Accessible, interactive and cloud-enabled genomic workflows integrated with the NCI Genomic Data Commons

Ling-Hong Hung
School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
Biodepot LLC, Seattle, WA
Email: lhhung@uw.edu

Bryce Fukuda
School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
Email: fukudab@uw.edu

Robert Schmitz
Biodepot LLC, Seattle, WA
Email: schmitzr1984@gmail.com

Varik Hoang
School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
Email: varikmp@outlook.com

Wes Lloyd
School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
Email: wlloyd@uw.edu

Ka Yee Yeung*
School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA
Biodepot LLC, Seattle, WA
Email: kayee@uw.edu

*Corresponding author

ABSTRACT

Large scale data resources such as the NCI's Cancer Research Data Commons (CRDC) and the Genotype-Tissue Expression (GTEx) portal have the potential to simplify the analysis of cancer data by providing data that can be used as standards or controls. However, comparisons with data that is processed using different methodologies or even different versions of software, parameters and supporting datasets can lead to artefactual results. Reproducing the exact workflows from text-based standard operating procedures (SOPs) is problematic as the documentation can be incomplete or out of date, especially for complex workflows involving many executables and scripts. We extend the Biodepot-workflow-builder (Bwb) platform to distribute the computational methodology with integrated data access to the National Cancer Institute (NCI) Genomic Data Commons (GDC). We have converted the GDC DNA sequencing (DNA-Seq), the GDC mRNA-Seq SOPs into reproducible, self-installing, containerized graphical workflows that users can apply to their custom datasets. Secure access to CRDC data is provided using the Data Commons Framework Services (DCFS) Gen3 protocol. The user can perform the analysis on their laptop, desktop or use their preferred cloud provider to access the computational and network resources available on the cloud. We demonstrate the impact of non-uniform analysis of control and treatment data for the inference of differentially expressed genes. Most importantly, we also provide a dynamic and practical solution for uniform and reproducible reprocessing of omics data allowing cancer researchers to take full advantage across multiple data resources such as the CRDC and GTEx.

KEYWORDS

Cancer Research Data Commons, cancer genomic workflows, cloud computing, FAIR, software, data integration

BACKGROUND

Massive amounts of high-throughput datasets have been generated to enhance the understanding and treatment of cancer. Large-scale programs such as The Cancer Genome Atlas (TCGA) [1], Therapeutically Applicable Research to Generate Effective Treatment (TARGET) [2], and Clinical Proteomic Tumor Atlas Consortium (CPTAC) [3] have generated multi-omics data resources across many diverse types of cancer. The National Cancer Institute (NCI) launched the Cancer Research Data Commons (CRDC) to connect these diverse datasets with analytical tools [4]. The CRDC provides access to different data-specific repositories, including the Genomic Data Commons (GDC) that stores raw sequencing data and derived results [5-7].

As of the v33.0 data release on May 3, 2022, the GDC consists of data from over 85 thousand cases across 72 projects [8]. Experimental strategies in the GDC include RNA sequencing (RNA-Seq), microRNA sequencing (miRNA-Seq), whole genome sequencing (WGS), whole exome sequencing (WXS), and targeted sequencing.

The CRDC provides different levels of data that reflect the amount of processing. The raw un-processed sequence data (level 1) derived directly from patients contain potentially identifying information and these data are typically controlled access. [9]. For the GDC, controlled access data requires dbGaP (database of Genotypes and Phenotypes) authorization and eRA Commons authentication. For data access, the GDC provides several methodologies that support authentication. The simplest is the web-portal that allows users to browse and query the database using a graphical interface. However, the web-portal is not really designed for downloading large datasets, such as raw sequence files which are typically in the range of 10-20 GB for GDC data [9]. For data download, the GDC has the standalone Data Transfer Tool which has an optional user interface (UI). However, this is an older tool that does not support the Gen3 authentication protocol which is the new standard for the CRDC databases. There is a newer Gen3 client that lacks the UI and is not documented on the GDC site. Finally, there is also a web API (Application Programming Interface) for programmatic access that can be used to download large datasets but requires more expertise to use.

In contrast to raw sequence data, processed data such as transcripts counts do not reveal the identity of patients and are often available with fewer restrictions. The datasets are smaller in size, and easier to use. However, methodologies for processing sequence data are heterogeneous and involve a wide variety of different software tools, parameters and supporting data that are constantly being updated. For example, the GDC has published detailed documentation of analytic pipelines developed to process raw data [10, 11] including its workflow for analyzing RNA-Seq data using the STAR aligner [12]. While STAR is a popular aligner, HISAT [13], Bowtie [14], Kallisto [15] and Salmon [16] are examples of other aligners and pseudo-aligners that are frequently used for RNA-seq analyses. The GDC uses the Genome Reference Consortium Human Build 38 (GRCh38) with additional viral decoy sequences to increase the accuracy of alignments. While GRCh38 is in widespread use, differences in the masking and decoys give rise to slightly different reference sequences. The GDC uses GENCODE [17] annotations to map alignment coordinates to transcripts. GENCODE is not universally used with RefSeq [18] being a manually curated annotation alternative. Different data releases from GDC can use different

pipelines. Early GDC data releases used the STAR aligner [12] that required two passes. Later releases used an improved version of STAR that aligned in one pass. Later data releases also use STAR to generate the transcript counts whereas earlier versions used HTSeq [19] to quantify counts. Earlier data releases use GENCODE v22 for annotations whereas the current release uses GENCODE v36.

Processing data using different pipelines can affect results. Arora et al. compared well-established RNA-Seq processing pipelines using 6690 human tumor and normal samples from the TCGA and GTEx projects and reported major discrepancies in the abundance estimates that include disease-associated genes [20]. Arora et al. called for a community wide effort to develop gold standards to estimate mRNA abundances that could be used to harmonize data from different projects [20]. However, results that we present in this paper show that even minor changes in the software, versions, parameters and supporting datasets can affect the identification of differentially expressed genes. Nor are the effects randomly distributed which would argue against the effectiveness of a data standards approach for harmonization. Furthermore, these changes are not capricious but reflect ongoing improvements in methodology and data. For example, we fully expect that new references and annotations derived the new telomere to telomere human sequence will eventually be incorporated into GDC pipelines. The solution that we propose is not to attempt to provide a static set of harmonized processed data but to facilitate and distribute the workflows so that they can be reproducibly applied to raw datasets and customized as methods and versions and supporting data are updated. To accommodate the size of the raw datasets and the computational demands of the processing, the tool must be cloud enabled to minimize data transfers, and take advantage of the throughput and scalable computational abilities available on the cloud. Finally, the workflows should be as accessible as possible to allow researchers to re-process their own datasets and expand the usage and utility across multiple data resources such as the CRDC.

