SOFTWARE TOOL ARTICLE

REVISED Arkas: Rapid reproducible RNAseq analysis [version 2; peer review: 2 approved]

Anthony R. Colombo, Timothy J. Triche Jr, Giridharan Ramsingh

Jane Anne Nohl Division of Division of Hematology and Center for the Study of Blood Diseases, Keck School of Medicine of University of Southern California, Los Angeles, CA, 90033, USA

v2 First published: 27 Apr 2017, 6:586
<https://doi.org/10.12688/f1000research.11355.1>
Latest published: 21 Jun 2017, 6:586
<https://doi.org/10.12688/f1000research.11355.2>

Abstract

The recently introduced Kallisto pseudoaligner has radically simplified the quantification of transcripts in RNA-sequencing experiments. We offer cloud-scale RNAseq pipelines *Arkas-Quantification*, and *Arkas-Analysis* available within Illumina's BaseSpace cloud application platform which expedites Kallisto preparatory routines, reliably calculates differential expression, and performs gene-set enrichment of REACTOME pathways. Due to inherent inefficiencies of scale, Illumina's BaseSpace computing platform offers a massively parallel distributive environment improving data management services and data importing. *Arkas-Quantification* deploys Kallisto for parallel cloud computations and is conveniently integrated downstream from the BaseSpace [Sequence Read Archive](#) (SRA) import/conversion application titled *SRA Import*. *Arkas-Analysis* annotates the Kallisto results by extracting structured information directly from source FASTA files with per-contig metadata, calculates the differential expression and gene-set enrichment analysis on both coding genes and transcripts. The *Arkas* cloud pipeline supports ENSEMBL transcriptomes and can be used downstream from the *SRA Import* facilitating raw sequencing importing, SRA FASTQ conversion, RNA quantification and analysis steps.

Keywords

transcriptome, sequencing, RNAseq, automation, cloud computing

Open Peer Review**Approval Status**

	1	2
version 2 (revision) 21 Jun 2017	 view	
	↑	
version 1 27 Apr 2017	? view	 view

1. **Harold Pimentel** , Stanford University, Stanford, USA2. **Ted Abel** , University of Iowa, Iowa, USA
Marie Gaine, University of Iowa, Iowa, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Container**
Virtualization in Bioinformatics collection.

Corresponding authors: Anthony R. Colombo (anthonycolombo60@gmail.com), Timothy J. Triche Jr (tim.triche@gmail.com)

Author roles: **Colombo AR:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **J. Triche Jr T:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation; **Ramsingh G:** Conceptualization, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project was funded by grants from Leukemia Lymphoma Society-Quest for Cures (0863-15), Illumina (San Diego), STOP Cancer and Tower Cancer Research Foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Colombo AR *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Colombo AR, J. Triche Jr T and Ramsingh G. **Arkas: Rapid reproducible RNAseq analysis [version 2; peer review: 2 approved]** F1000Research 2017, **6**:586 <https://doi.org/10.12688/f1000research.11355.2>

First published: 27 Apr 2017, **6**:586 <https://doi.org/10.12688/f1000research.11355.1>

REVISED Amendments from Version 1

This revised manuscript eliminates previously unrelated discussion points such as, an in-depth examination of Docker and its role in data sharing. This version 2 is concise, explicitly stating motivations for developing the Docker software. It includes updated [Figure 1](#), which has swapped the order of images from the previous version [Figure 1](#). In addition, [Supplementary Figure S1](#) and [Supplementary Figure S2](#) were added showing the application interface. Further, this revised manuscript discusses more relevant topics such as comparing cloud computing platforms.

See referee reports

Introduction

High-performance computing based bioinformatic workflows have three main subfamilies: in-house computational packages, virtual-machines (VMs), and cloud based computational environments. The in-house approaches are substantially less expensive when raw hardware is in constant use and dedicated support is available, but internal dependencies can limit reproducibility of computational experiments. Specifically, “superuser” access needed to deploy container-based, succinct code encapsulations (often referred to as “microservices” elsewhere) can run afoul of normal permissions, and the maintenance of broadly usable sets of libraries across nodes for users can lead to shared code dynamically linking to different libraries under various user environments. By contrast, modern cloud-based approaches and parallel computing are forced by necessity to offer a user-friendly platform with high availability to the broadest audience. Platform-as-a-service approaches take this one step further, offering controlled deployment and fault tolerance across potentially unreliable instances provided by third parties such as [Amazon Web Service Elastic Compute Cloud](#) (AWS EC2) and enforcing a standard for encapsulation of developers’ services such as [Docker](#). Within this framework, the user or developer cedes some control of the platform and interface, in exchange for the platform provider handling the details of workflow distribution and execution. This has provided the best compromise of usability and reproducibility when dealing with general audiences. In this regard, the lightweight-container approach exemplified by Docker lead to rapid development and deployment compared to VMs. Combined with versioning of deployments, it is feasible for users to reconstruct results from an earlier point in time, while simultaneously re-evaluating the generated data under state-of-the-art implementations.

Docker offers advantages for reproducible research practices, and also is the principal infrastructure to leading platforms such as [Illumina’s BaseSpace](#) platform, [Google Genomics](#), [Galaxy](#) and [SevenBridges](#). Cloud computational ecosystems preserve developmental environments using Docker containerization framework, and improves bioinformatic validation. Containerized cloud applications form part of the global distributive effort and are favorable over local in-house computational pipelines because they offer rapid access to numerous public workflows, easy interfacing to archived read databases, and accelerate the upholding process of raw data.

A major bottleneck in RNAseq analysis is the processing steps for importing raw data. The majority of RNAseq analysis pipelines consist of read preparation steps, followed by computationally expensive alignment against a reference. Software for calculating transcript abundance and assembly can surpass 30 hours of computational time¹. If known or putative transcripts of defined sequences are the primary interest, then pseudoalignment, which is defined as near-optimal RNAseq transcript quantification, is achievable in minutes on a standard laptop using Kallisto software¹. *Arkas* was developed using a simple framework, yet massively parallel, for RNAseq transcript quantification that would allow users to expedite pseudoalignment on arbitrary datasets, and significantly reduce the amount of required preparatory routines.

