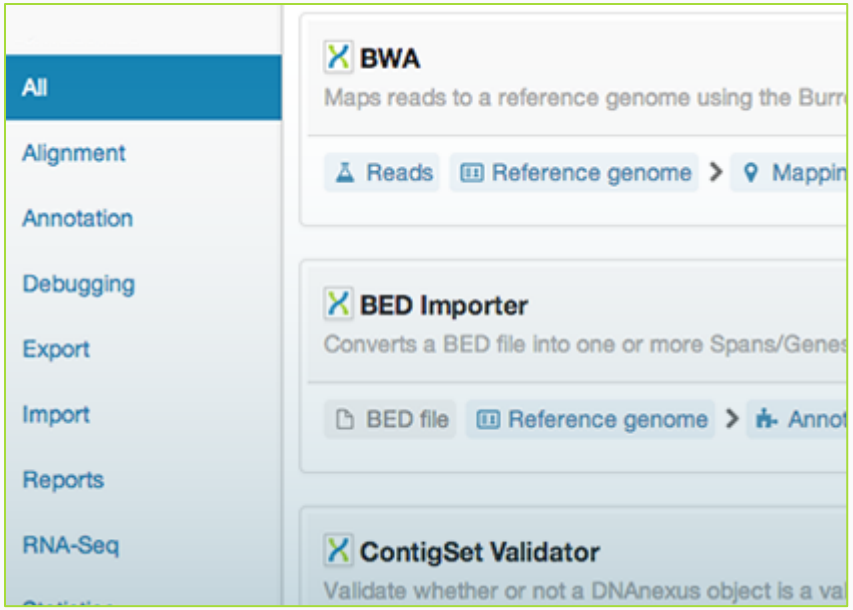
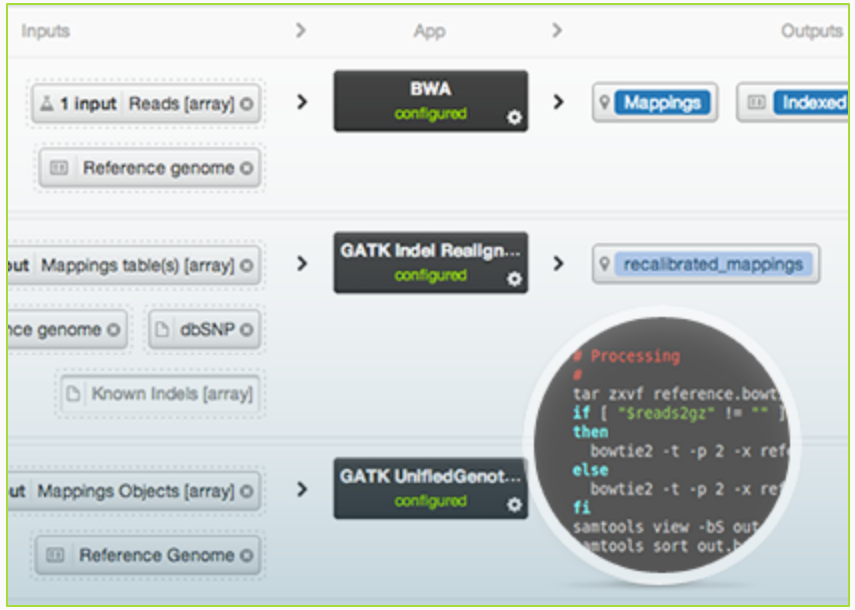
Genomics applications in the cloud with the DNAnexus Platform