Our Contributions

In this paper, we have converted the GDC descriptions of their data process pipelines to graphical workflows for the Biodepot-workflow-builder (Bwb) [21] platform. In particular, we added GDC DNA sequencing (DNA-Seq) [11], the GDC mRNA-Seq workflows [10], as well as the Data Commons Framework Services (DCFS) Gen3 authentication [22] to the Bwb to provide integrated access to the CRDC. A workflow in the Bwb consists of a sequence of graphical widgets, each of which represents a modular and containerized task. Bwb is an *interactive graphical* platform

for biomedical scientists with features including automated installation, the form-based entry of parameters, and the ability to add new modules via drag-and-drop, and support for graphical output for applications such as Jupyter notebooks. Bwb is easily customizable through and also includes widgets to support the execution of Common Workflow Language (CWL) [23] and Workflow Description Language (WDL) [24] workflows. Bwb can execute locally or on a cloud provider such as AWS, Google Cloud, Azure and IBM cloud without any need for account registration. We demonstrate the utility of our dynamic platform harmonizing RNA-Seq data from TCGA and the Genotype-Tissue Expression (GTEx) [25] projects. Most importantly, we empirically studied the impact of uniform re-processing of data versus direct use of processed data on the inference of differentially expressed genes. Instead of developing a static gold standard data processing pipelines for genomics data, we believe that our platform for reproducibly distributing the computational protocols will enhance will increase the flexibility, and ease of integration across multiple data sources.

Related Work

In addition to enabling data sharing, the GDC provides software tools from the web-based data portal that support data analysis, visualization, and exploration (DAVE) [26]. The NCI also supports the development of cloud-based platforms to analyze data hosted by the CRDC, including the Broad Institute's Terra (formerly known as Firecloud) [27], the Institute for Systems Biology's Cancer Gateway in the Cloud (ISB-CGC) [28, 29], and Seven Bridges Cancer Genomics Cloud [30, 31]. Both Terra and the ISB-CGC leverage Google Cloud to support cancer genomic analysis. Terra provides integrated access to the CRDC while providing pre-configured workspaces that support common use cases, such as the GATK best practices workflow. The ISB-CGC supports interactive web-based applications, Google Cloud APIs, and custom scripts and APIs for CRDC data access [28]. Seven Bridges is a commercial service using Amazon Web Services (AWS) or Google Cloud for bioinformatics analyses [30]. Users can drag multiple "apps" and parameters onto a canvas and connect them to define an executable workflow.

Most existing workflow execution platforms such as NextFlow [32] were designed around traditional batch workflows and scripting methods where a user interface such as DolphinNext [33] was appended afterwards. Other CWL execution engines, such as Seven Bridges with integrated access to the CRDC, leverage on power tools for CWL, such as Rabix that supports editing of CWL scripts and visualization of CWL workflows. Galaxy is a web-server that provides a common web interface for users to create and execute workflows in a consistent hardware and

software environment on a server or cluster [34]. While most Galaxy workflows are not containerized, Galaxy can use Bio-Docklets [35] to execute Docker workflows. However, importing tools and containers from non-Galaxy sources is not trivial and requires modifying a set of configuration files and scripts whereas Bwb provides specific GUI tools for customizing existing workflows. In contrast, Bwb provides a form-based user interface to facilitate the import of R, Python scripts, and Docker containers. Additionally, Bwb supports graphical output as well as interaction with workflow modules that have a GUI. Bwb can also export workflows as bash scripts of Docker commands that can be run without Bwb and easily incorporated into other scripts.

IMPLEMENTATION

Each step in our workflows in the Bwb is encapsulated in a Docker container, with specific version tags to ensure software dependencies and compatibility. Users can ***interactively*** start, stop, and modify these workflows through a drag-and-drop, point-and-click user interface in Bwb. Parameters in each module can be changed via a form-based user interface in Bwb. Bwb supports modules with their own graphical output and interfaces, including gnumeric spreadsheets [36], the Integrated Genome Viewer (IGV) [37], and Jupyter notebooks [38]. We tested our workflows with controlled-access data from The Cancer Genome Atlas (TCGA) [1] and open-access data from the Cancer Cell Line Encyclopedia (CCLE) [39, 40] projects from the GDC. Our widgets, workflows, and documentation are available from our GitHub repository (https://github.com/Biodepot-LLC/workflows).

Integration with the Cancer Research Data Commons

The NCI Data Commons Framework Services (DCFS), powered by Gen3, is a set of software services that facilitate the hosting, management, and sharing of cancer datasets in the cloud [41, 42]. Among these software services, authentication and authorization services are available so that controlled access data can be shared in the CRDC cloud infrastructure. In particular, the Gen3-client provides a command-line interface to upload and download files to and from a Gen3 data commons [22]. In this work, we encapsulated the binary executable of this Gen3-client in a Docker container and provided a form-based user interface by creating a graphical widget in the Bwb for downloading files.

Before researchers could obtain access to controlled data in the GDC, PIs must first apply for access to specific projects or datasets through the NIH dbGaP, and then grant access to

individuals in their labs. Therefore, the Gen3 client only supports downloading files belonging to projects that the user has access to in the dbGaP. Files, such as the Panel of Normals (PON) used in the DNA-Seq pipeline, that require controlled access but belong to no specific project, cannot be downloaded with the Gen3 client. These files, however, can be downloaded using the GDC API. Consequently, we have consolidated these two methods into a Gen3 widget. **Figure 1** shows a screenshot of Bwb's support for downloading controlled access files using the Gen3 client and the GDC API**.** The user authenticates for Gen3 by signing into the dbGaP via the NIH eRA Commons to obtain the credentials file. Users can also provide an access token for the GDC API. This widget will attempt to use Gen3 to download the file (or manifest of files) and if that fails the widget will attempt to use the GDC API. This ensures that all controlled access files can be downloaded. Parallel downloads are supported using Bwb's internal list-based scheduler if the user enters multiple files or manifests. Using this new widget, downloads of controlled access files can now be incorporated into Bwb workflows, such as the Panel of Normals file needed for the GDC DNA-seq pipeline, and sequencing read data from the GDC. A demonstration video for the Gen3 widget in the Bwb is available at https://youtu.be/8upzPouRGys.

mRNA-Seq workflow from the Genomic Data Commons

The GDC mRNA-Seq workflow [10] aligns raw sequence files to the GRCh38.d1.vd1 reference sequence using the STAR (Spliced Transcripts Alignment to a Reference) aligner [12], followed by the quantification step that outputs raw read counts and normalized read counts. In GDC Data Release version 15 to 31, STAR [12] version 2.6.0c was used to compute the index and alignment, counts were obtained using HTSeq [19] using GENCODE [17] v22 as the reference annotation. Starting in GDC Data Release version 32, STAR version 2.7.5c is now used with an additional input parameter, reference annotation is updated from GENCODE v22 to v36, and HTSeq is no longer used in the mRNA-Seq workflow. **Figure 2** shows a screenshot of our implementation in the Bwb consisting of the following steps: download the reference and sample data; create a genome index using the reference sequence; align reads to the reference, quantify the number of reads mapped to each gene, and calculate normalized gene expression values. A demonstration video of the GDC mRNA-seq workflow is available at https://youtu.be/YzFa9Een7Tc.