In collaboration with Illumina (San Diego, USA) the available [BaseSpace](#) platform was already well-suited for parallel transformation of raw sequencing data into analytical results. BaseSpace has an available application *SRA Import* which automates SRA importing and FASTQ conversion pre-processing steps. The application *SRA Import* is simple requiring the SRA accession number and limits imports to 25gb per application call. *Arkas* can ingest successfully imported samples avoiding all raw data handling. For example, if one were interested in re-analyzing an experiment from SRA with reads totaling 141.3GB, *Arkas* facilitates SRA processing and state-of-the-art pseudoalignment by reducing raw sequencing data to summary quantifications totaling 1.63GB and includes an extensive analysis report of less than 10MB. The total data reduction exceeds 4 orders of magnitude with little or no loss of user-visible information. Moreover, the untouched original data is never discarded unless the user explicitly demands it. The appropriate placement of *Arkas* applications adjacent to the origin of sequencing data removes cumbersome data relocation costs and greatly facilitates sequencing archive re-analysis using state-of-the-art pseudoalignment.

Arkas, encapsulates Kallisto, automates the construction of composite transcriptomes from, quantifies transcript abundances, and implements reproducible rapid differential expression analysis coupled with gene set enrichment analysis. The *Arkas* workflow is versionized into Docker containers and publicly deployed within Illumina’s BaseSpace platform which ingests raw RNA sequencing data and completes a full analysis in approximately 2 hours.

Methods

The first step in the *Arkas* pipeline requires *Arkas-Quantification* to transform all the raw RNA sequencing data of an entire experiment into Kallisto pseudoaligned quantification output data. The second step, *Arkas-Analysis*, requires the pseudoaligned data to be input with respect to a comparison and control group, and returns a comprehensive analysis including differential expression, and gene-set enrichment.

If the user selects the defaults, *Arkas-Quantification* will complete pseudoalignment in approximately 43–60 minutes. *Arkas-Quantification* completion time is independent of the number of samples input, but is restricted to node availability (AWS EC2 node

availability is fairly high). *Arkas-Analysis*, will complete using a single node in approximately 1–1.5 hours for moderate sample group sizes ($N \leq 20$), and under 2–2.5 hours for much larger designs.

Arkas-Quantification implementation

Arkas is a two-step cloud pipeline. *Arkas-Quantification* is the first step, which reduces the computational steps required to quantify and annotate large numbers of samples against large catalogs of transcriptomes. *Arkas-Quantification* calls Kallisto for on-the-fly transcriptome indexing and quantification recursively for numerous sample directories. Kallisto quantifies transcript abundance from input RNAseq reads by using pseudoalignment, which identifies the read-transcript compatibility matrix¹. The compatibility matrix is formed by counting the number of reads with the matching alignment; the equivalence class matrix has a much smaller dimension compared to matrices formed by transcripts and read coverage. Computational speed is gained by performing the Expectation Maximization (EM) algorithm over a smaller matrix.

For RNAseq projects with many sequenced samples, *Arkas-Quantification* encapsulates expensive transcript quantification preparatory routines, while uniformly preparing Kallisto execution commands within a versionized environment encouraging reproducible protocols. The quantification step automates the index caching, annotation, and quantification associated while running the Kallisto pseudoaligner integrated within the BaseSpace environment. For users interested in quality control checks, BaseSpace offers an independent application *FastQC* which performs *fastqc* on sequencing data. The first step in the pipeline can process raw reads into transcript and pathway collection results within Illumina's BaseSpace cloud platform, quantifying against default transcriptomes such as ERCC spike-ins, ENSEMBL non-coding RNA, or cDNA build 88 for both *Homo sapiens* and *Mus musculus*; further the first step supports user uploaded FASTA files for customized analyses. *Arkas-Quantification* can support microRNAs (miRNA), however we encourage users to analyze miRNAs separately because pseudoalignment requires reducing k-mer size in the Target-DeBruijn Graph (TDBG) to miRNA sequence lengths (ranging from 16–22) which can increase path ambiguities. *Arkas-Quantification* is packaged into a Docker container and is publicly available as a cloud application within BaseSpace.

Arkas-Analysis implementation

Previous work² has revealed that filtering transcriptomes to exclude lowly-expressed isoforms can improve statistical power, while more-complete transcriptome assemblies improve sensitivity in detecting differential transcript usage. Based on earlier work by Bourgon *et al.*³, we included this type of filtering for both gene- and transcript-level analyses within *Arkas-Analysis*. The analysis pipeline automates annotations of quantification results, resulting in more accurate interpretation of coding and transcript sequences in both basic and clinical studies by just-in-time annotation and visualization.

Arkas-Analysis integrates quality control analysis for experiments that include Ambion spike-in controls, multiple normalization selections for both coding gene and transcript differential expression analysis, and differential gene-set analysis. If ERCC spike-ins,

defined by the External RNA Control Consortium⁴, are detected then *Arkas-Analysis* will calculate Receiver Operator Characteristic (ROC) plots using 'erccdashboard'⁵. The ERCC analysis reports average ERCC Spike amount volume, comparison plots of ERCC volume amount, and normalized ERCC counts (Figure 1).

Subsequent analyses import the data structure from SummarizedExperiment and creates a sub-class titled KallistoExperiment that preserves the S4 structure and is convenient for handling assays, phenotypic and genomic data. KallistoExperiment includes GenomicRanges⁶, preserving the ability to handle genomic annotations and alignments, supporting efficient methods for analyzing high-throughput sequencing data. The KallistoExperiment sub-class serves as a general-purpose container for storing feature genomic intervals and pseudoalignment quantification results against a reference genome called by Kallisto. By default KallistoExperiment couples assay data such as the estimated counts, effective length, estimated median absolute deviation, and transcript per million count where each assay data is generated by a Kallisto run; the stored feature data is a GenomicRanges object from ⁶, storing transcript length, GC content, and genomic intervals.