Andrey Kislyuk

BOSC 2013

The DNAnexus Platform



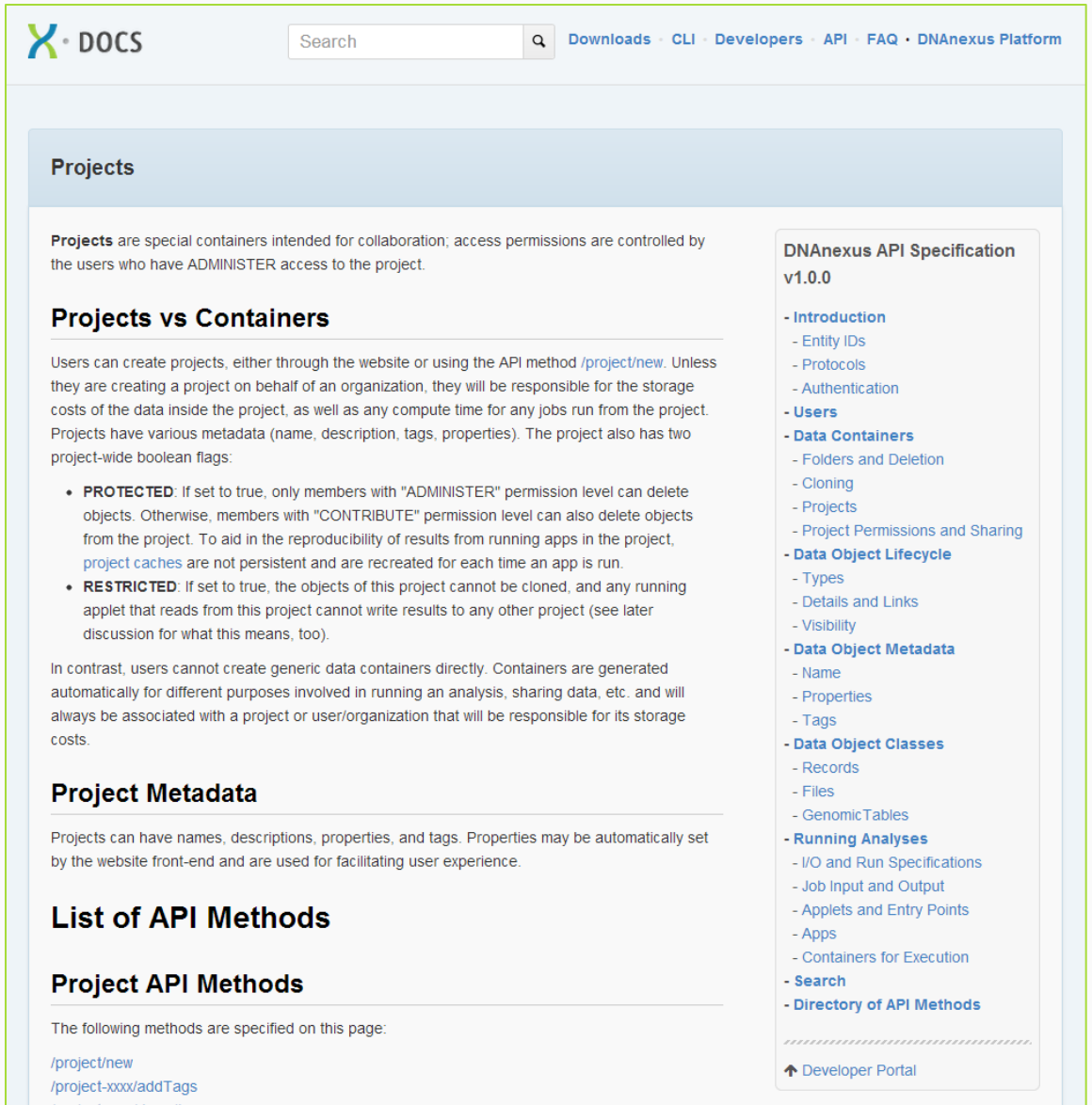


Configurable cloud infrastructure for genomics

DNAnexus

The DNAnexus Platform

Open and comprehensive API



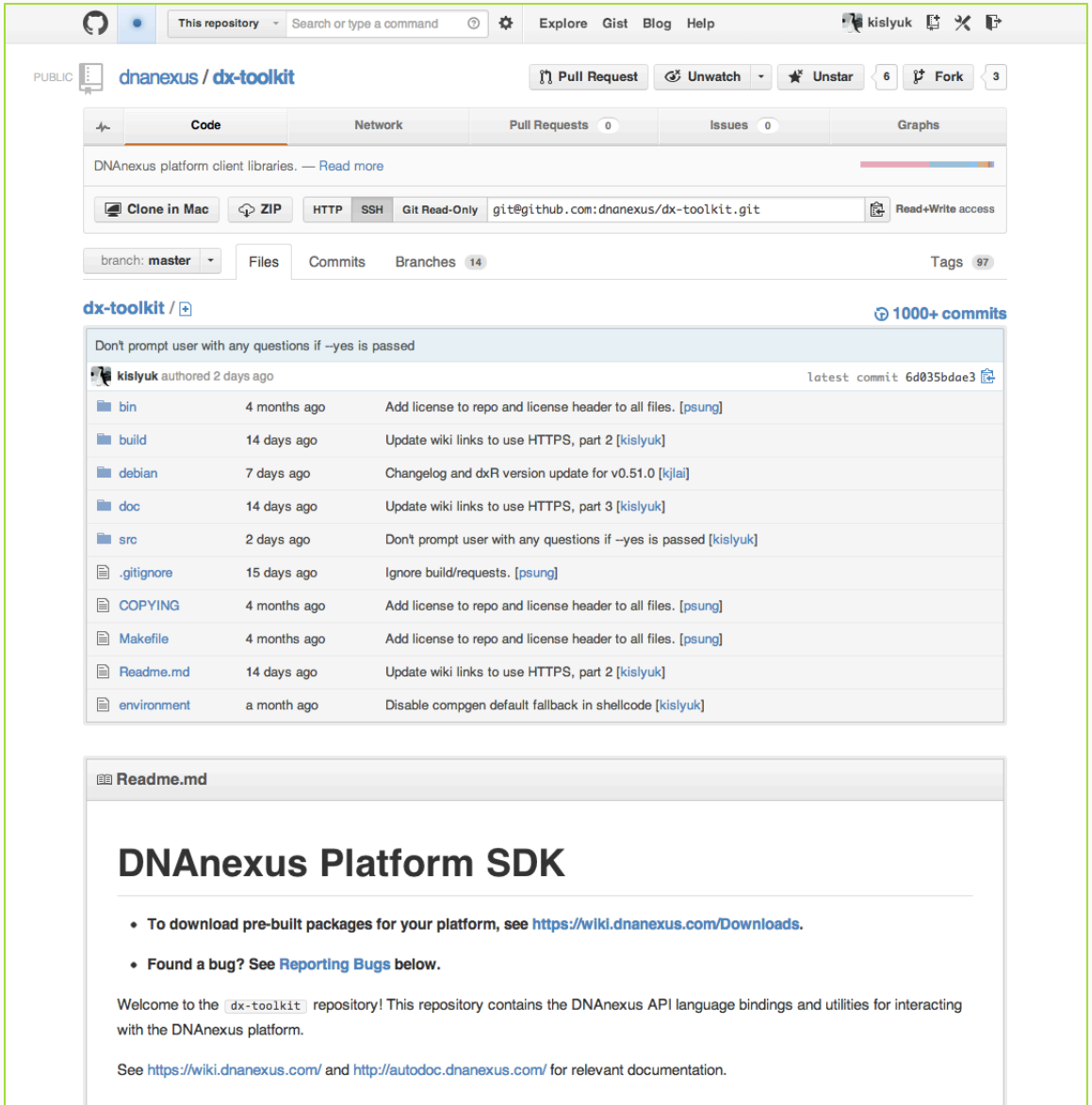
DNAnexus

The DNAnexus Platform

Open and comprehensive SDK

Linux
OS X
Windows

Upload agents (Dropbox-like)