DNA-Seq workflow from the Genomic Data Commons

The GDC DNA-Seq workflows consist of six main procedures that identify somatic variants from WXS and WGS data by comparing allele frequencies in normal versus tumor samples [11]. Our

implementation in the Bwb, shown in **Figure 3,** consists of the following main steps: 1) download the reference and sample data; 2) convert BAM input files to fastq format; 3) align read groups to the reference genome using *bwa mem*; 4) perform variant calling using multiple callers; 5) annotate raw somatic mutations based on biological context and known variants from external mutation databases; 6) convert VCF to MAF files; and 7) display results in the Integrated Genome Viewer (IGV). A novel contribution from our team is that we have added functionality to create a batch file to load multiple variant files and regions of interest. A demonstration video of the GDC DNA-seq workflow is available at https://youtu.be/M7MCI83Q7_A.

Support of CWL and WDL workflows

For bioinformatics, there are two commonly used languages to describe containerized workflows, the Workflow Description Language (WDL) [24, 43] used primarily by the Broad Institute, and the Common Workflow Language (CWL) [23, 44]. The Toil tool developed by Vivian et al. [45] supports execution of workflows described using WDL or CWL in addition to supporting job schedulers. We have implemented four new widgets, namely Toil/CWL, Toil/WDL, Cromwell/WDL and Cromwell/CWL widgets, to allow users to easily import existing CWL/WDL workflows as reproducible modules into the Bwb. These widgets allow users of CWL/WDL workflows to take advantage of the Bwb's features such as the integration with CRDC data through Gen3 or interactive Juptyer notebooks.

A simple example to showcase these benefits is the Sanger Institute WGS variant calling workflow from the GDC DNA-seq workflows page. We have used our Toil workflow to support the CWL description of this workflow shown in **Figure 5** where the workflow is connected to a widget that automatically downloads the necessary files. A demonstration video is available at https://youtu.be/XRdhJRvVecM.

RESULTS

In this section, we demonstrate the utility of our dynamic platform harmonizing RNA-Seq data from cancer samples in TCGA and non-diseased tissue-specific samples in the Genotype-Tissue Expression (GTEx) [25] project. Specifically, we empirically studied the impact of different pipeline variations on the estimation of transcript abundance and inference of differentially expressed genes.

Data

For TCGA RNA-Seq data, we downloaded fastq files from the GDC Legacy Archive [46], BAM files and counts data generated by the STAR aligner (Data Release version 32) from the GDC Data Portal [47]. Since the HTSeq files were removed from the GDC Data Portal starting in Data Release 32, we used patient case IDs to identify HTSeq count files' case ID by cross-referencing to the manifest available from the GDC documentation GitHub [48]. For the GTEx RNA-Seq data, we downloaded the v8 processed counts data from the GTEx Portal [49] and the controlled access BAM files from the AnVIL repository [50, 51].

Comparison of different versions of GDC data release

Since the RNA-Seq workflow had been changed substantially in Data Release version 32 compared to versions 15 to 31, our first step is to quantitatively compare the published raw counts from Data Release v15 HTSeq output counts file with the Data Release v32 STAR (v2.7.5c) output counts file. In particular, we downloaded published counts for sample TCGA-AB-2821 with case UUID: f6f9ed0d-2b3c-45b7-b214-853b5a207bac from the TCGA Acute Myeloid Leukemia (TCGA-LAML) project. After removing version numbers from the stable Ensembl gene IDs, there are a total of 56,485 common gene IDs in both count files. We observe that the published counts from GDC Data Release v15 are quite different from GDC Data Release v32, with 31,804 (56%) genes showing identical unnormalized counts. We computed the relative change ((v32 - v15)/v15) for each gene for which the v15 counts are non-zero. Among the 37,179 genes with non-zero counts, the median of the relative change is 0.0185, the 90th percentile of the relative change is 0.5, and 55 genes show a relative change above 10. Table 1 shows the 8 genes with relative change above 100. This comparison illustrates that minor updates in data processing workflows could lead to major changes in the output counts of some genes. Our results highlight the need for a ***dynamic solution*** to re-process the raw data as workflows are being updated to adopt the latest version of aligners and annotation references.

Ensembl gene ID	gene symbol	GDC v15 counts	GDC v32 counts
ENSG00000157654	PALM2AKAP2	7	2536
ENSG00000197753	LHFPL5	2	659
ENSG00000245864	MEF2C-AS2	1	147
ENSG00000250891	LINC02208	1	102
ENSG00000253194	AL137009.1	1	635
ENSG00000260007	AC107871.1	1	496
ENSG00000265817	FSBP	1	157
ENSG00000279170	TSTD3	3	568

Table 1. Genes with relative change ((v32 - v15)/v15) over 100 when comparing the GDC Data Release version 15 to version 32. Note that the stable Ensembl gene IDs are shown without the version numbers.

Comparison of published vs. reprocessed counts from GTEx

The GTEx project is a public resource profiling tissue-specific gene expression in non-diseased individuals [25]. The GTEx version 8 RNA-Seq processing workflow [52] used STAR v2.5.3a to align reads to the human reference genome GRCh38/hg38, based on the GENCODE v26 annotation. Subsequently, read counts and normalized TPM values were produced with RNA-SeQC v1.1.9 [53]. Comparing to the GDC workflow, different versions of STAR and GENCODE annotations were used in addition to using RNA-SeQC to generate read counts. Before we integrate the GTEx data with TCGA data, we studied the impact of the GTEx RNA-Seq pipeline versus the GDC RNA-Seq pipeline on unnormalized counts. Towards this end, we downloaded data from three whole blood samples, namely GTEX-N7MS-0007-SM-2D7W1, GTEX-NFK9-0006-SM-3GACS, and GTEX-O5YT-0007-SM-32PK7. The publicly available v8 processed counts were downloaded from the GTEx data portal [49], while the BAM files were downloaded from AnVIL [50, 51]. We applied the GDC version 32 pipeline to process the downloaded BAM files from GTEx.