Given a KallistoExperiment containing the Kallisto sample abundances, principal component analysis (PCA) is performed⁷ on trimmed mean of M-value (TMM) normalized counts⁸ (Figure 2A). Differential expression (DE) is calculated on the library normalized transcript expression values, and the aggregated transcript bundles of corresponding coding genes using limma/voom linear model⁹ (Figure 3A). In addition to library normalization, we wished to add an optional data driven normalization. In the analysis pipeline, an unsupervised normalization method would not require more than a two group experimental design which was favorable due to its simplicity. Alternatively, supervised data driven normalization is a specialized task which requires users to define batch groups, and/or additional experimental groups. Further, the adjusted data must be evaluated in the context of the experiment. In-silico normalization, using factor analysis, effectively removes unwanted variation driven entirely by data¹⁰.

The analysis report returns comprehensive visualization results. PCA and DE analysis of both transcripts and coding genes is performed with easily interpretable images (Figure 2B, Figure 3B, Figure 3C). In each DE analysis FDR filtering method is defaulted to 'Benjamini-Hochberg', if there are no resultant DE genes/transcripts the FDR methods is switched to 'none'. *Arkas-Analysis* consumes the Kallisto data output from *Arkas-Quantification*, and automates DE analysis using TMM normalization and in-silico normalization on both transcript and coding gene expression in a defaulted two group experimental design, allowing customized selections. One must examine and compare the PCA sample clustering, and sample boxplots between the two methods to determine the improvement of in-silico normalization (Figure 2). If RUV improves the PCA clustering within the context of an experiment, and reduces the number of outliers observed in boxplots then it is likely that the normalization weights are useful.

Gene set differential expression, which includes gene-gene correlation inflation corrections, is calculated using Qusage¹¹. Qusage calculates the variance inflation factor, which corrects the inter-gene

Figure 1. Arkas-Analysis ERCC spike-in Controls Report. A) The Receiver Operator Characteristic plot of (detected) ERCC ratios in gene expression experiments. The X-axis shows the False Positive Rate, the Y-axis shows True Positive Rate. **B)** and **C)** shows the dynamic range of abundances of ERCC RNA amounts with a linear model fit, and ERCC RNA counts. **D)** shows a dispersion of mean transcript abundances and the estimated dispersion.

Figure 2. Arkas-Analysis Normalization Report: Normalization Analysis Using TMM and RUV. A) TMM normalization is performed on sample data and depicts the sample quartiles on normalized sample expression, PCA plot, and histogram of the adjusted p-values calculated from the DE analysis. Orange is the comparison group and green is the control group. **B)** A similar analysis is performed with RUV in-silico normalization.

correlation that results in high type 1 errors using pooled or non-pooled variances between experimental groups. The gene set enrichment is conducted using [Reactome](#) pathways constructed using ENSEMBL transcript/gene identifiers ([Figure 4](#) and [Table 1](#)); REACTOME

gene sets are not as large as other databases, so *Arkas-Analysis* outputs DE analysis in formats compatible with more exhaustive databases such as [Advaita](#). The DE files are compatible as a custom upload into Advaita iPathway guide, which offers an extensive

Figure 4. Arkas-Analysis Gene-Set Enrichment Plot. Gene-Set enrichment output report, each point represents the differential mean activity of each gene-set with 95% confidence intervals. The X-axis are individual gene-sets. The Y-axis is the $\log_2$ fold change.

Table 1. Arkas-Analysis Gene-Set Enrichment Statistics. The columns represent the Reactome pathway name corresponding to the depicted pathways in [Figure 4](#), the $\log_2$ fold change, p-value, adjusted FDR, and an active link to the Reactome website with visual depictions of the gene/transcript pathway. *Arkas-Analysis* will output a similar report testing transcript-level sets.

Pathway name	Log fold change	P.value	FDR	Gene URL
R-HAS-1989781	-0.87	0.0008	0.06	http://www.reactome.org/PathwayBrowser/#/R-HSA-1989781
R-HAS-2173796	-0.51	0.007	0.217	http://www.reactome.org/PathwayBrowser/#/R-HSA-2173796
R-HAS-6804759	-1.62	0.009	0.217	http://www.reactome.org/PathwayBrowser/#/R-HSA-6804759
R-HAS-381038	-0.43	0.013	0.226	http://www.reactome.org/PathwayBrowser/#/R-HSA-381038
R-HAS-2559585	-0.4	0.032	0.341	http://www.reactome.org/PathwayBrowser/#/R-HSA-2559585
R-HAS-4086398	-0.95	0.033	0.341	http://www.reactome.org/PathwayBrowser/#/R-HSA-4086398
R-HAS-4641265	-0.95	0.033	0.341	http://www.reactome.org/PathwayBrowser/#/R-HSA-4641265
R-HAS-422085	-1.17	0.04	0.361	http://www.reactome.org/PathwayBrowser/#/R-HSA-422085
R-HAS-5467345	-0.56	0.069	0.389	http://www.reactome.org/PathwayBrowser/#/R-HSA-5467345
R-HAS-6804754	-0.57	0.07	0.389	http://www.reactome.org/PathwayBrowser/#/R-HSA-6804754
R-HAS-6803204	-1.19	0.081	0.389	http://www.reactome.org/PathwayBrowser/#/R-HSA-6803204

Gene Ontology (GO) pathway analysis. Pathway enrichment analysis can be performed from the BaseSpace cloud system downstream from parallel differential expression analysis and can integrate with other pathway analysis software tools.

Data variance between software versions

We wished to show the importance of enforcing matching versions of Kallisto when quantifying transcripts because there is deviation of data between versions. Due to updated versions and improvements of Kallisto software, there obviously exists variation of data between algorithm versions (Figure 5, Supplementary Table 1,

Supplementary Table 2). We calculated the standardized mean differences, and the variation of the differences between the same 5 samples from Kallisto (setting bootstraps = 42) versions 0.43 and 0.43.1 (Supplementary Table 2), and found large variation of differences between raw values generated by differing Kallisto versions, signifying the importance of version analysis of Kallisto results.

Supplementary Table 1 shows the variation of the errors of the raw values such as estimated counts, effective length, and estimated median absolute deviation using the same Kallisto version 0.43.0.

Figure 5. Quantile-Quantile Plots of Data Variation Comparing Differences in Kallisto Data from Versions 0.43.1 and 0.43.0. The X-axis depicts the theoretical quantiles of the standardized mean differences. The Y-axis represents the observed quantiles of standardized mean differences.