DNAnexus

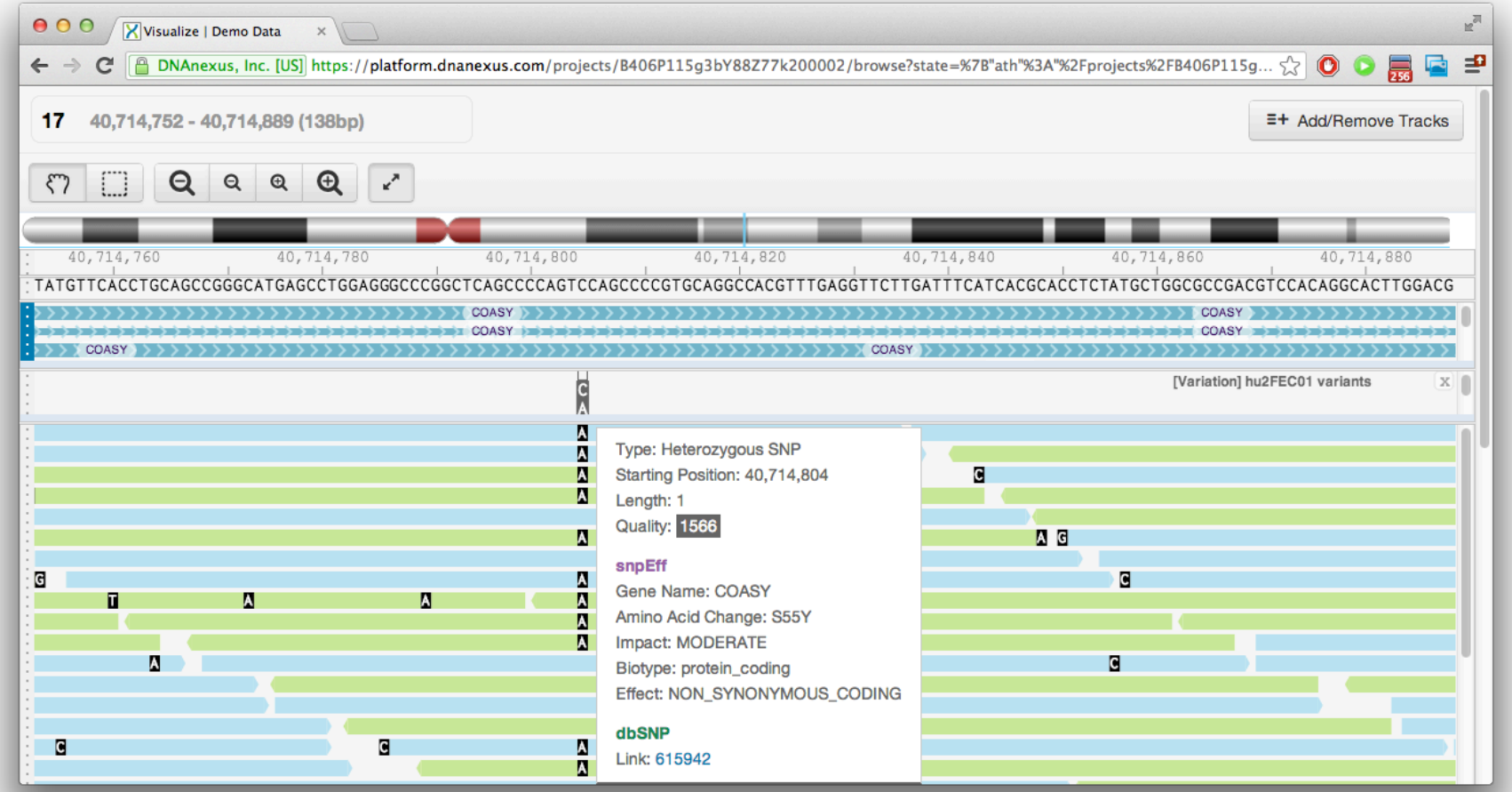
Batteries are included

The DNAnexus Platform

Powerful visualization tools built in

DNAnexus

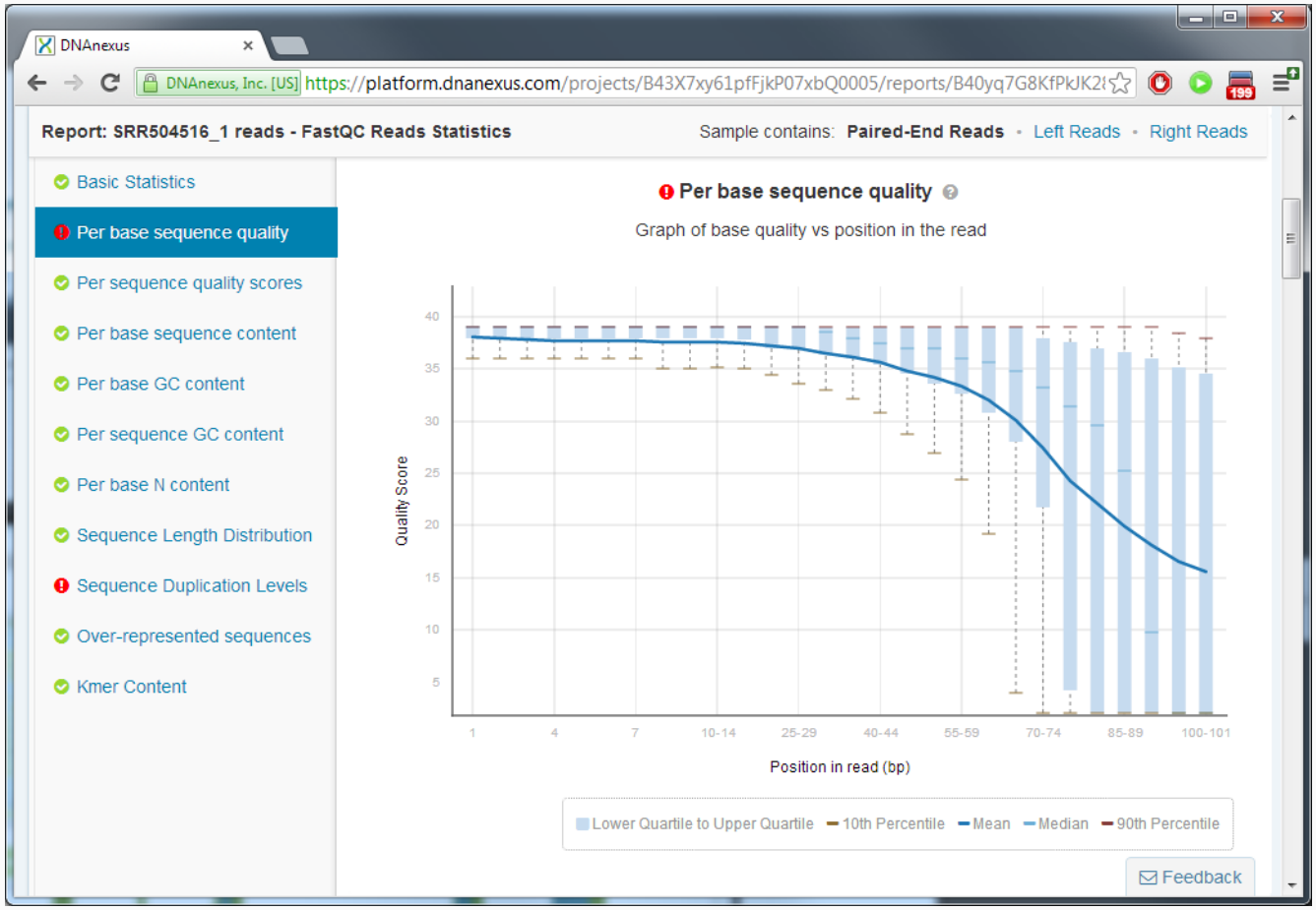
The DNAnexus Platform



Powerful visualization tools built in

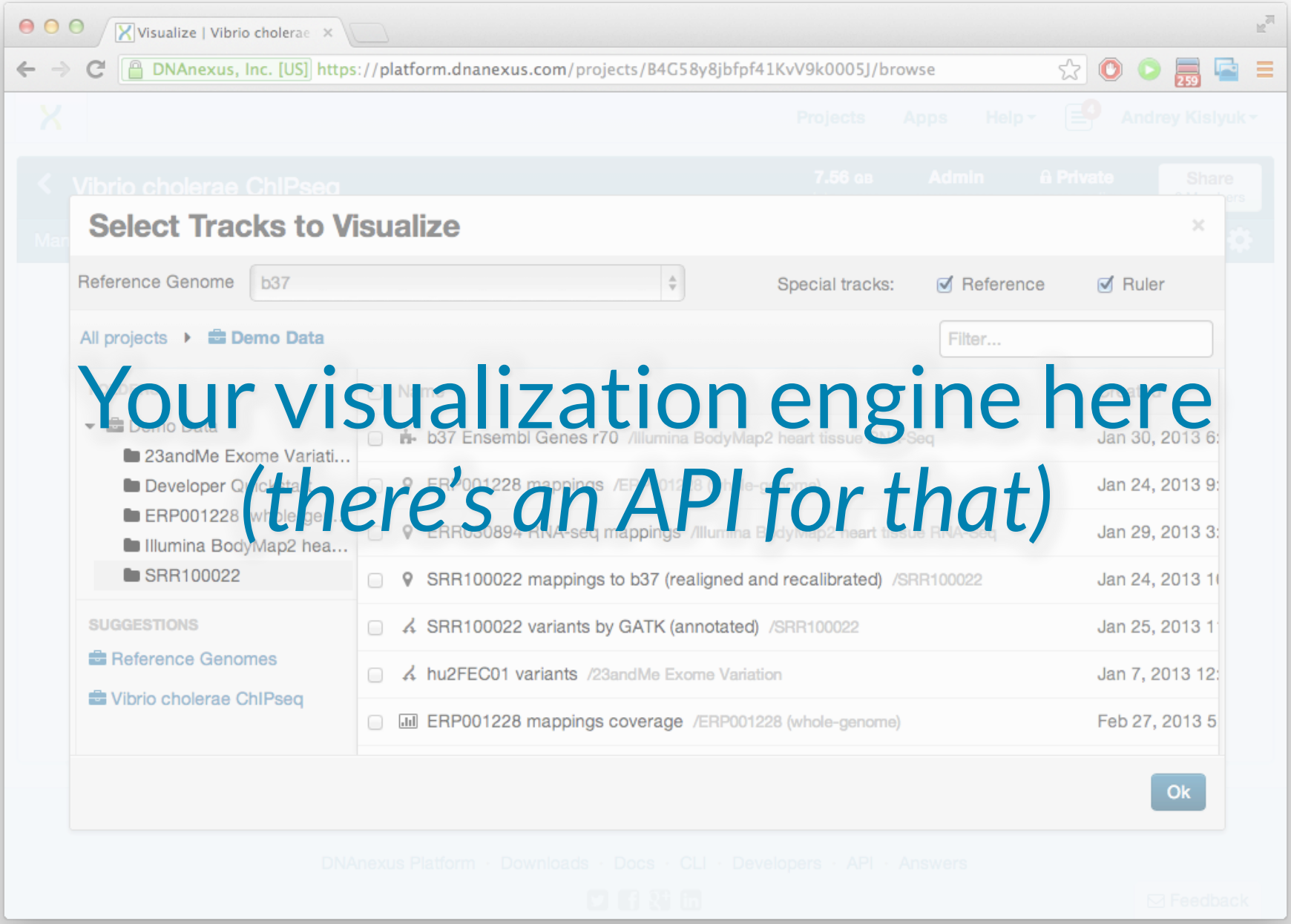
DNAnexus

The DNAnexus Platform



Quality control reports

DNAnexus

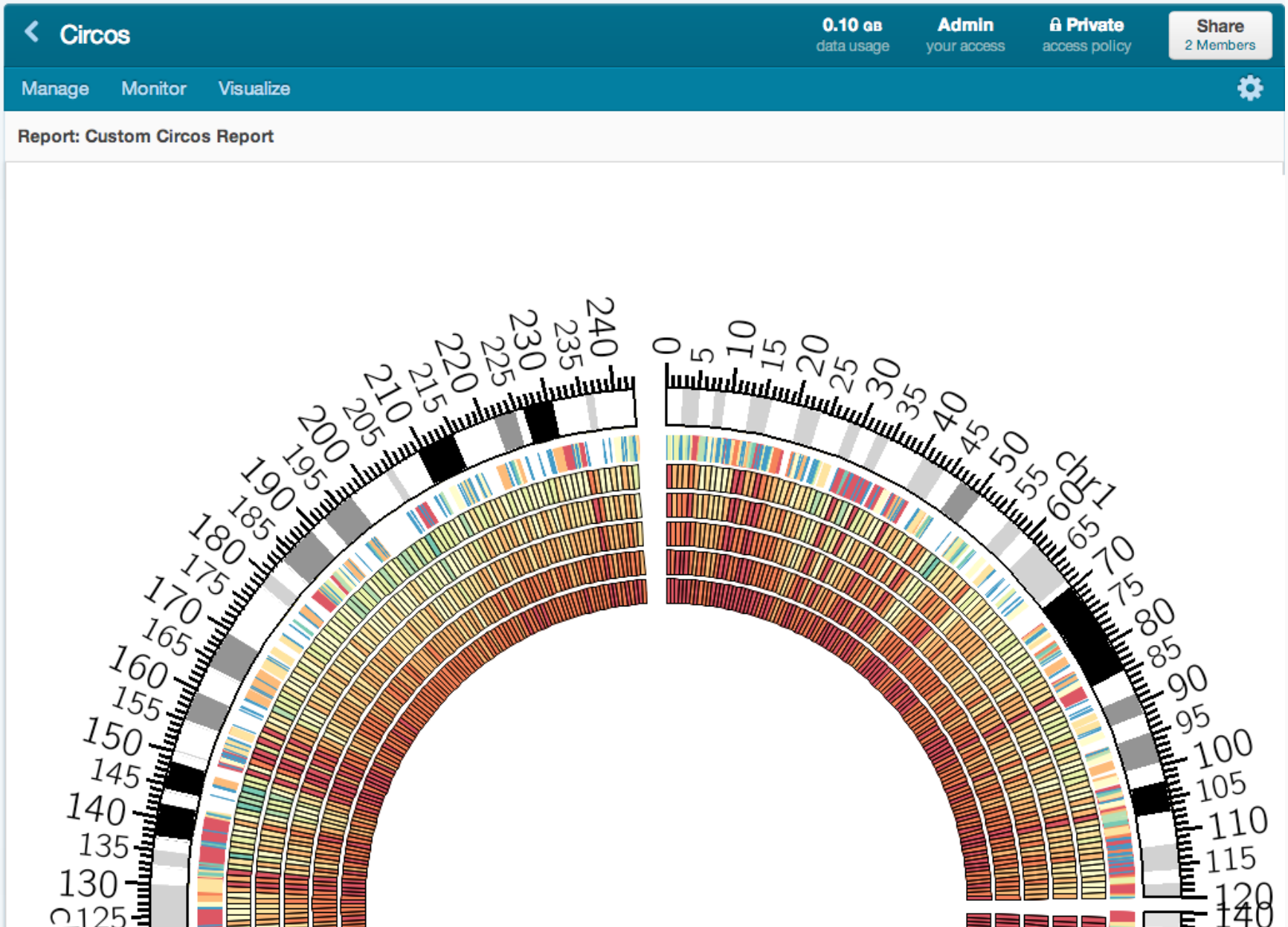


Your visualization engine here

(there's an API for that)

DNAnexus

Yes, we have Circos



DNAnexus is ready for clinical data

Secure, compliant, audited

Data compliance in accordance with:

HIPAA	GCP	CLIA
21 CFR Part 11	21 CFR Part 58	21 CFR Part 493

European Data Privacy laws and regulations

DNAnexus

Those are not just acronyms...