First, we computed the Pearson's correlation coefficient between the published and reprocessed counts for each of these three whole blood samples. The correlation coefficients are 0.957, 0.954, and 0.994 respectively for each of the N7MS, NFK9, and O5YT samples. We observed that the number of non-zero published counts are 8592, 7092, and 8333 respectively for each of these three samples out of a total of 35,117 genes. Next, we computed the relative change, defined as ((reprocessed - published)/published), for each gene with non-zero published count. The median of the relative change are 0.441, 0.452, and 0.455 respectively for each of these three blood samples. The 99th percentile of the relative change are 5.018, 5.000, and 5.307 respectively. As a control, we compared the three published samples to the three reprocessed samples by applying DESeq2 [54], and observed that 197 genes have adjusted p-values under 0.05. To summarize, while the correlation coefficients between the published and reprocessed counts are high, there are some genes with substantial changes when the raw data was reprocessed with the GDC workflow. This empirical study highlights the need to harmonize the RNA-Seq data using the identical workflow before integration.

Integration of cancer and normal RNA-seq data using our dynamic platform

As an example, we illustrate that the GDC Data Release v32 RNA-Seq workflow implemented in our dynamic platform can be used to integrate cancer data from the GDC and non-diseased data from the GTEx project. Specifically, we downloaded BAM files from three cases (TCGA-AB-2821, TCGA-AB-2828, TCGA-AB-2839) in the TCGA-LAML project. We harmonized the transcript abundance of these three cancer patients and the three whole blood non-diseased samples discussed in the previous sub-section using our implementation of the GDC version 32 workflow. Subsequently, we used DESeq2 [54] to infer differentially expressed genes. Out of the 27,310 Ensembl gene IDs, 7087 gene IDs show an adjusted p-value under 0.01. We then mapped the Ensembl gene IDs to gene symbols using biomaRT [55]. **Figure 6** shows the volcano plot of the DESeq2 output.

As a control, we concatenated the v8 published counts from GTEx and v32 published counts from GDC, applied DESeq2 and recorded the top 10 differentially expressed genes. **Table 2** compares the differentially expressed genes inferred from concatenation of published counts versus those inferred from harmonized uniform GDC re-processing. We observe that CDCR1 has the most significant (smallest) adjusted p-values in both scenarios with zero change in rankings. As another example, FCGR3B is the second most significant differentially expressed gene in the harmonized re-processed scenario, with the reduction of one rank in the concatenation of published counts scenario. We observe that 5 of the top 10 (CD68, GPS2, ARL6IP4, GABARAP, CHKB) differentially expressed genes in the concatenation of published counts scenario exhibit dramatic change in rankings (over 10,000). In other words, half of the top 10 differentially expressed genes could be bogus if we don't reprocess the raw RNA-Seq data using uniform pipelines. This example illustrates the importance of harmonized uniform processing of RNA-Seq data.

	Published vs Published			Reprocessed vs Reprocessed	
padj	**Gene**	**rank Δ**	**rank Δ**	**Gene**	**padj**
5.64E-70	CXCR1	0	0	CXCR1	5.26E-69
1.35E-57	CD68	26180	1	FCGR3B	1.91E-56
1.11E-56	FCGR3B	-1	4	KCNJ15	6.68E-45
1.36E-56	GPS2	20906	5	FAM157A	7.42E-44
1.36E-56	ARL6IP4	19617	13	TREML3P	4.54E-36
3.15E-51	RNASEK	379	15	R3HDM4	5.45E-36
4.64E-46	KCNJ15	-4	10	CCNJL	6.41E-36
1.77E-44	GABARAP	10206	16	LCN2	7.05E-30

1.85E-44	FAM157A	-5	35	YPEL3	2.65E-26
1.30E-43	CHKB	26113	22	CD177	7.51E-26

Table 2: Comparison of the top 10 differentially expressed genes inferred from concatenation of published counts versus those inferred from harmonized uniform GDC re-processing. The column "rank delta" corresponds to the change in rank. In particular, a zero rank delta means no change in rank, e.g. CXCR1.

CONCLUSIONS

Using RNA-Seq data as a case study, we demonstrate the need to reprocess raw sequencing data since published counts change over different data releases with updated versions of aligners and reference annotations. We also show the need to harmonize raw sequencing data generated by different projects by re-processing the data with the same RNA-Seq workflow. Our observations echo the findings by Arora et al. [56]. However, instead of calling for a concerted, community-wide gold-standard for data processing, we provide ***a dynamic solution*** that distributes the computational methodology in a reproducible, accessible, customizable, and cloud-enabled manner to facilitate the reprocessing of data. Our platform is containerized and ready to be deployed on any cloud platform or local host with Docker installed. We support cancer genomic workflows published by the NCI Genomic Data Commons (GDC). Our platform leverages the NCI DCFS Gen3 framework to enable integration of controlled access data from the Cancer Research Data Commons. New widgets and workflows can be added to the Bwb using the form-based user interface without writing additional GUI code. We demonstrate the utility of our dynamic platform harmonizing RNA-Seq data from TCGA and the Genotype-Tissue Expression (GTEx) [25] projects.

ABBREVIATIONS

AMI: Amazon Machine Image

API: application programming interface

AWS: Amazon Web Services

Bwb: Biodepot-workflow-builder

CPTAC: Clinical Proteomic Tumor Atlas Consortium

CRDC: Cancer Research Data Commons

CWL: Common Workflow Language

DCFS: Data Commons Framework Services
dbGaP: database of Genotypes and Phenotypes
DNA-Seq: DNA sequencing
DTT: Data Transfer Tool
EC2: Elastic Compute Cloud
GDC: Genomic Data Commons
IGV: Integrated Genome Viewer
miRNA-Seq: micro RNA sequencing
NCI: National Cancer Institute
NGS: Next-generation sequencing
PON: Panel of Normals
RNA-Seq: RNA sequencing
TARGET: Therapeutically Applicable Research to Generate Effective Treatment
TCGA: The Cancer Genome Atlas
WDL: Workflow Description Language
WGS: whole genome sequencing
WXS: whole exome sequencing

Competing interests

LHH and KYY have equity interest in Biodepot LLC, which receives compensation from NCI SBIR contract numbers 75N91020C00009 and 75N91021C00022. The terms of this arrangement have been reviewed and approved by the University of Washington in accordance with its policies governing outside work and financial conflicts of interest in research.

Funding

LHH, WL, and KYY are supported by NIH grant R01GM126019, NCI SBIR contracts 75N91020C00009 and 75N91021C00022. RS, BF and VH are supported by NCI SBIR contracts 75N91020C00009 and 75N91021C00022. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

The content is solely the responsibility of the authors, and does not necessarily represent the official views of the National Institutes of Health.

Authors' contributions

Conceptualization: L.H.H. and K.Y.Y. Software development and testing: L.H.H., R.S., B.F. and V.H. Code documentation: R.S., B.F., and V.H. Cloud resources: L.H.H., W.L., and R.S. Computational benchmarking: L.H.H., R.S., V.H., and B.F. Data analysis: B.F. Design of case studies: K.Y.Y. and L.H.H. Figure preparation: L.H.H. Writing - original draft preparation: K.Y.Y. Writing - reviewing and editing: All authors. Funding acquisition: L.H.H. and K.Y.Y. The authors read and approved the final manuscript.