As expected, Kallisto data generated by the same Kallisto version had very low variation of errors within the same version 0.43.0 for every transcript across all samples. However, upon comparing Kallisto version 0.43.1 to version 43.0 using the raw data such as estimate abundance counts, effective length, estimated median absolute deviation, and transcript per million values, we found, as expected, large variation of data.

Supplementary Table 2 shows that there is large variation of the differences of Kallisto data calculated between differing versions. **Figure 5** depicts the standardized mean differences, i.e. errors, between Kallisto versions fitted to a theoretical normal distribution. The quantile-quantile plots show that the errors are marginally normal, with a consistent line centered near 0 but also large outliers (**Figure 5**). As expected, containerizing analysis pipelines will enforce versionized software, which benefits reproducible analyses.

The Dockerization of Arkas BaseSpace applications versionizes the Kallisto reference index to enforce that the Kallisto software versions are identical, and further documents the Kallisto version used in every cloud analysis. The enforcement of reference versions and Kallisto software versions prevents errors when comparing experiments.

Operation

Arkas-Quantification instructions are provided within BaseSpace (details for new users can be found [here](#)). *Arkas* is a web style format, but can also be launched using the command line using **BaseSpace Command Line Interface**. The inputs are RNA sequencing samples, which may include SRA imported reads, and the outputs include the Kallisto data, .tar.gz files of the Kallisto sample data, and a report summary (**Supplementary Figure 1** and **Supplementary Figure 2**). Users may select for species type (*Homo sapiens* or *Mus musculus*), optionally correct for read length bias, and optionally select for the generation of pseudoBAMs. More significantly, users have the option to use the default transcriptome (ENSEMBL build 88) or to upload a custom FASTA of their choosing. For users that wish for local analysis, they can download the sample .tar.gz Kallisto files and analyze the data locally.

The *Arkas-Analysis* instructions are provided within the BaseSpace environment. The input for the analysis app is the *Arkas-Quantification* sample data, and the output files are separated into corresponding folders. The analysis also depicts figures for each respective analysis (**Figure 1–Figure 4**) and the images can be downloaded as a HTML format.

Annotation of coding genes and transcripts

The extraction of genomic and functional annotations directly from FASTA contig comments, eliding sometimes-unreliable dependencies on services such as **BioMart**, are calculated rapidly. The annotations were performed with a run time of 2.336 seconds (**Supplementary Table 3**) which merged the previous Kallisto data from 5 samples, creating a KallistoExperiment class with feature data containing a GenomicRanges¹¹ object with 213782 ranges and 9 metadata columns. The system runtime for creating a merged

KallistoExperiment class for 5 samples was 23.551 seconds (**Supplementary Table 4**).

Discussion

Complete transcriptomes enrich annotation information, improving downstream analyses

The choice of catalog, and the type of quantification performed, influence the results of sequencing analysis. ENSEMBL reference genomes are provided to GENCODE as a merged database from Havana's manually curated annotations with ENSEMBL's automatic curated coordinates. AceView, UCSC, RefSeq, and GENCODE have approximately twenty thousand protein coding genes, however AceView and GENCODE have a greater number of protein coding transcripts in their databases. RefSeq and UCSC references have less than 60,000 protein coding transcripts, whereas GENCODE has 140,066 protein coding loci. AceView has 160,000 protein coding transcripts, but this database is not manually curated. GENCODE is annotated with special attention given to long non-coding RNAs (lncRNAs) and pseudogenes, improving annotations and coupling automated labeling with manual curating. The database selected for protein coding transcripts can influence the amount of annotation information returned when querying gene/transcript level databases.

Although previously overlooked, lncRNAs have been shown to share features and alternate splice variants with mRNA, revealing that lncRNAs play a central role in metastasis, cell growth and cell invasion¹². LncRNA transcripts have been shown to be functional and are associated with cancer prognosis. *Arkas*' default transcriptomes include ENSEMBL (build 88) cDNA and non-coding RNA reference sequences.

Arkas cloud pipeline: modern and simple

Recent developments for virtualized operating systems, such as Docker, allow for local software environments to be preserved, whereas cloud platforms deploy the preserved software. Docker allows users to build layers of read/write access files, creating a portable operating system which exhaustively controls software versions and data, while systematically preserving the pipeline software. Currently, Docker is the principal infrastructure for cloud bioinformatic computational software platforms such as Illumina's BaseSpace platform, **Google Genomics**, **SevenBridges**, and **Galaxy**.

The Google Cloud Platform supports popular languages such as Python, Node, and Ruby with services related to computing, and storage. Google Genomics Platform has a steeper learning curve recommending familiarity with services such as **Compute Engine**, and **Cloud Storage**. Google Genomics hosts cloud storage transfer services for importing source data to storage buckets from HTTP/HTTPS locations. Data management services outside the Google Genomics platform, such as **Globus**, serves SRA database which can interact with Google Genomics applications reducing the bottleneck of SRA downloads.

Google offers very cost effective means for analyzing data, but requires expensive preparatory routines. Tatlow *et al.* employed Kallisto to pseudoalign 12,307 RNA-sequencing samples by renting

preemptible VMs from Google Cloud Platform for as little as \$0.09 per sample. Tatlow *et al.* pseudoaligned 1,811 breast carcinoma samples completing on average in 101 minutes, and 934 Cancer Cell Line Encyclopedia (CCLE) BAM files completing in 84.7 minutes on average. However large scale efforts require specialized knowledge for controlling containers (e.g. [Kubernetes](#)), manage resources, and queues. Although Tatlow *et al.* skillfully employed a cost-effective implementation of RNA-Seq analysis of massive databases, they mention a critical need for reducing the preprocessing routines involved in cloud computing¹³.

Galaxy offers shared workflows and analytical pipelines but is limited in the services related to storage due to the usage of public servers. In this light private storage platforms can flexibly store experimental data, although the range of analysis tools is not as wide compared to open-source platforms. Galaxy offers many usable tools with a wide range of visualization pipelines. In contrast, BaseSpace offers tools to accomplish specific tasks at the expense of lowering the learning curve, which may be attractive for researchers interested in immediate, and verifiable, results.