- All data encrypted with full-disk AES-256 at rest, SSL on the move
- Production access controls
- Third-party security audits
- Optional 2-Factor Auth
- LXC (Linux Containers) hypervisor
- Auditable by user

Your data is yours

We will **never** hold your data hostage

- Always exportable
- Always downloadable
- We're not allowed to look at it

The DNAnexus Platform

Reliability

- Triple data redundancy
- Geographically distributed
- Job-level hardware fault tolerance
- Reproducible and auditable results for 6+ years

DNAnexus is ready for clinical data

DNAnexus

Blazing fast development

Open-source stack

App wizard

walks you through app creation

Learn by example

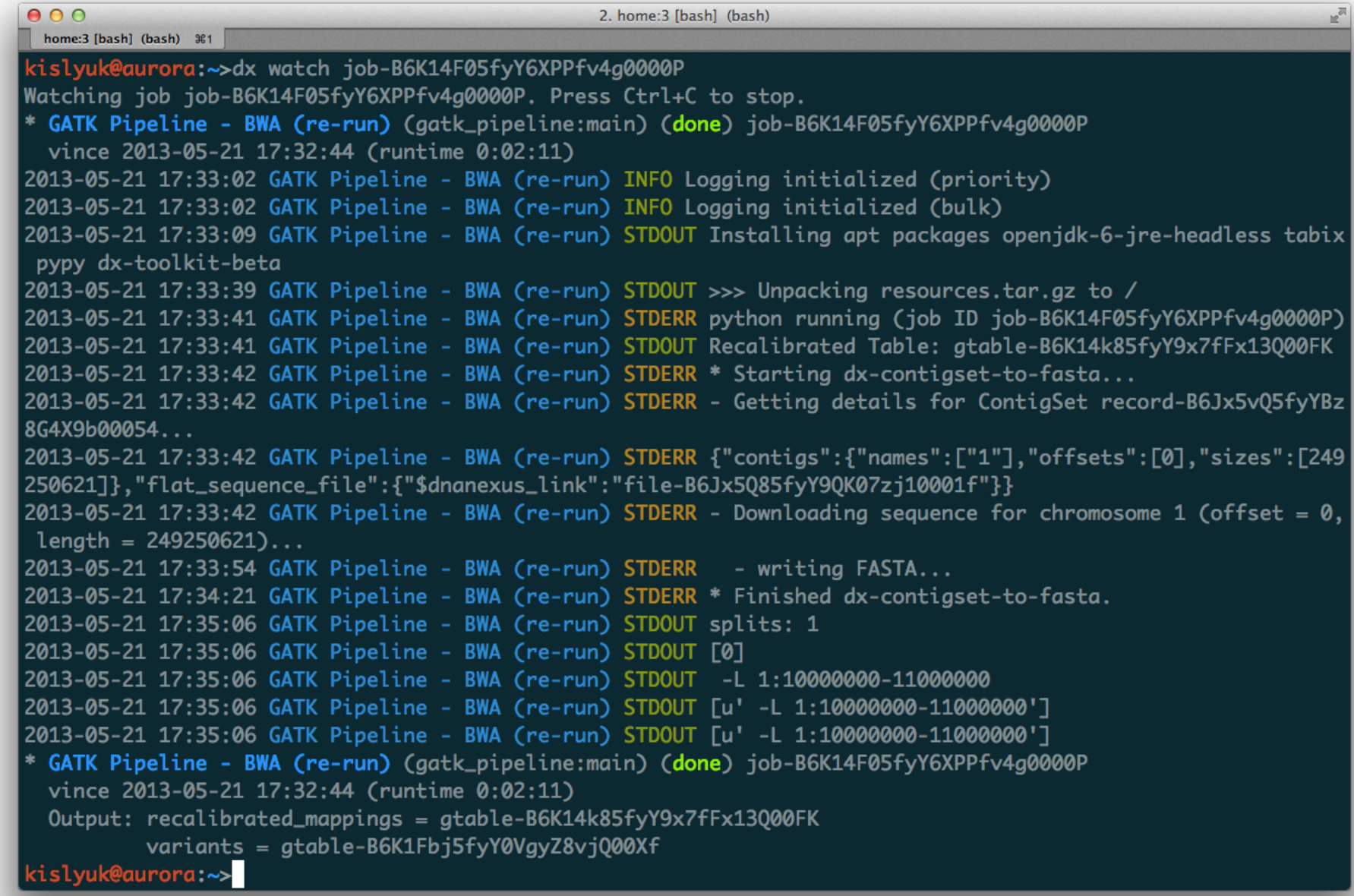
fork our repos

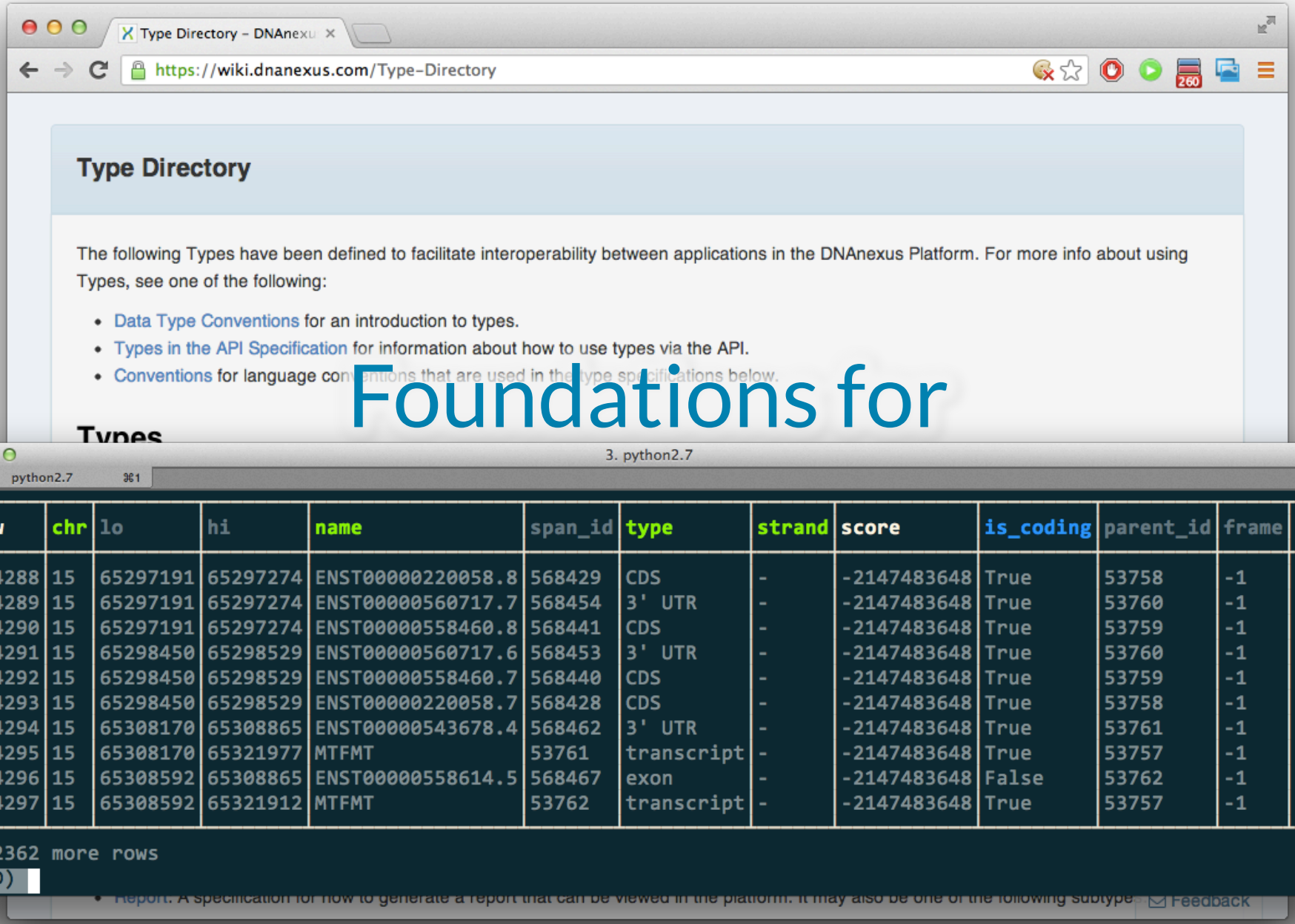
Collaborate

deploy apps from GitHub

DNAnexus

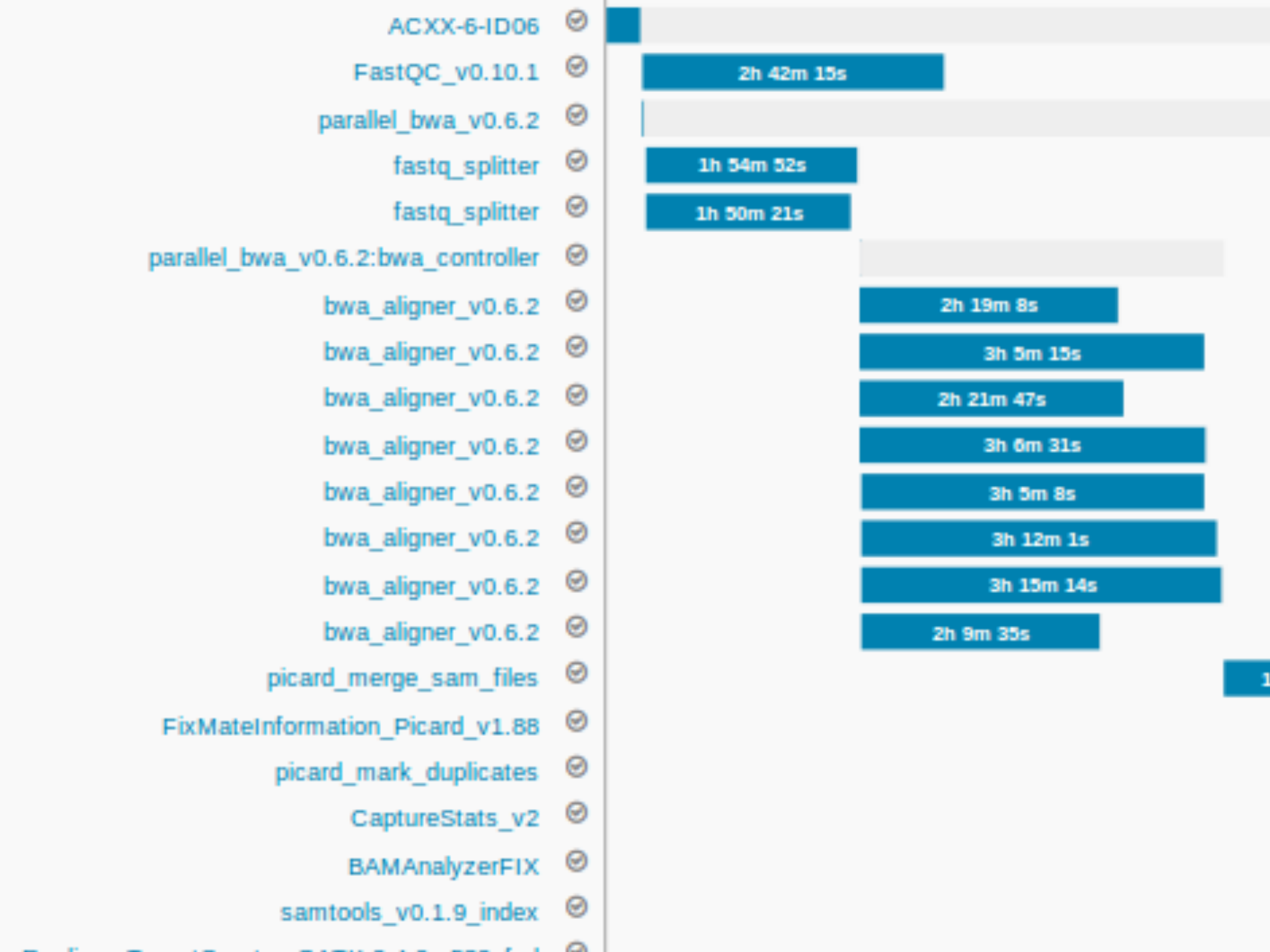
Debug quickly




Type Directory – DNAnexu
https://wiki.dnanexus.com/Type-Directory
Type Directory
The following Types have been defined to facilitate interoperability between applications in the DNAnexus Platform. For more info about using Types, see one of the following:
Data Type Conventions for an introduction to types.
Types in the API Specification for information about how to use types via the API.
Conventions for language con
Types
Foundations for
3. python2.7
python2.7 ⌘1
Row chr lo hi name span_id type strand score is_coding parent_id frame desc
524288 15 65297191 65297274 ENST00000220058.8 568429 CDS - -2147483648 True 53758 -1