Acknowledgements

We would like to acknowledge cloud credits from Amazon Web Services and IBM Cloud.

Figure 1 A screenshot of the panels from the Gen3 download widget.
The required and optional entries panels from the Gen3 download widgets are shown. The user enters the location of the Gen3 credentials file and the desired profile to be used from the file. These are obtained by signing into the dbGaP database via the NIH eRA Commons. The user also has the option of entering the token file for use with the GDC API for files that cannot be currently downloaded by the Gen3 client. Multiple GUIDs or manifests can be entered and parallel downloads are supported using Bwb's built-in parallelism or using Gen3's multithreaded downloading of manifests. We also include the option of decompressing the files on the fly as they are being downloaded.

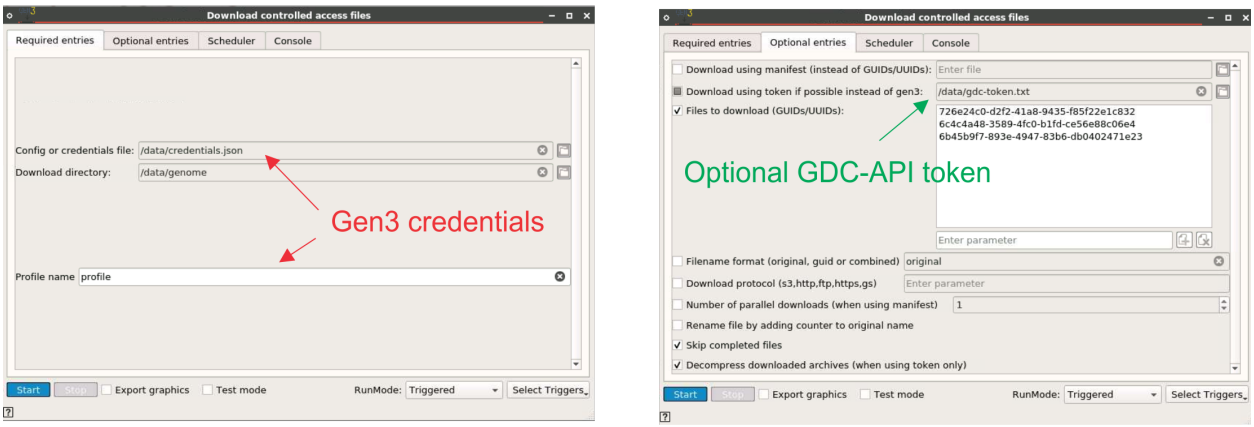
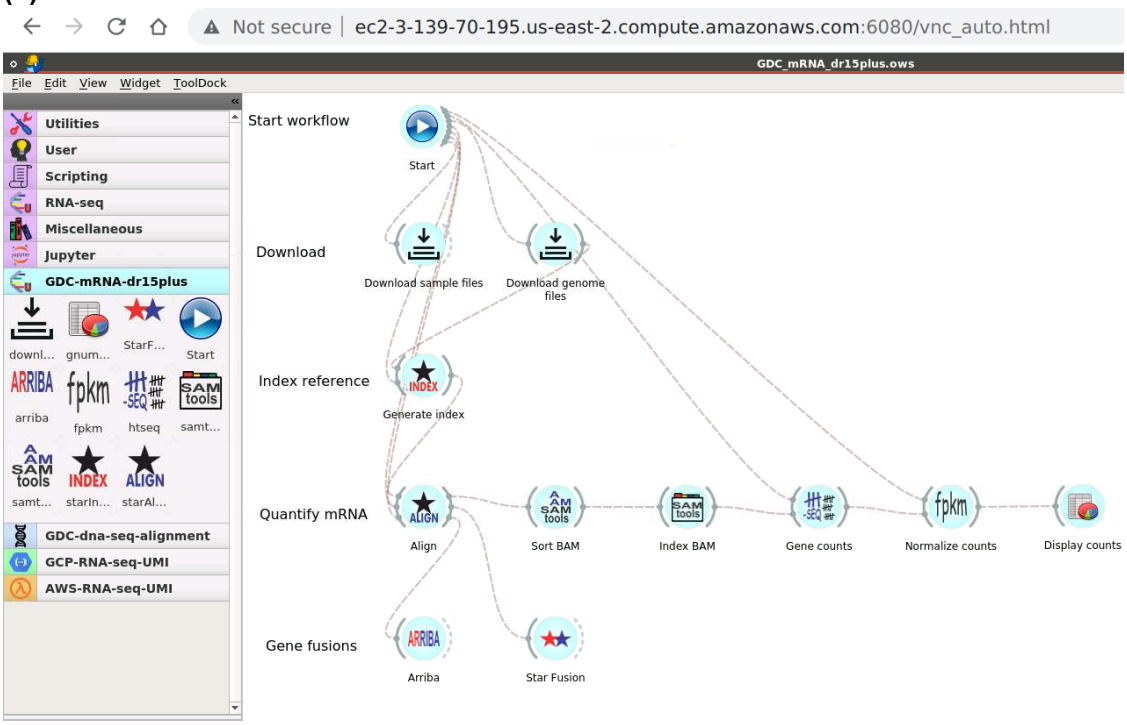


Figure 2 Screenshots of the GDC mRNA-Seq analysis workflows implemented in the Bwb. (a) GDC Data Release v15 mRNA-Seq workflow. (b) GDC Data Release v32 mRNA-Seq workflow. Each icon (widget) controls a separate containerized module. Double-clicking on a widget reveals graphical elements for parameter entry, starting and stopping execution, and displaying intermediate output. Lines connecting widgets indicate data flow between the execution modules. Connections and widgets can be added and removed using a drag-and-drop interface. The workflow itself is started by double-clicking on the start widget.

(a)



(b)