BaseSpace offers other RNAseq tools and another analysis pipeline *RNAExpress* which reduces preparatory routines. *RNAExpress* runs DESeq2 and can be used to cross validate *Arkas-Analysis*. DESeq2 uses a negative binomial distribution to model differential expression, whereas *Arkas* implements limma/voom empirical Bayes analysis pipeline. *RNAExpress* completed in 109 minutes comparing 4 controls and 4 comparison samples. Using the same samples, *Arkas-Quantification* completed in 42 minutes, and *Arkas-Analysis* completed in 54 minutes. Illumina's BaseSpace catalog of modern, yet simple, tools are attractive for users wishing share sessions, and to rapidly (re)analyze entire experiment(s).

Conclusion

Arkas integrates the Kallisto pseudoalignment algorithm into the BaseSpace cloud computation ecosystem that can implement large-scale parallel ultra-fast transcript abundance quantification. We reduce a computational bottleneck by freeing inefficiencies from utilizing rapid transcript abundance calculations and connecting accelerated quantification software to the Sequencing Read Archive. We remove the second bottleneck because we reduce the necessity of database downloading; instead we encourage users to download aggregated analysis results. We also expand the range of common sequencing protocols to include an improved gene-set enrichment algorithm, Qusage, and allow for exporting into an exhaustive pathway analysis platform, Advaita, over the AWS EC2 field in parallel.

Supplementary material

Supplementary Table 1: Data variation with matching Kallisto versions. This shows the variation of mean differences between data using the matching Kallisto version 0.43.0. The rows represent the samples from the first run using version 0.43.0. The columns represent the samples from an additional run with version 0.43.0.

[Click here to access the data.](#)

Data availability

[Data used in testing variation between versions](#)

Controls: SRR1544480 [Immortal-1](#)

SRR1544481 [Immortal-2](#)

SRR1544482 [Immortal-3](#)

Comparison: SRR1544501 [Qui-1](#)

SRR1544502 [Qui-2](#)

Software availability

Latest source code:

<https://github.com/RamsinghLab/Arkas-RNASeq>

Archived source code as at the time of publication:

DOI: [10.5281/zenodo.545654](https://doi.org/10.5281/zenodo.545654)²¹

License:

MIT license

Reference FASTA annotation files

For Homo-sapiens and Mus-musculus ENSEMBL FASTA files were downloaded [here](#) for release 88.

ERCC sequences

The ERCC sequences are provided in a SQL database format located [here](#)

Author contributions

AC wrote the manuscript, and developed the web-application and related software. TJ developed software, and helped the project design. GR wrote the manuscript and contributed to the development of software.

Competing interests

No competing interests were disclosed.

Grant information

This project was funded by grants from Leukemia Lymphoma Society-Quest for Cures (0863-15), Illumina (San Diego), STOP Cancer and Tower Cancer Research Foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Table 2: Data variation with non-matching Kallisto versions. Variation of mean differences between non-matching Kallisto versions and a randomly selected run previously generated ([Supplementary Table 1](#)). The rows are samples run using version 0.43.0, the columns are runs using version 0.43.1.

[Click here to access the data.](#)

Supplementary Table 3: Annotation runtime. System runtime for full annotation of a merged KallistoExperiment (seconds). The columns represent system runtime, the Elapsed Time is the total runtime.

[Click here to access the data.](#)

Supplementary Table 4: KallistoExperiment formation runtime. System runtime for the creation of a merged KallistoExperiment (seconds). The columns are similar to [Supplementary Table 3](#).

[Click here to access the data.](#)

Supplementary Figure S1: Arkas-Quantification Web-Style user interface. The input field for **A) Arkas-Quantification** and **B) Arkas-Analysis** demonstrating SRA re-quantification. The control and comparison samples were obtained using BaseSpace SRA Import application, and were input into the Arkas pipeline.

[Click here to access the data.](#)

Supplementary Figure S2: Arkas-Quantification output directory. **A)** Depicts the Arkas-Quantification output directory which includes sample folders containing Kallisto data. **B)** Depicts the contents of a specific folder output by Arkas-Quantification.

[Click here to access the data.](#)

References

- Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Soneson C, Matthes KL, Nowicka M, *et al.*: **Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage.** *Genome Biol.* 2016; **17**: 12.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bourgon R, Gentleman R, Huber W: **Independent filtering increases detection power for high-throughput experiments.** *Proc Natl Acad Sci U S A.* 2010; **107**(21): 9546–9551.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Baker SC, Bauer SR, Beyer RP, *et al.*: **The External RNA Controls Consortium: a progress report.** *Nat Methods.* 2005; **2**(10): 731–734.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Munro SA, Lund SP, Pine PS, *et al.*: **Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures.** *Nat Commun.* 2014; **5**: 5125.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lawrence M, Huber W, Pagès H, *et al.*: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol.* 2013; **9**(8): e1003118.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Risso D, Schwartz K, Sherlock G, *et al.*: **GC-content normalization for RNA-Seq data.** *BMC Bioinformatics.* 2011; **12**: 480.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ritchie ME, Phipson B, Wu D, *et al.*: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; **43**(7): e47.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Risso D, Ngai J, Speed TP, *et al.*: **Normalization of RNA-seq data using factor analysis of control genes or samples.** *Nat Biotechnol.* 2014; **32**(9): 896–902.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Yaari G, Bolen CR, Thakar J, *et al.*: **Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations.** *Nucleic Acids Res.* 2013; **41**(18): e170.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mitra SA, Mitra AP, Triche TJ: **A central role for long non-coding RNA in cancer.** *Front Genet.* 2012; **3**: 17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tatlow PJ, Piccolo SR: **A cloud-based workflow to quantify transcript-expression levels in public cancer compendia.** *Sci Rep.* 2016; **6**: 39259.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 07 August 2017

<https://doi.org/10.5256/f1000research.12854.r23689>

© 2017 Pimentel H. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Harold Pimentel

Department of Genetics, Stanford University, Stanford, CA, USA

Hello Colombo *et al.*,

Firstly: I am so very sorry for such a late review.

Anyway, the new manuscript looks much better. Thanks for the revisions.

I just have one nitpick: in Figure 2 you show the p-value distribution over the range (0, 0.05). Perhaps I missed something, but I'm not sure I completely understand the value of showing over this interval rather than the whole interval (0, 1). Does it have to do with the normalization adjusting p-values specifically in this range?