524289 15 65297191 65297274 ENST00000560717.7 568454 3' UTR - -2147483648 True 53760 -1
524290 15 65297191 65297274 ENST00000558460.8 568441 CDS - -2147483648 True 53759 -1
524291 15 65298450 65298529 ENST00000560717.6 568453 3' UTR - -2147483648 True 53760 -1
524292 15 65298450 65298529 ENST00000558460.7 568440 CDS - -2147483648 True 53759 -1
524293 15 65298450 65298529 ENST00000220058.7 568428 CDS - -2147483648 True 53758 -1
524294 15 65308170 65308865 ENST00000543678.4 568462 3' UTR - -2147483648 True 53761 -1
524295 15 65308170 65321977 MTFMT 53761 transcript - -2147483648 True 53757 -1
524296 15 65308592 65308865 ENST00000558614.5 568467 exon - -2147483648 False 53762 -1
524297 15 65308592 65321912 MTFMT 53762 transcript - -2147483648 True 53757 -1
1212362 more rows
(END)
Feedback

DNAnexus

Massive on-demand compute


ACXX-6-ID06
FastQC_v0.10.1
2h 42m 15s
parallel_bwa_v0.6.2
fastq_splitter
1h 54m 52s
fastq_splitter
1h 50m 21s
parallel_bwa_v0.6.2:bwa_controller
bwa_aligner_v0.6.2
2h 19m 8s
bwa_aligner_v0.6.2
3h 5m 15s
bwa_aligner_v0.6.2
2h 21m 47s
bwa_aligner_v0.6.2
3h 6m 31s
bwa_aligner_v0.6.2
3h 5m 8s
bwa_aligner_v0.6.2
3h 12m 1s
bwa_aligner_v0.6.2
3h 15m 14s
bwa_aligner_v0.6.2
2h 9m 35s
picard_merge_sam_files
FixMateInformation_Picard_v1.88
picard_mark_duplicates
CaptureStats_v2
BAMAnalyzerFIX
samtools_v0.1.9_index

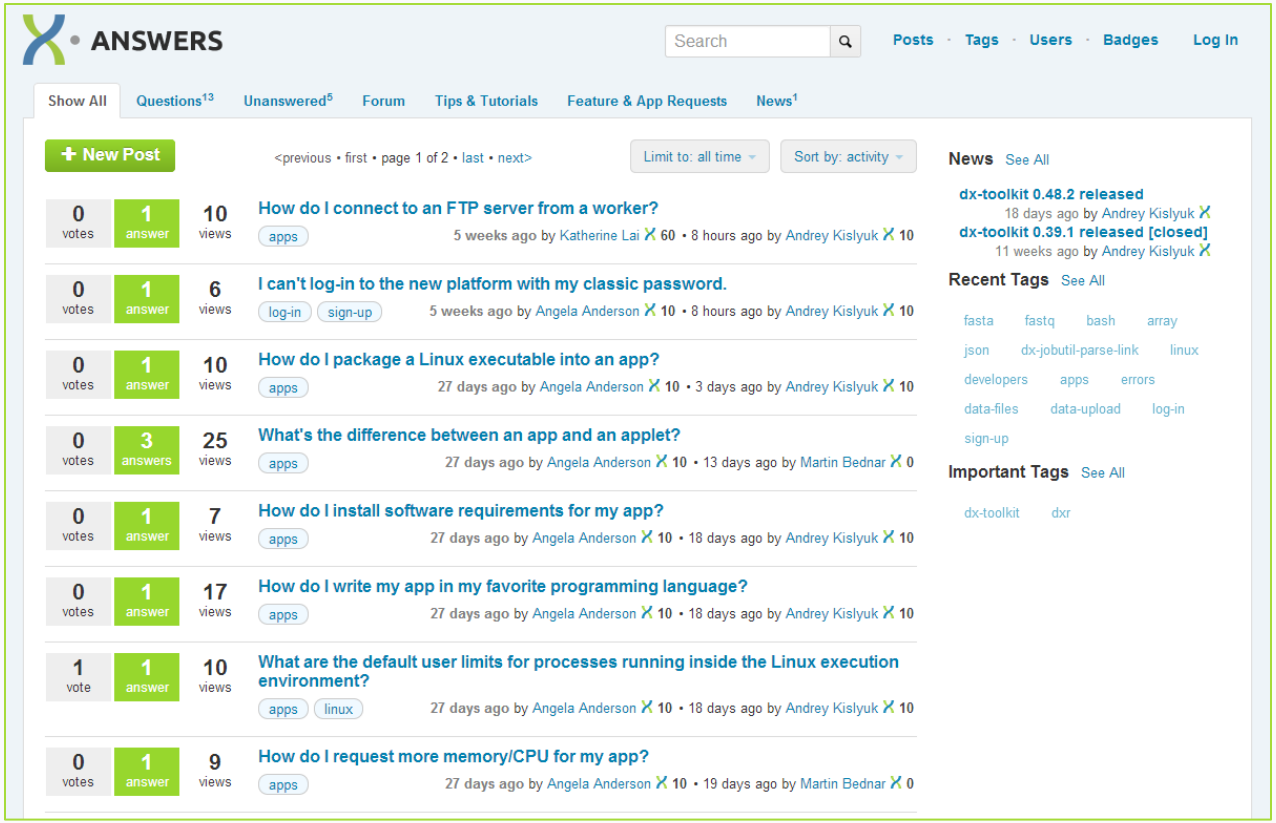
Instant-on supercomputer
at your fingertips…