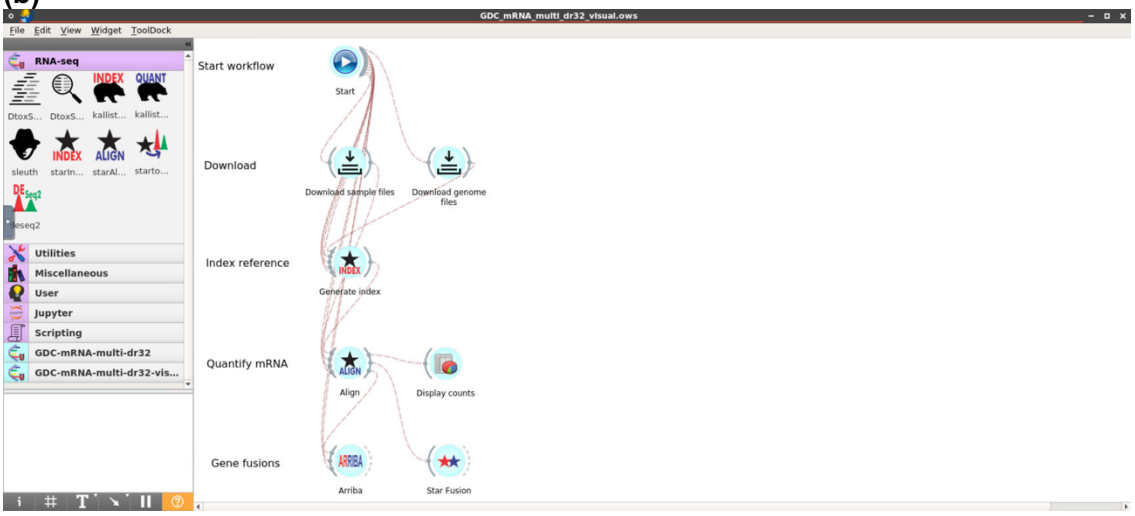


Figure 3 A screenshot of the GDC DNA-seq analysis workflow implemented in the Bwb. At the end of the workflow we have added an IGV widget which will automatically pop up with the regions of interest pre-loaded to allow the user to quickly evaluate the final MAF file.

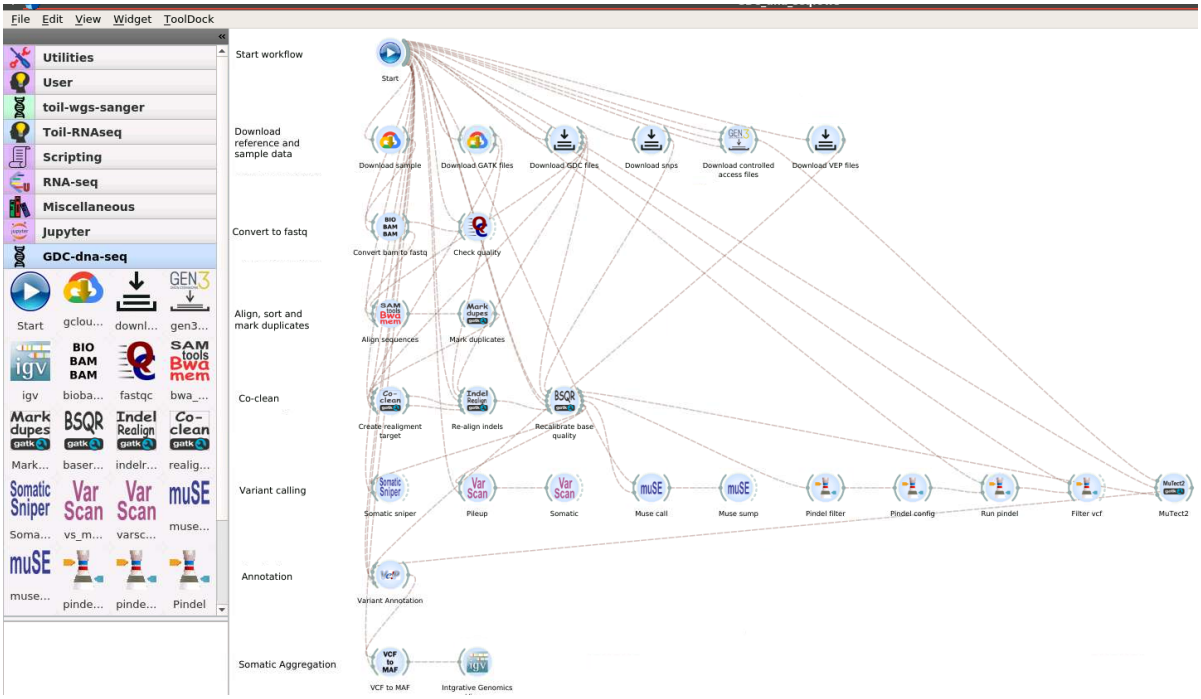


Figure 4 A screenshot of using the Cromwell widget in the Bwb to perform sequence file format validation. The results of this workflow were displayed in an open-source light-weight text editor “Geany”.

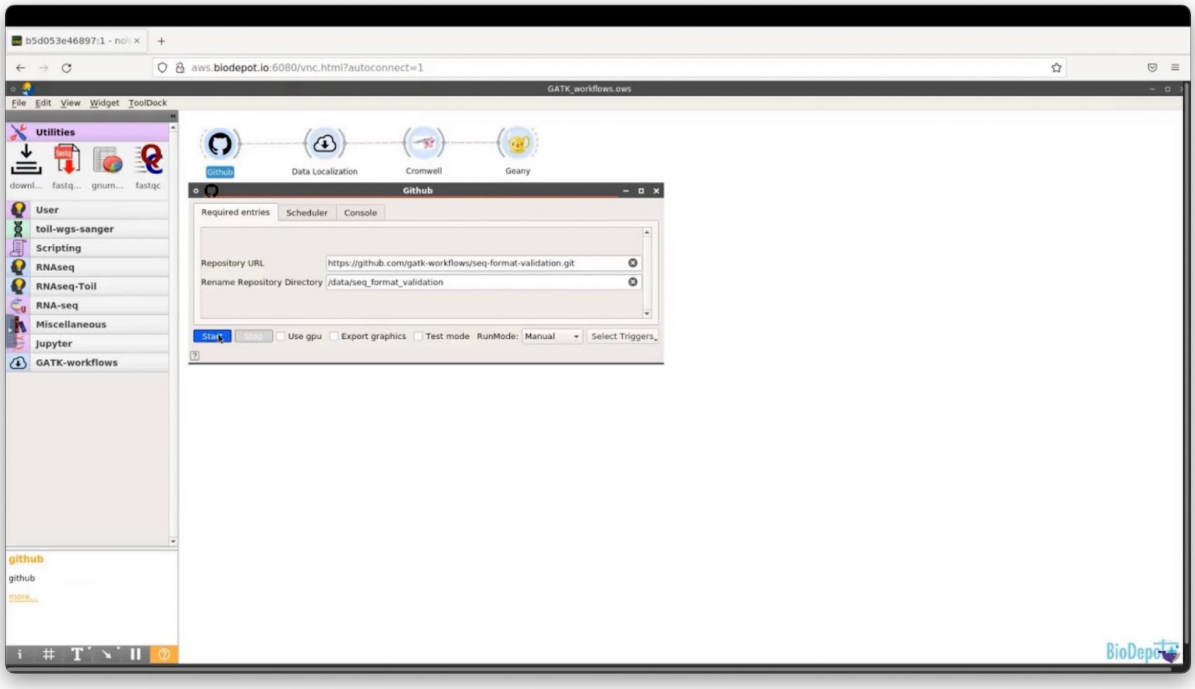


Figure 5 A screenshot of the WGS Variant Calling workflow using the Toil CWL widget. This figure depicts a screenshot of the WGS Variant Calling pipeline from the GDC DNA-seq documentation implemented using Toil-CWL. We have 3 versions of Toil that take Python, CWL, or WDL scripts along with a YAML or JSON file to specify objects and files. The widget is shown connected to a download widget which fetches the CWL and JSON files for execution using the Toil cwl-runner.

