的工作項目及相對應的時間與花費。



【圖二：在 Seven Bridges Genomics-Cancer Genomics Cloud 平台上使用的工作項目，以及相對應的工具、操作時間與花費費用。】

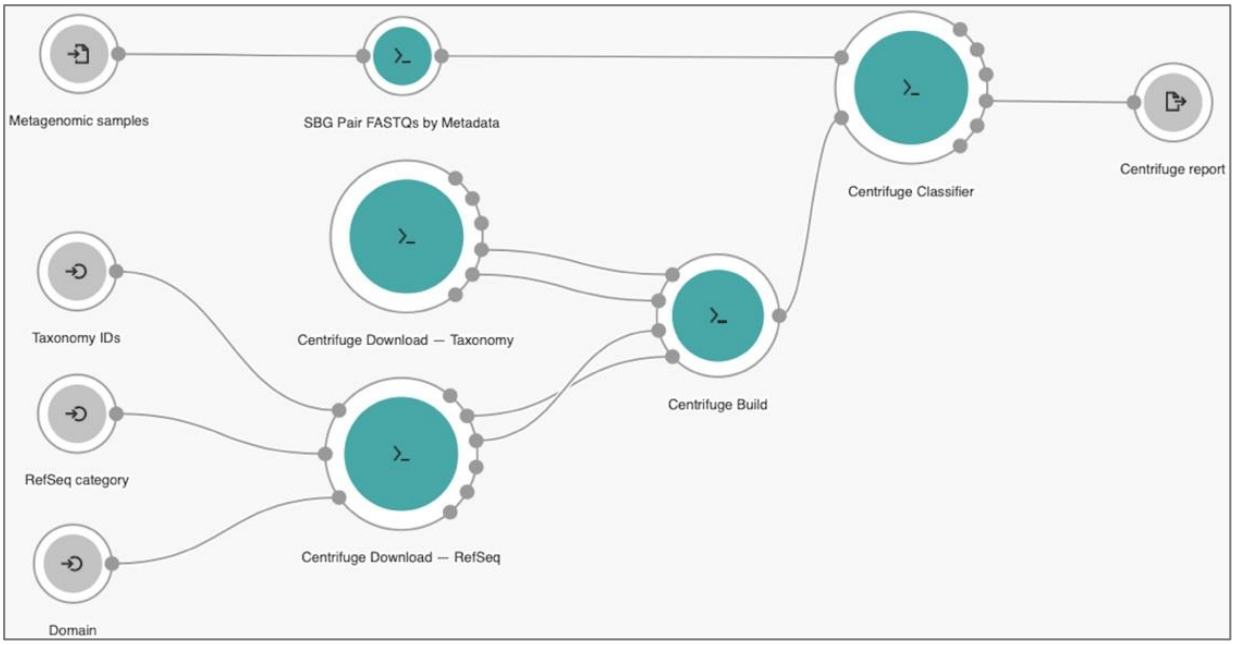
一般來說，雲服務可以分為四種類型：數據即服務 (DaaS, data as a service)、軟體即服務 (SaaS, software as a service)、平台即服務 (PaaS, platform as a service) 和基礎設施即服務

(IaaS, infrastructure as a service)。而此生物領域的雲服務的應用的確完全符合：公共生物數據庫進行 DaaS、工具或工作流用於 SaaS、分析和編程環境用於 PaaS，以及虛擬化資源，如用於 IaaS 的虛擬中央處理器 (vCPU)。

我們透過此平台建置了許多專案計畫，其中一個有關 Covid-19 序列分析的專案計畫已於 2021 年 10 月份發表(2)。以下內容我們介紹如何利用 SBG-CGC 雲平台，改良優化一個原有的演算法，成為一套合乎我們病毒鑑識使用的工作流，憑藉著索引序列 (Index) 有效的將數百條 SARS-CoV2 (即引發 COVID-19 的病毒株) 與 H1N1 病毒株進行區隔。在時間足夠的前提下，鑑定病毒序列是否為 SARS-CoV2 不是一件困難的事情，但是在大流行期間，短時間內湧入上百、或是上千條序列，而定序本身已經要花一些時間，再加上定序完以後的比對 (mapping)、組裝 (assembly) 皆需耗時。傳統的 RT-PCR 技術，更是需耗費半天以上的時間，而且以 PCR 的兩端引子 (probes) 夾出的序列更會因為病毒株的突變而出現偽陰性的判讀。因此次世代定序是一個較快速的方式，但我們並不像以往需要全基因組的比對、組裝，而是根據參考基因 (reference genome) 採用索引 (index) 序列的部分，因此能夠在一完成部分片段定序 (如 short-reads) 時就能即時判定是否是 SARS-CoV2。

Centrifuge 演算法原本是一個內建在平台上的工具，此演算法原本是用來區分腸胃道中微生物菌叢種類的工具，Centrifuge 是根據 Burrows-Wheeler transformation(3) 以及 Ferragina-Manzini index 索引(4) 的索引演算法，自參考基因組 (Reference genome sequences) 對應於微生物總體基因分類。但跳脫原本的分類 (Classification) 目的，而改寫、並為了更穩健的識別 (Identification) 目的重新修改設計出符合快速、準確的鑑別、與區分病毒株的工具與工作流。我們整套工作流是以 Rabix 開發, 在通用工作語言 (common workflow language, CWL)下，能移植、擴充、重複性、再現性，並跨不同硬體、軟體平台環境使用，包括工作站 (workstations)、運算叢集 (clusters)、雲以及高效能電腦。

Rabix 是一個用於編碼、測試和除錯 CWL 應用程序的整合式軟體環境。CWL 應用程序可以被描述為工具或工作流。該工具可以獨立執行，也可以內建到工作流中，工作流是一個或多個連接工具的集合。一個工作流也可以是其他工作流的集合。在 Rabix 中，CWL 應用程序變成節點 (node) 和邊 (edge)，以指示連接工具之間的數據元素或變量的流動。節點 (node) 作為可以並行 (parallel) 執行的單個或一組命令工作，可以是輸入、工具或輸出。同時，邊 (edge) 代表從上游節點傳遞到下游節點的數據元素或變量、文件或參數。CWL 工作流代碼可以描述為可導出為 JavaScript Object Notation (JSON) 方案的步驟。我們使用 Rabix 在 CGC 平台上所執行的工作流的如圖三所示。其中圖示、及 分別代表輸入 (input)、軟體工具、和輸出 (output) 節點。



【圖三：顯示我們所使用的病毒鑑定雲工作流。
其中、及 分別代表輸入 (input)、軟體工具和輸出 (output)。】

至於詳細的分析細節就不在此贅述，有興趣的人可以透過次頁附註的論文瞭解內容詳情，本文僅藉此機會向大家分享一些在 CGC 平台上的使用經驗，讓大家認識此一便利的平台資源。

參考文獻：

1. Lau et al., The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Res. 2017. 77(21):e3-e6.
2. Lim et al., Orchestrating an Optimized Next-Generation Sequencing-Based Cloud Workflow for Robust Viral Identification during Pandemics. Biology (Basel), 2021. 10(10):1023.
3. Burrows, et al., In Technical Report 124; Digital Equipment Corporation: Palo Alto, CA, USA, 1994.
4. Ferragina et al., In Proceedings of the 41st Annual Symposium on Foundations of Computer Science, Redondo Beach, CA, USA, 12–14 November 2000; p. 390.