…only when you need it

Spin up thousands of instances
Specify instance types
Pay by the second

Community Collaboration

DNAnexus Answers

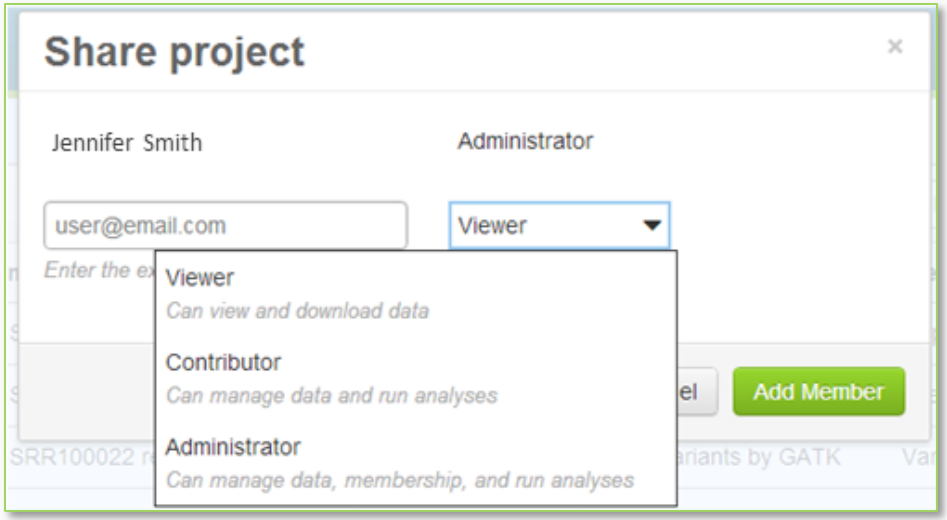


A community working together

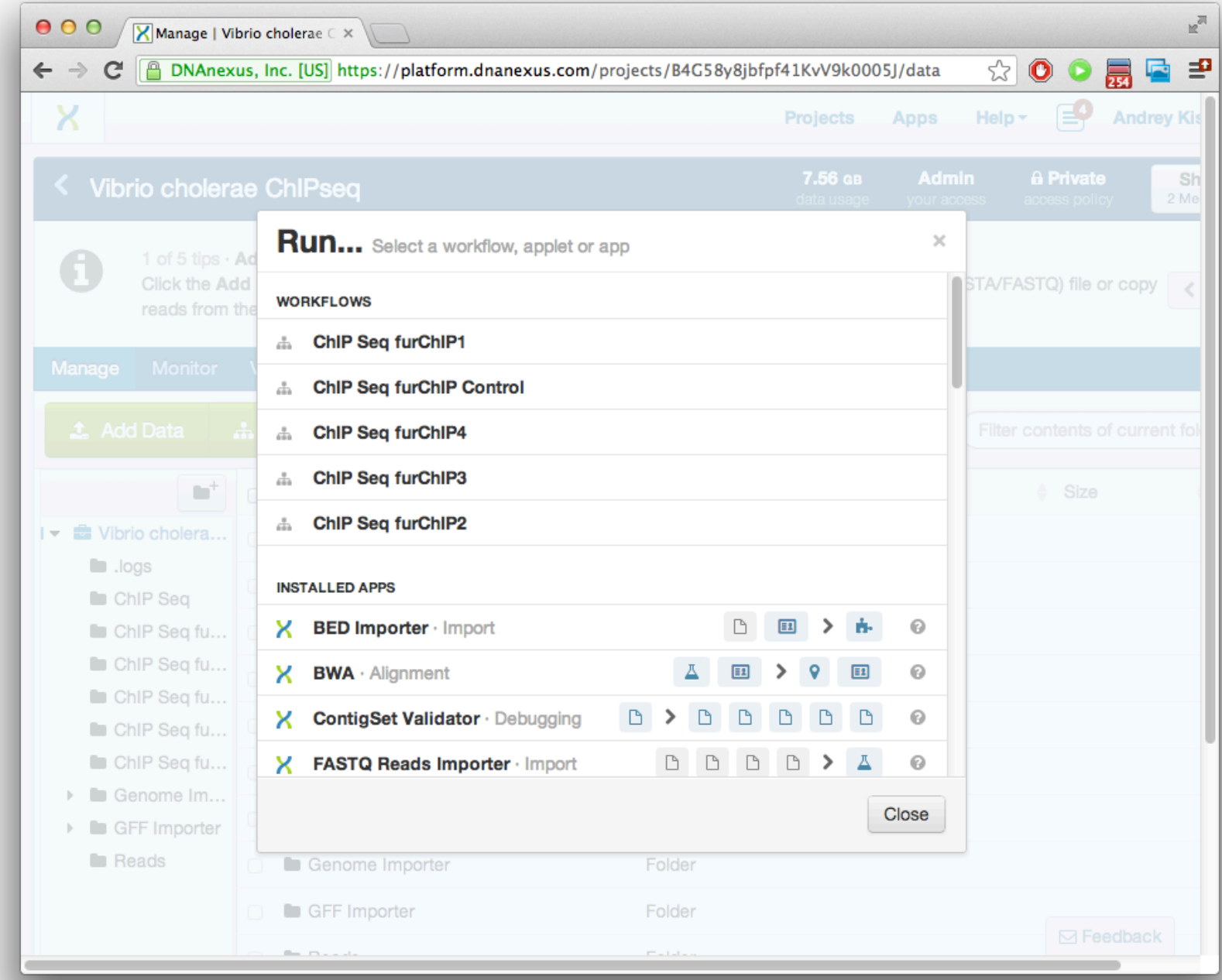
DNAnexus

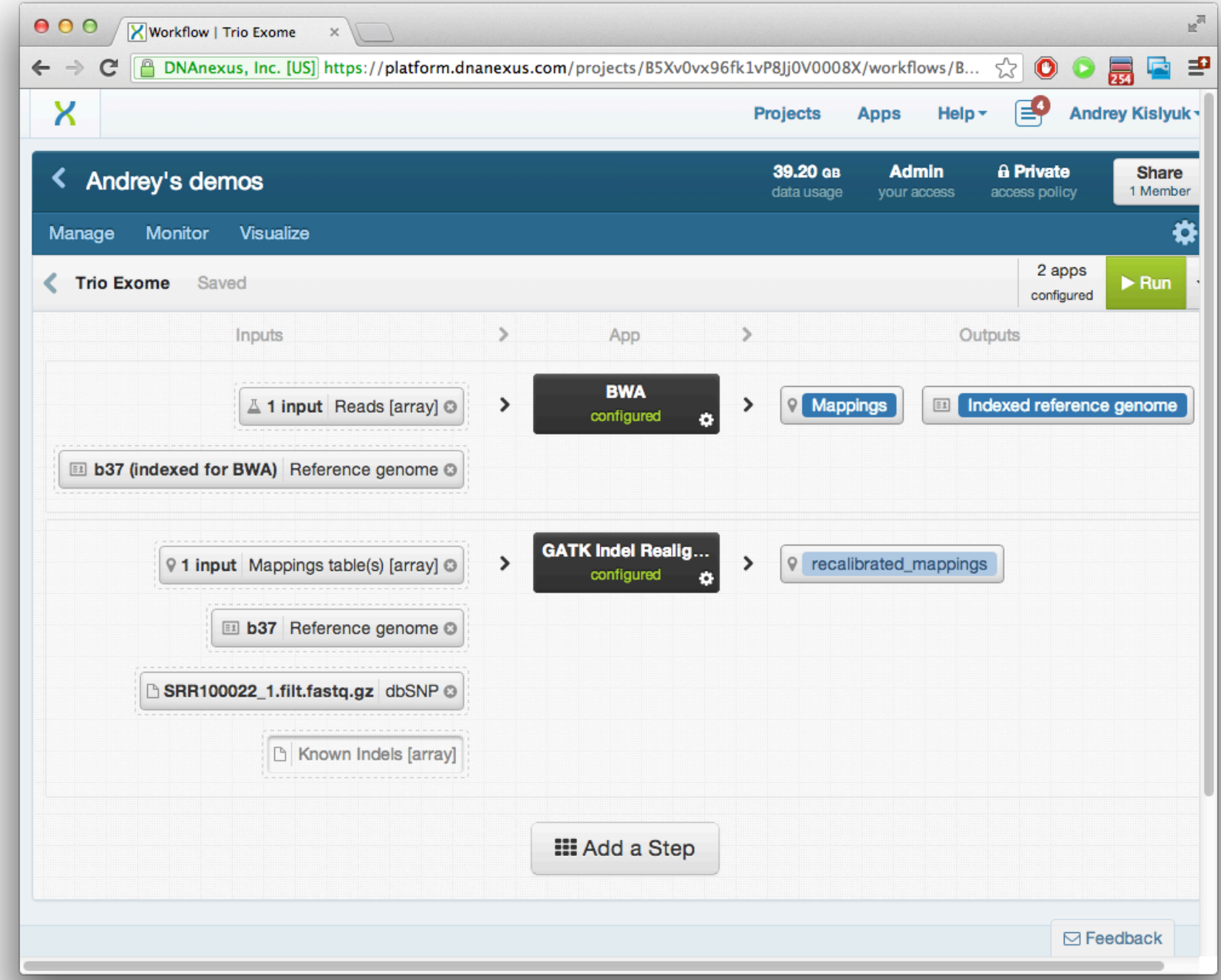
Instant collaboration

- **Eliminate** data transfer headaches
- **Collaborate** on data, tools, workflows in one environment
- **Enable** bioinformatics experts to deliver tools to biologists



DNAnexus


Manage | Vibrio cholerae
DNAnexus, Inc. [US] https://platform.dnanexus.com/projects/B4G58y8jbfpf41KvV9k0005J/data
Projects
Apps
Help
Vibrio cholerae ChIPseq
7.56 GB
data usage
Admin
your access
Private
access policy
Run... Select a workflow, applet or app
WORKFLOWS
ChIP Seq furChIP1
ChIP Seq furChIP Control
ChIP Seq furChIP4
ChIP Seq furChIP3
ChIP Seq furChIP2
INSTALLED APPS
BED Importer · Import
BWA · Alignment
ContigSet Validator · Debugging
FASTQ Reads Importer · Import
Close
Manage
Monitor
Add Data
Genome Importer
Folder
GFF Importer
Folder
Feedback