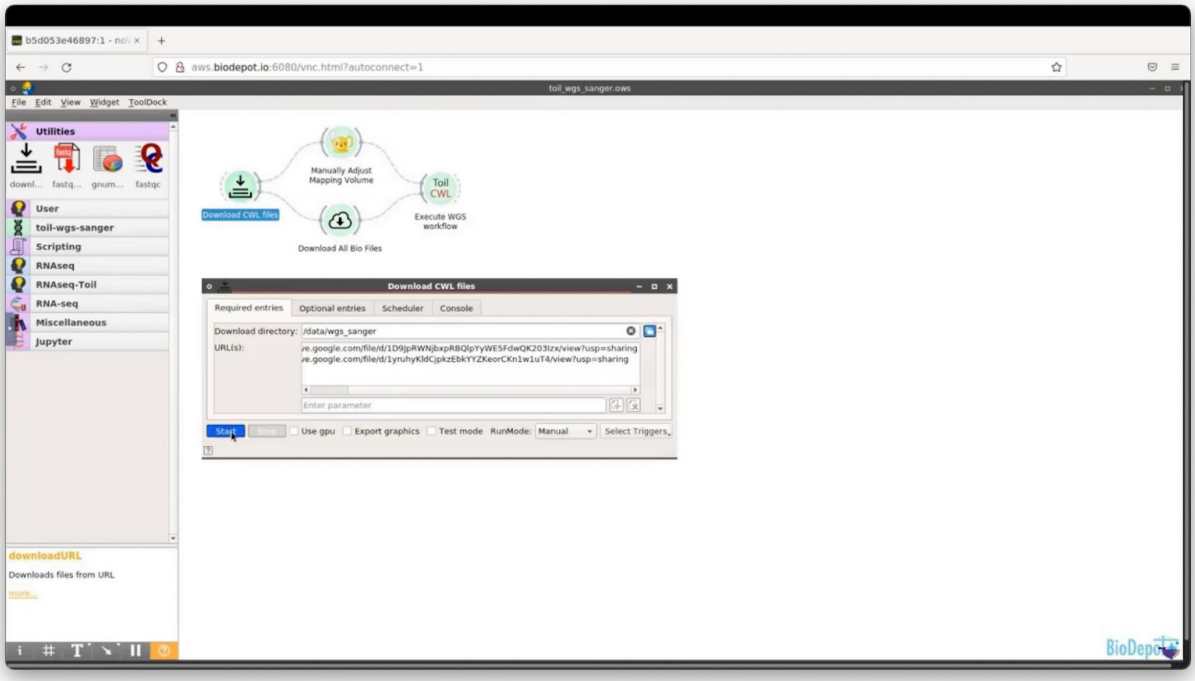
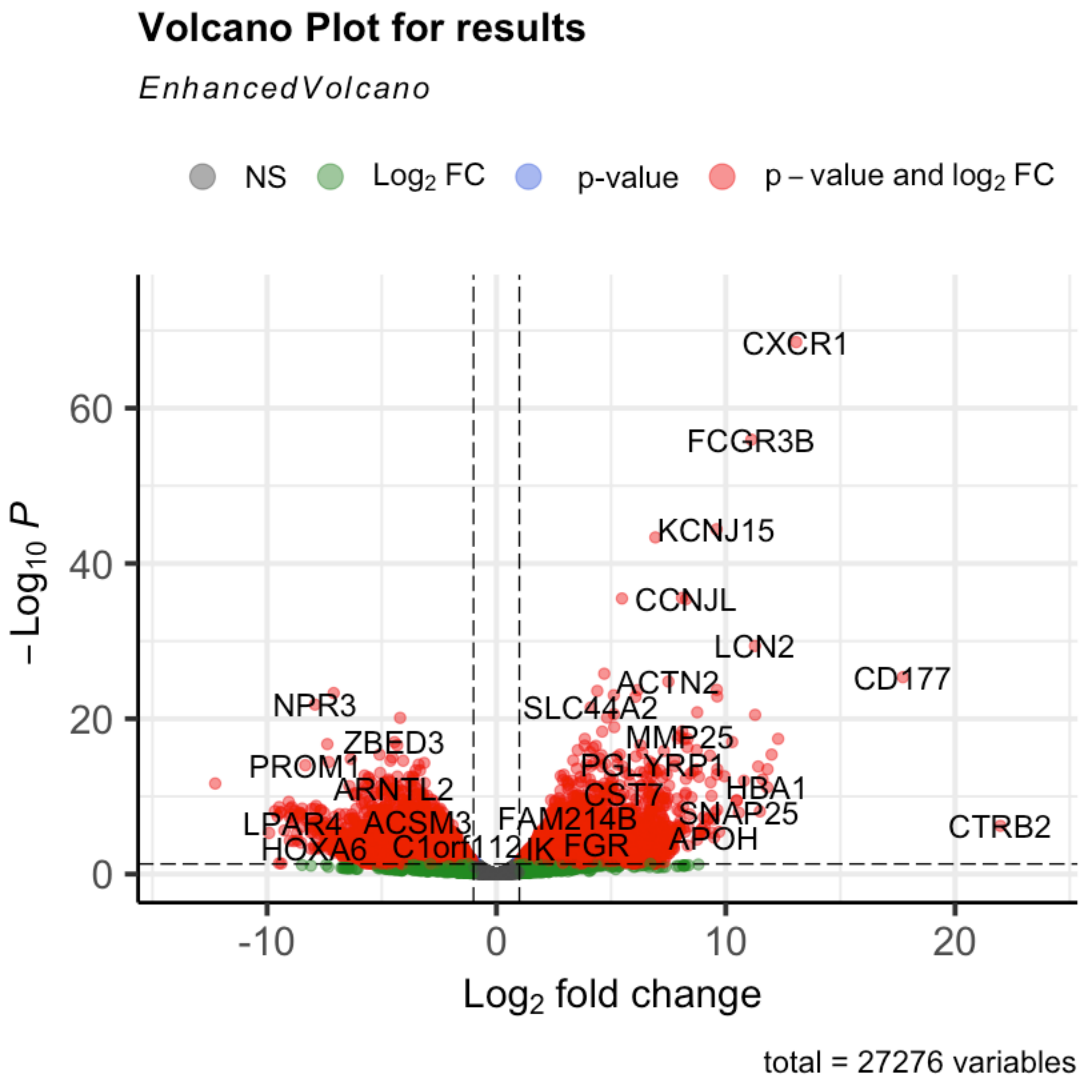


Figure 6 Volcano plot for differential genes comparing cancer vs. normal blood samples.