Regardless -- nice work and congrats!

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: RNA-Seq analysis methods and data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 24 May 2017

<https://doi.org/10.5256/f1000research.12258.r22616>

© 2017 Abel T et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ted Abel

Iowa Neuroscience Institute, University of Iowa, Iowa, USA

Marie Gaine

Iowa Neuroscience Institute, University of Iowa, Iowa, USA

This paper introduces a RNA-Seq analysis pipeline, Arkas, which combines currently available tools typically used in RNA-Seq studies. The novelty of this pipeline is the encapsulation of tools needed to prepare the data, run quality control checks, analyze the data and perform secondary analyses. This is especially beneficial for investigators new to RNA-Seq analysis with little experience navigating through computational tools. The authors take care to outline the rationale behind creating an easy-to-use interface and how this will increase reproducibility and consistency across RNA-Seq studies. They emphasize the importance of consistency with versions by showing differing results between two Kallisto versions.

However, there are some minor limitations also found in this study:

It would be beneficial to include quality control checks at the beginning of the pipeline to generate data regarding the inputted sequencing files.

It would be interesting to see more processing time information to show the benefit of using this pipeline compared to similar methods.

As is discussed, the inclusion of lncRNAs increases the amount of potentially interesting results from this pipeline. However, the authors have chosen to ignore microRNAs, an important regulator of cellular function. The inclusion of microRNAs as a default option in this pipeline would provide even more potentially interesting results.

The normalization steps and Figure 2 should be discussed in more detail. Specifically, expand on the reasons for choosing these two methods and the differences between the methods and their outputs. In addition, a note about how a user should select a normalization type would help new users.

Whilst the authors suggest that the integration of Docker will help produce reproducible research methods, the in-depth look into Docker is unnecessary, as no data has been provided to show its benefit above other options.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Molecular neuroscience

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 08 Jun 2017

Giridharan Ramsingh, Keck School of Medicine of University of Southern California, Los Angeles, USA

Thank you very much Dr. Abel for your insightful review.

The revised manuscript removed the in-depth discussion of Docker because it was too broad. The revised version included a discussion section that compares processing times between Google Genomics, and another BaseSpace application.

Your comments helped address the analysis of microRNAs. For example, Kallisto can process smaller FASTA sequences, however this invokes limitations to the construction of the Target DeBruijn Graph by increasing the path ambiguity of longer read sequences. The revised manuscript now addressed this limitation, and suggested that users analyze microRNAs separately. This analysis feature is not yet a default, but would be a great future addition. We further address details in regard to normalization motivation and selection.

As suggested by the first reviewer Dr. Pimentel, we have significantly reduced the broad discussion section, and explicitly described the motivation for the development of *Arkas*. We have additionally revised the 'Methods' section to provide a brief overview of the applications, and clearer descriptions of the interface style that included Supplementary Figures depicting both interfaces.

"This paper introduces a RNA-Seq analysis pipeline, Arkas, which combines currently available tools typically used in RNA-Seq studies. The novelty of this pipeline is the encapsulation of tools needed to prepare the data, run quality control checks, analyze the data and perform secondary analyses. This is especially beneficial for investigators new to RNA-Seq analysis with little experience navigating through computational tools. The authors take care to outline the rationale behind creating an easy-to-use interface and how this will increase reproducibility and consistency across RNA-Seq studies. They emphasize the importance of consistency with versions

by showing differing results between two Kallisto versions.

However, there are some minor limitations also found in this study:

It would be beneficial to include quality control checks at the beginning of the pipeline to generate data regarding the inputted sequencing files."

Thank you for this suggestion. Analyzing read quality will guide users into the important decision to filter low quality reads, however *Arkas* was not designed to address this. In the revised manuscript, we have now mentioned another independent BaseSpace application *FastQC* which can assess read quality. For users interested in manually uploading sequencing data to BaseSpace, each read must pass a quality filter. This quality filter will automatically reject poor quality reads, and for this we designed *Arkas* with the assumption that sequenced reads input were of good quality.

"It would be interesting to see more processing time information to show the benefit of using this pipeline compared to similar methods."

Thank you very much for addressing processing times. The revised manuscript significantly reduced the discussion section to comparisons of processing times. Your remarks inspired the addition of processing times of *Arkas*. We've included further information comparing the processing time to another BaseSpace application *RNAExpress*. Further, we added processing time information of a different Kallisto analysis pipeline implemented over Google Genomics Platform. The discussion section now is far more concise with greater relevance toward the functionality of our developed software.

"As is discussed, the inclusion of lncRNAs increases the amount of potentially interesting results from this pipeline. However, the authors have chosen to ignore microRNAs, an important regulator of cellular function. The inclusion of microRNAs as a default option in this pipeline would provide even more potentially interesting results."

Including microRNAs is a very great idea. *Arkas* can quantify microRNAs, but we decided not include microRNAs as default yet. In the revised manuscript we address that the small sequence sizes are a potential limitation to quantification of cDNAs/ncRNAs because it may increase path ambiguities during the construction of the Target DeBruijn graphs. Hence, we suggest that users analyze microRNAs separately and locally. This would be a great additional feature for the next version of *Arkas*.

"The normalization steps and Figure 2 should be discussed in more detail. Specifically, expand on the reasons for choosing these two methods and the differences between the methods and their outputs. In addition, a note about how a user should select a normalization type would help new users."

Thank you for addressing this. The revised manuscript has now explicitly stated how end-users may decide a selection of the normalization type. We further provide a brief explanation to why unsupervised normalization was selected.

"Whilst the authors suggest that the integration of Docker will help produce reproducible research methods, the in-depth look into Docker is unnecessary, as no data has been provided to show its benefit above other options."

We agree that the discussion of Docker was too broad, and the revised discussion is focused on comparative performance from other cloud platforms.

Competing Interests: None.

Reviewer Report 18 May 2017

<https://doi.org/10.5256/f1000research.12258.r22282>

© 2017 Pimentel H. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Harold Pimentel

Department of Genetics, Stanford University, Stanford, CA, USA

Note: I am a co-author of the kallisto tool, one of the tools that is used in this pipeline.

Colombo et al. describe Arkas, a tool that takes raw RNA-Seq data and produces several different types of downstream analyses. Arkas leverages existing analysis tools (e.g. kallisto and limma) and platforms (Illumina BaseSpace) to create an easy to use, fast, and reproducible pipeline. A very useful (unique?) feature is that it documents software versions and enforces consistent software versions allowing users to see the potential differences with different software versions. This is made explicit in the "Results" section.

Having all of these tools together greatly reduces the time to setup analyses and also reduces the complexity for RNA-Seq novices who might have no idea where to start. Arkas makes all of the typical figures one might make in a standard RNA-Seq analysis. It also provides gene-set analyses which are often excluded from other pipelines. In my experience, gluing together analyses from differential expression to gene-set analyses can often be an annoyance due to inconsistencies and annotations and versions of these annotations. Arkas nicely solves this problem.

While I think the idea is very good and the tool seems comprehensive, I feel the manuscript needs a bit of work. Here are a few points:

- There are a few areas where the scope seems too broad. In general, I feel that the manuscript can be shortened to be more clear as well as more precise. In particular, the Docker section in the discussion is too broad and the role of Arkas seems lost. I strongly recommend shortening this section and discussing the role of Docker in Arkas more clearly.
- While the abstract and introduction provide a description of Arkas in RNA-Seq analysis, they do not provide a motivation. It is sort of hinted in several sections in the paper, but it is not explicit.

The motivation of building another pipeline should be explicit.

- How does this pipeline compare to other pipelines such as Galaxy, DNANexus, etc.? Should probably be noted in the introduction/discussion.
- Perhaps I missed it, but the interface of Arkas does not appear to be described. There is a short subsection "Operation" that doesn't describe the type of interface. It appears to be available on Illumina BaseSpace, but does this make it a commandline tool or an online web form style tool? A short description of this interface and possibly supplementary figures (if it is a web form style) should be provided. This is unclear to folks who are not familiar with BaseSpace.
- It should be greater emphasized how this tool can be used to reanalyze existing SRA data with relative ease. In my opinion this is a very strong argument as to why one might want a tool like this.

Areas that can be shortened:

- "Data variance between software versions" can be shortened as some of this is repeated in "Results."
- "Complete transcriptomes enrich annotation information..." Specifics of annotations can probably be removed/condensed. It is probably sufficient to say that some are 3x times larger which can change results drastically.
- "Docker as a cornerstone of reproducible research" The role of Docker in general can probably be shortened and how Arkas leverages it should be made more clear.

More minor points:

- A short sentence at the beginning of "Methods" should give an overview of the two-step process.
- The Galaxy Project (<https://usegalaxy.org/>) should probably be cited even though the scope is a bit different.
- Figure 1a: "Receiver Operator Characteristic plot" of what? This is stated in the main text, but should also be stated in the figure caption.
- Swap Figure 1d and 1c.
- It seems like BaseSpace sessions can easily be shared? If so, this is an additional strong point of using BaseSpace in Arkas.

Overall, I'm very excited to see this comprehensive tool exist and be described in this paper.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: I am a co-author of the kallisto tool, one of the tools that is used in this pipeline.

Reviewer Expertise: RNA-Seq analysis methods and data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 08 Jun 2017

Giridharan Ramsingh, Keck School of Medicine of University of Southern California, Los Angeles, USA

Thank you very much Dr. Pimentel for your thorough review. We have significantly reduced the broad discussion section, and narrowed the manuscript to the most important features. The 'Abstract' and 'Introduction' section was reduced to explicitly state the motivations for the design of *Arkas*. In the revised manuscript, the 'Methods' section provides a brief overview of the applications, and the 'Operation' section describes the interface style and includes Supplementary Figures depicting both apps.

The second reviewer Dr. Abel also suggested that the in-depth discussion of Docker was too broad. The revised version includes a discussion section that compares processing times between Google Genomics, and another BaseSpace application. We also have now included brief points in regard to Galaxy.

Your helpful comments helped the manuscript become much more concise. In addition to your remarks, we have addressed important features regarding microRNAs on behalf of the second reviewer. Kallisto can process smaller FASTA sequences, however we have now addressed that users can analyze microRNAs, but we suggest a separate analysis for this.

We thank you very much for your revisions and appreciate your thoughtful remarks. We believe that addressing your remarks the manuscript is greatly elevated. Below are point-by-point responses to your questions.

"Note: I am a co-author of the kallisto tool, one of the tools that is used in this pipeline."

Colombo et al. describe Arkas, a tool that takes raw RNA-Seq data and produces several different types of downstream analyses. Arkas leverages existing analysis tools (e.g. kallisto and limma) and platforms (Illumina BaseSpace) to create an easy to use, fast, and reproducible pipeline. A very useful (unique?) feature is that it documents software versions and enforces consistent

software versions allowing users to see the potential differences with different software versions. This is made explicit in the "Results" section.

Having all of these tools together greatly reduces the time to setup analyses and also reduces the complexity for RNA-Seq novices who might have no idea where to start. Arkas makes all of the typical figures one might make in a standard RNA-Seq analysis. It also provides gene-set analyses which are often excluded from other pipelines. In my experience, gluing together analyses from differential expression to gene-set analyses can often be an annoyance due to inconsistencies and annotations and versions of these annotations. Arkas nicely solves this problem.

While I think the idea is very good and the tool seems comprehensive, I feel the manuscript needs a bit of work. Here are a few points:

- There are a few areas where the scope seems too broad. In general, I feel that the manuscript can be shortened to be more clear as well as more precise. In particular, the Docker section in the discussion is too broad and the role of Arkas seems lost. I strongly recommend shortening this section and discussing the role of Docker in Arkas more clearly."

Thank you very much for your input. In the revised manuscript, we have narrowed the Docker discussion section to the scope of BaseSpace platform, and have described Arkas' relationship to Docker as an applied infrastructure to this platform. The previous version of the manuscript detailed the role of Docker in the broad concept of reproducible research. We have omitted these details. The revised manuscript describes the interdependent relationship between Arkas and Docker in the context of BaseSpace. For example, Arkas containerized Node.js and R to parse the BaseSpace JSON input information relating to BaseSpace's input fields. The new manuscript explained that Docker and Arkas are not independent entities, and pertain specifically to BaseSpace.