REFERENCES

1. **The Cancer Genome Atlas (TCGA)** [https://www.cancer.gov/tcga]
2. **TARGET: Therapeutically Applicable Research to Generate Effective Treatments** [https://ocg.cancer.gov/programs/target]
3. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, Madhavan S, Ketchum KA: **The CPTAC Data Portal: A Resource for Cancer Proteomics Research**. *J Proteome Res* 2015, **14**(6):2707-2713.
4. **NCI Cancer Research Data Commons (CRDC)** [https://datascience.cancer.gov/data-commons]

5. Wilson S, Fitzsimons M, Ferguson M, Heath A, Jensen M, Miller J, Murphy MW, Porter J, Sahni H, Staudt L *et al*: **Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API**. *Cancer Res* 2017, **77**(21):e15-e18.
6. Zhang Z, Hernandez K, Savage J, Li S, Miller D, Agrawal S, Ortuno F, Staudt LM, Heath A, Grossman RL: **Uniform genomic data analysis in the NCI Genomic Data Commons**. *Nat Commun* 2021, **12**(1):1226.
7. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM: **Toward a Shared Vision for Cancer Genomic Data**. *N Engl J Med* 2016, **375**(12):1109-1112.
8. **NCI Genomic Data Commons (GDC)** [https://gdc.cancer.gov/]
9. **GDC Data Access Processes and Tools** [https://gdc.cancer.gov/access-data/data-access-processes-and-tools]
10. **NCI GDC documentation: mRNA-seq analysis pipeline** [https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/]
11. **NCI GDC documentation: DNA-seq analysis pipeline** [https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/]
12. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.
13. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods* 2015, **12**(4):357-360.
14. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
15. Bray NL, Pimentel H, Melsted P, Pachter L: **Near-optimal probabilistic RNA-seq quantification**. *Nat Biotechnol* 2016, **34**(5):525-527.
16. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression**. *Nat Methods* 2017, **14**(4):417-419.
17. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J *et al*: **GENCODE reference annotation for the human and mouse genomes**. *Nucleic Acids Res* 2019, **47**(D1):D766-D773.

18. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D *et al*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**(D1):D733-745.
19. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data**. *Bioinformatics* 2015, **31**(2):166-169.
20. !!! INVALID CITATION !!! [12].
21. Hung LH, Hu J, Meiss T, Ingersoll A, Lloyd W, Kristiyanto D, Xiong Y, Sobie E, Yeung KY: **Building Containerized Workflows Using the BioDepot-Workflow-Builder**. *Cell Syst* 2019, **9**(5):508-514 e503.
22. **Gen3 Data Commons** [https://gen3.org/resources/user/gen3-client/]
23. **Common Workflow Language (CWL)** [https://www.commonwl.org/v1.2/]
24. **Workflow Description Language (WDL)** [https://github.com/openwdl/wdl/blob/main/versions/1.0/SPEC.md]
25. Consortium GT: **The Genotype-Tissue Expression (GTEx) project**. *Nat Genet* 2013, **45**(6):580-585.
26. **GDC Dave Tools** [https://gdc.cancer.gov/analyze-data/gdc-dave-tools]
27. **Terra** [https://terra.bio/]
28. **ISB-CGC** [https://isb-cgc.appspot.com/]
29. Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, Hahn A, Gibbs DL, Slagel J, Longabaugh WJ *et al*: **The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research**. *Cancer Res* 2017, **77**(21):e7-e10.
30. **Seven Bridges** [https://www.sevenbridges.com/]
31. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M *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.
32. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C: **Nextflow enables reproducible computational workflows**. *Nat Biotechnol* 2017, **35**(4):316-319.
33. Yukselen O, Turkyilmaz O, Ozturk AR, Garber M, Kucukural A: **DolphinNext: a distributed data processing platform for high throughput genomics**. *BMC Genomics* 2020, **21**(1):310.

34. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Gruning BA *et al*: **The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update**. *Nucleic Acids Res* 2018, **46**(W1):W537-W544.
35. Kim B, Ali T, Lijeron C, Afgan E, Krampis K: **Bio-Docklets: virtualization containers for single-step execution of NGS pipelines**. *Gigascience* 2017, **6**(8):1-7.
36. **Gnumeric** [http://www.gnumeric.org/]
37. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Brief Bioinform* 2013, **14**(2):178-192.
38. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J, Corlay S: **Jupyter Notebooks-a publishing format for reproducible computational workflows**, vol. 2016; 2016.
39. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D *et al*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature* 2012, **483**(7391):603-607.
40. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D *et al*: **Addendum: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature* 2019, **565**(7738):E5-E6.
41. **Gen3 Data Commons** [https://gen3.org/]
42. **Introducing the Data Commons Framework** [https://datascience.cancer.gov/news-events/blog/introducing-data-commons-framework]
43. Voss K, Gentry J, Van Der Auwera G: **Full-stack genomics pipelining with GATK4+ WDL+ Cromwell [version 1; not peer reviewed]. F1000Research (2017), 4**. In.; 2017.
44. Amstutz P, Tijanić N, Soiland-Reyes S, Kern J, Stojanovic L, Pierce T, Chilton J, Mikheev M, Lampa S, Ménager H: **Portable workflow and tool descriptions with the CWL (Common Workflow Language)**. *F1000Research* 2015, **4**.
45. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, Pfeil J, Narkizian J, Deran AD, Musselman-Brown A *et al*: **Toil enables reproducible, open source, big biomedical data analyses**. *Nat Biotechnol* 2017, **35**(4):314-316.
46. **GDC Legacy Archive** [https://portal.gdc.cancer.gov/legacy-archive/search/f]
47. **GDC Data Portal** [https://portal.gdc.cancer.gov/]

48. **GDC documentation GitHub** [https://github.com/NCI-GDC/gdc-docs]
49. **GTEx Portal** [https://gtexportal.org/home/datasets]
50. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, Culotti A, Ellrott K, Goecks J, Grossman RL *et al*: **Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space**. *Cell Genom* 2022, **2**(1).
51. **AnVIL: NHGRI Analysis Visualization and Informatics Lab-space** [https://anvilproject.org/]
52. **RNA-seq pipeline for the GTEx Consortium** [https://github.com/broadinstitute/gtex-pipeline/tree/master/rnaseq]
53. DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G: **RNA-SeQC: RNA-seq metrics for quality control and process optimization**. *Bioinformatics* 2012, **28**(11):1530-1532.
54. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. *Genome Biol* 2014, **15**(12):550.
55. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**. *Nat Protoc* 2009, **4**(8):1184-1191.
56. Arora S, Pattwell SS, Holland EC, Bolouri H: **Variability in estimated gene expression among commonly used RNA-seq pipelines**. *Sci Rep* 2020, **10**(1):2734.