"- While the abstract and introduction provide a description of Arkas in RNA-Seq analysis, they do not provide a motivation. It is sort of hinted in several sections in the paper, but it is not explicit. The motivation of building another pipeline should be explicit."

Thank you for this suggestion. We have now explicitly provided the motivation for Arkas' development by mentioning bottlenecks in RNA-sequencing such as sequencing importing and pre-processing steps, and how Arkas rectifies those bottlenecks. In the revised version, we illustrate how Arkas was developed downstream from BaseSpace *SRA Import* to greatly reduce importing and conversion steps. Also, we now explicitly stated the motivation for Arkas-Quantification such that Kallisto was implemented in parallel, which now scales quantification speed to the Amazon AWS EC2 cluster node availability rate. In addition, the revised manuscript explicitly stated the motivation for Arkas-Analysis, which provides a comprehensive analysis.

"- How does this pipeline compare to other pipelines such as Galaxy, DNANexus, etc.? Should probably be noted in the introduction/discussion."

Thank you for this suggestion. In the revised discussion section, we now compare features of other cloud platforms, and other BaseSpace RNA-Seq applications. The revised discussion now included processing times of a large scale RNA-seq analysis that implemented Kallisto using Google Genomics Platform. In addition to Google Genomics, the revised manuscript briefly compares features offered by Galaxy to BaseSpace. Further we compare Arkas to other BaseSpace RNA-Seq applications.

"- Perhaps I missed it, but the interface of Arkas does not appear to be described. There is a short subsection "Operation" that doesn't describe the type of interface. It appears to be available on Illumina BaseSpace, but does this make it a commandline tool or an online web form style tool? A short description of this interface and possibly supplementary figures (if it is a web form style) should be provided. This is unclear to folks who are not familiar with BaseSpace."

Thank you again for this suggestion. We have included a description explicitly stating that Arkas is a web form style. In addition, we included two Supplementary Figures to address the web input forms. Supplementary Figure 1 shows the input form for both web style apps, and Supplementary Figure 2 shows the output folder directory of the Arkas-Quantification.

"- It should be greater emphasized how this tool can be used to reanalyze existing SRA data with relative ease. In my opinion this is a very strong argument as to why one might want a tool like this."

Thank you for addressing reanalysis of SRA data. In the updated manuscript, we now mention that Arkas' design was motivated by the BaseSpace application *SRA Import*. The revised introduction now explicitly stated that Arkas is SRA compatible and we have provided citations for readers interested in utilizing this SRA application.

"Areas that can be shortened:

- "Data variance between software versions" can be shortened as some of this is repeated in "Results." "

We combined the "Data variance between software versions" and "Results" section into an appropriate concise section.

- "Complete transcriptomes enrich annotation information..." Specifics of annotations can probably be removed/condensed. It is probably sufficient to say that some are 3x times larger which can change results drastically."

We reduced this discussion to brief specifics of database sizes. While obvious, we believe that a brief overview provides motivation for the default transcriptomes chosen by Arkas. In the revised manuscript, we provide a very concise explanation behind the selection of default transcriptomes.

- "Docker as a cornerstone of reproducible research" The role of Docker in general can probably be shortened and how Arkas leverages it should be made more clear."

Thank you again for this comment. We agree that this broad discussion went off topic and may distract future readers. The manuscript is greatly improved with the removal of the discussion about democratization of research efforts, and biotechnology. We significantly revised the discussion to a comparison of differing cloud platforms and corresponding processing times of other cloud applications.

"More minor points:

- A short sentence at the beginning of "Methods" should give an overview of the two-step process.
"

We provided an overview of *Arkas* in the section described.

"- The Galaxy Project (<https://usegalaxy.org/>) should probably be cited even though the scope is a bit different."

Galaxy is briefly mentioned in the discussion. The revised manuscript reviewed and compared processing times of Google Genomics Platform and another RNAseq application within BaseSpace.

"- Figure 1a: "Receiver Operator Characteristic plot" of what? This is stated in the main text, but should also be stated in the figure caption.

- Swap Figure 1d and 1c."

Thank you for pointing this out. The revised Figure 1a now states that the Receiver Operator Characteristic plot is for ratios of detected and actual spiked ERCC sequences. We have swapped Figure 1d and Figure 1c.

"- It seems like BaseSpace sessions can easily be shared? If so, this is an additional strong point of using BaseSpace in Arkas."

We now mention this brief point in the discussion.

"Overall, I'm very excited to see this comprehensive tool exist and be described in this paper."

Thank you very much Dr. Pimentel.

Competing Interests: None

Comments on this article

Version 2

Author Response 07 Mar 2021

Anthony Colombo, Keck School of Medicine of University of Southern California, Los Angeles, USA

Hi,
i am an author to this software.

The software is available on the zenodo link.

The software used for this publication is available on the Ramsingh github here
(<https://github.com/RamsinghLab>)

The downstream developer code I have my own branch on my own personal github that is not as maintained.

<https://github.com/arcolombo/arkas>

All code are available.

Note: TxDbLite annotates the repeat element transcriptome and is needed. the arkas software implements the kallisto package.

The Arkas-RNASeq package is what we developed for illumina for cloud computing.

please let me know if there are any issues.

Competing Interests: No competing interests were disclosed.

Author Response 31 May 2019

Anthony Colombo, Keck School of Medicine of University of Southern California, Los Angeles, USA

For an updated version of the arkas and TxDbLite software please see my GitHub account.

Arkas : <https://github.com/arcolombo/arkas>

TxDbLite: <https://github.com/arcolombo/TxDbLite>

Competing Interests: No competing interests were disclosed.

Author Response 05 Oct 2018

Anthony Colombo, Keck School of Medicine of University of Southern California, Los Angeles, USA

Hi just saw this. All the relevant software is stored in the below link.

<https://zenodo.org/record/545654#.W7fd2XVKjCI>

Competing Interests: no competing interests.

Reader Comment 04 May 2018

Lavinia Gordon, University of Melbourne, Australia

Hello, the Github repository no longer appears to exist? - <https://github.com/RamsinghLab/Arkas-RNASeq>

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research