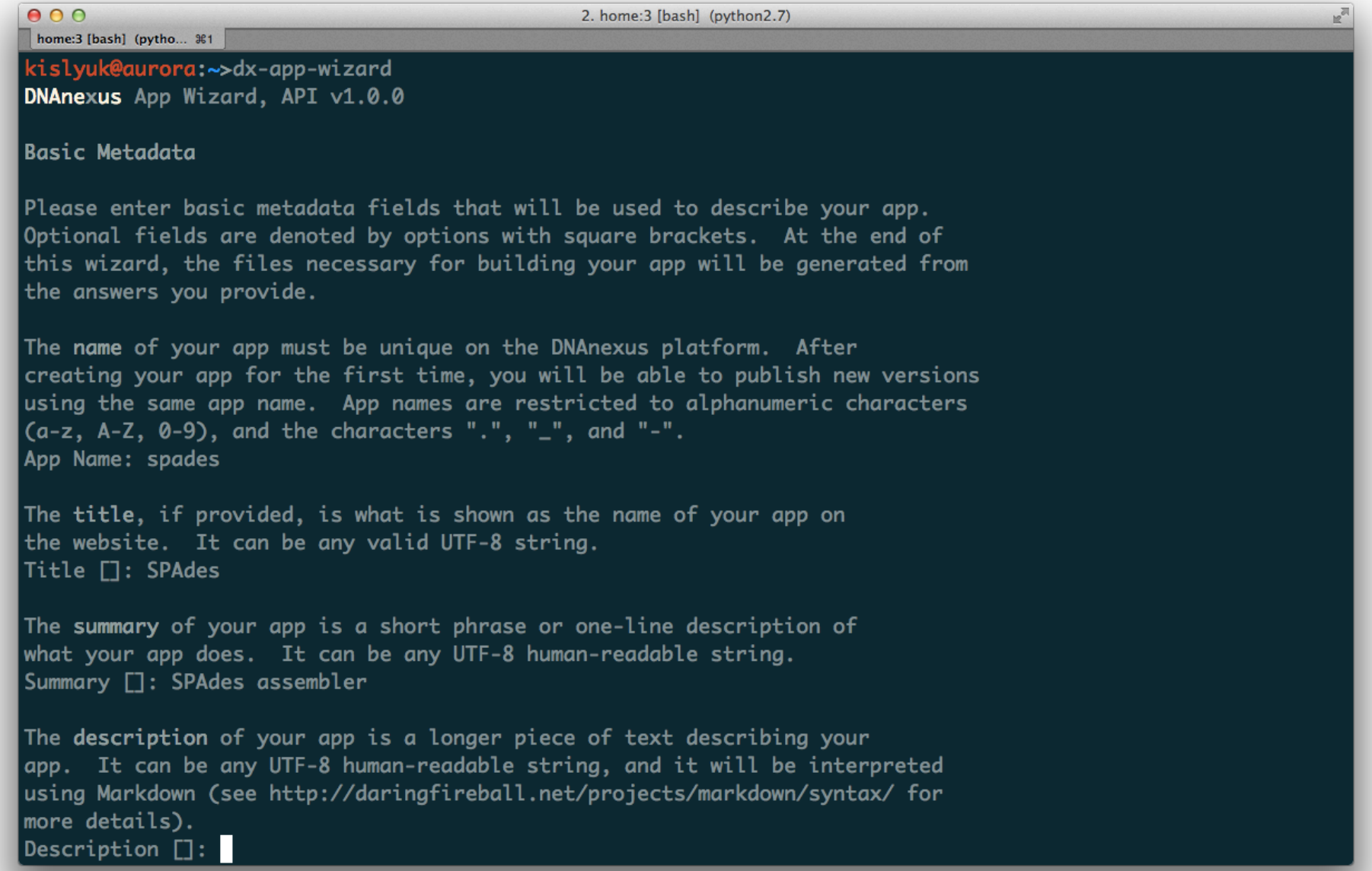

Workflow | Trio Exome
DNAnexus, Inc. [US] https://platform.dnanexus.com/projects/B5Xv0vx96fk1vP8Jj0V0008X/workflows/B...
Projects
Apps
Help
Andrey Kislyuk
Andrey's demos
39.20 GB
data usage
Admin
your access
Private
access policy
Share
1 Member
Manage
Monitor
Visualize
Trio Exome
Saved
2 apps
configured
Run
Inputs
App
Outputs
1 input Reads [array]
BWA
configured
Mappings
Indexed reference genome
b37 (indexed for BWA) Reference genome
1 input Mappings table(s) [array]
GATK Indel Realig...
configured
recalibrated_mappings
b37 Reference genome
SRR100022_1.filt.fastq.gz dbSNP
Known Indels [array]
Add a Step
Feedback

So how do I add my app?

DNAnexus

1. Run App Wizard

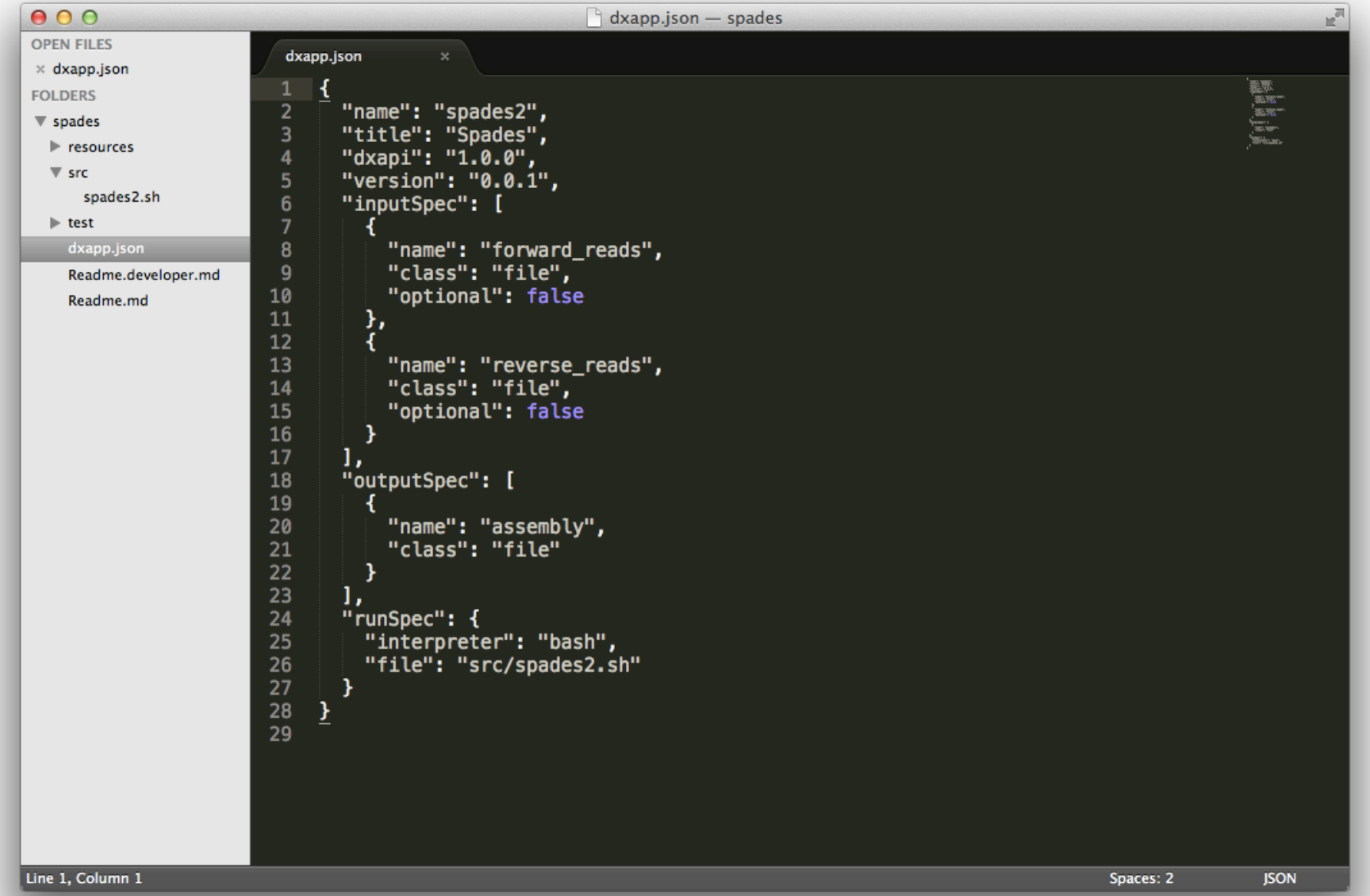
DNAnexus


2. home:3 [bash] (python2.7)
home:3 [bash] (pytho... ⌘1
kislyuk@aurora:~>dx-app-wizard
DNAnexus App Wizard, API v1.0.0
Basic Metadata
Please enter basic metadata fields that will be used to describe your app.
Optional fields are denoted by options with square brackets. At the end of
this wizard, the files necessary for building your app will be generated from
the answers you provide.
The name of your app must be unique on the DNAnexus platform. After
creating your app for the first time, you will be able to publish new versions
using the same app name. App names are restricted to alphanumeric characters
(a-z, A-Z, 0-9), and the characters ".", "_", and "-".
App Name: spades
The title, if provided, is what is shown as the name of your app on
the website. It can be any valid UTF-8 string.
Title []: SPAdes
The summary of your app is a short phrase or one-line description of
what your app does. It can be any UTF-8 human-readable string.
Summary []: SPAdes assembler
The description of your app is a longer piece of text describing your
app. It can be any UTF-8 human-readable string, and it will be interpreted
using Markdown (see http://daringfireball.net/projects/markdown/syntax/ for
more details).
Description []:

DNAnexus

app spec

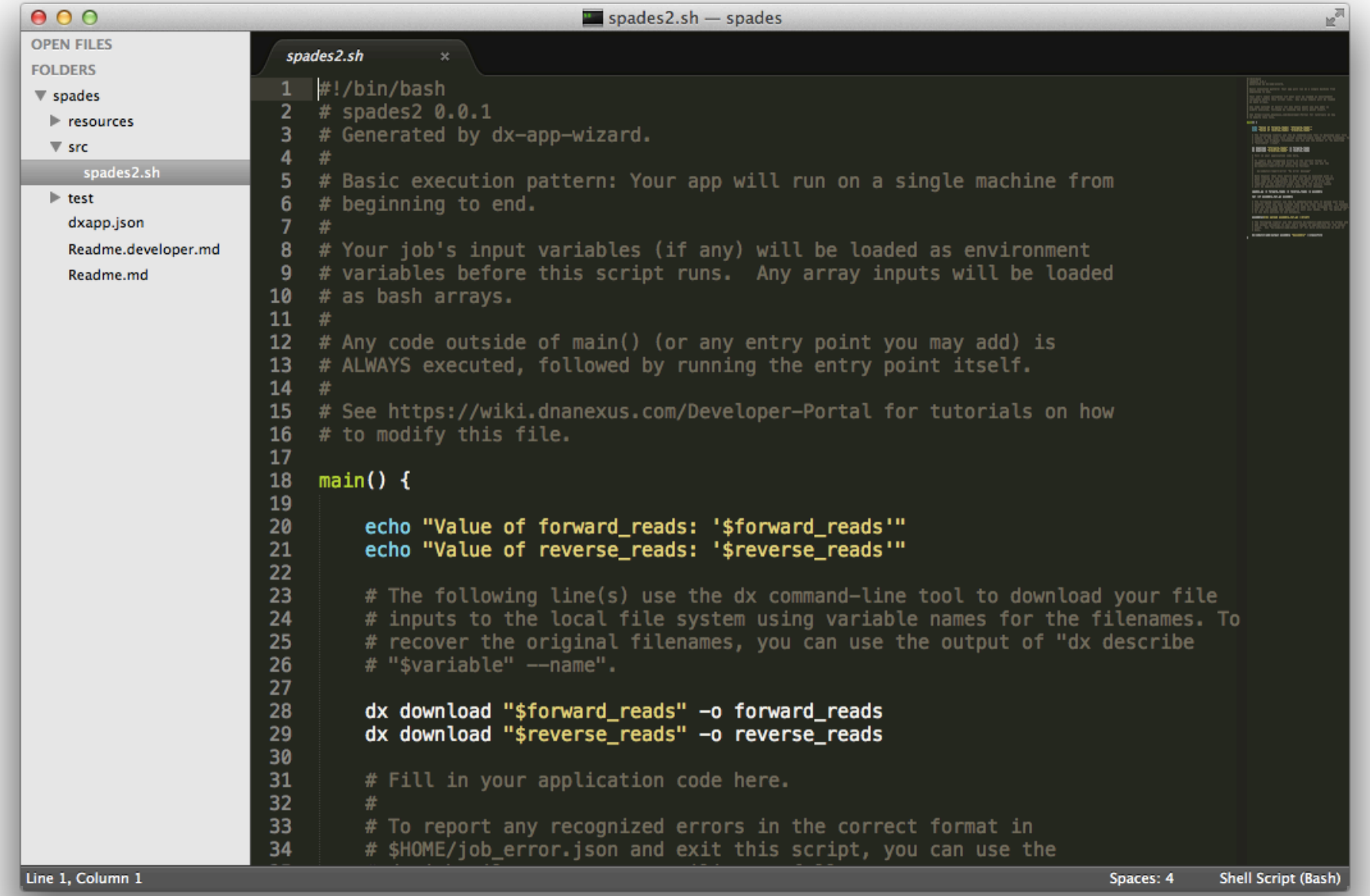
DNAnexus


dxapp.json — spades
OPEN FILES
dxapp.json
FOLDERS
spades
resources
src
spades2.sh
test
dxapp.json
Readme.developer.md
Readme.md
dxapp.json
{
"name": "spades2",
"title": "Spades",
"dxapi": "1.0.0",
"version": "0.0.1",
"inputSpec": [
{
"name": "forward_reads",
"class": "file",
"optional": false
},
{
"name": "reverse_reads",
"class": "file",
"optional": false
}
],
"outputSpec": [
{
"name": "assembly",
"class": "file"
}
],
"runSpec": {
"interpreter": "bash",
"file": "src/spades2.sh"
}
}
Line 1, Column 1
Spaces: 2
JSON

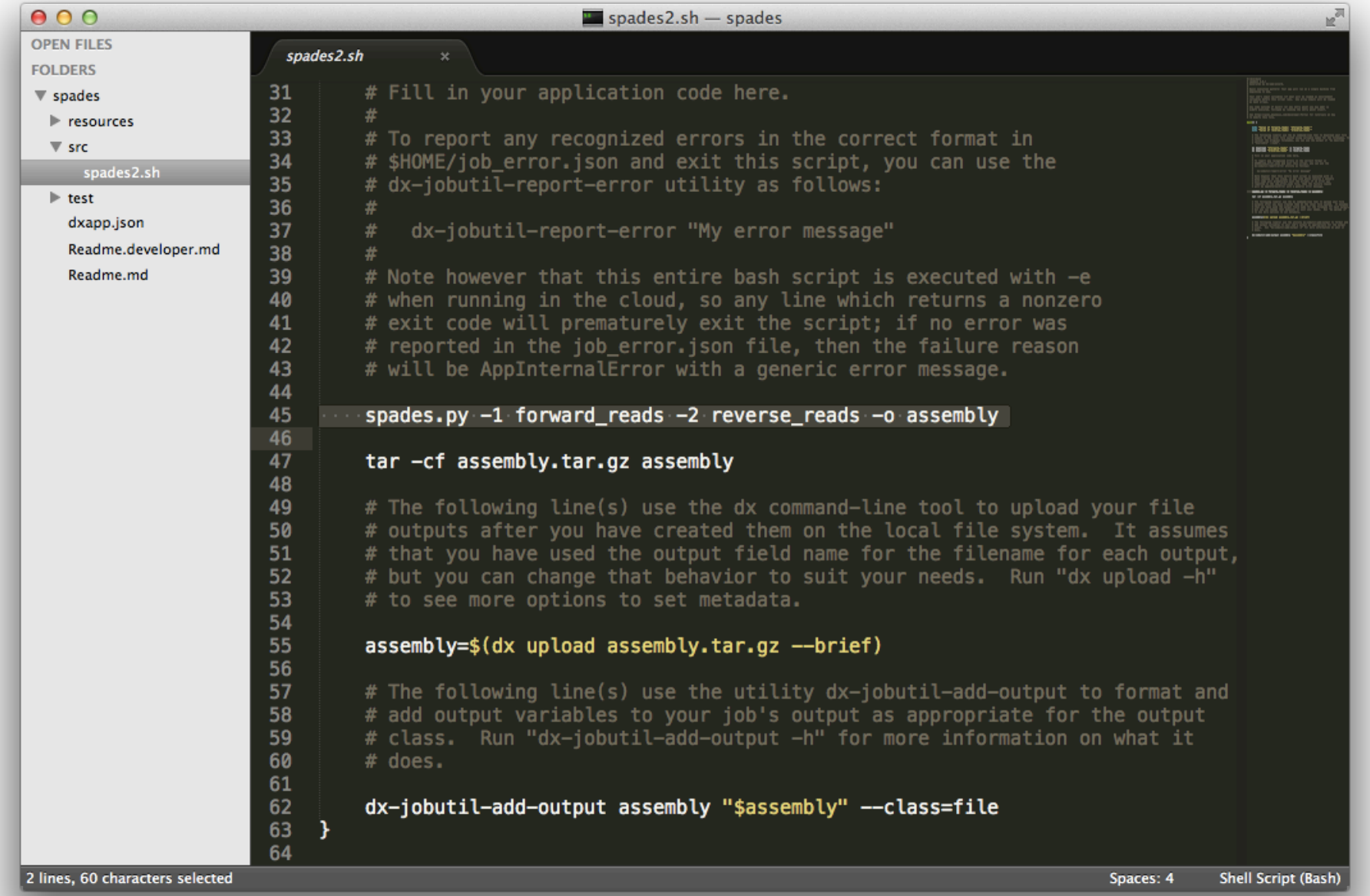
DNAnexus

2. Add entry point code

DNAnexus


spades2.sh — spades
OPEN FILES
FOLDERS
spades
resources
src
spades2.sh
test
dxapp.json
Readme.developer.md
Readme.md
spades2.sh
#!/bin/bash
spades2 0.0.1
Generated by dx-app-wizard.
#
Basic execution pattern: Your app will run on a single machine from
beginning to end.
#
Your job's input variables (if any) will be loaded as environment
variables before this script runs. Any array inputs will be loaded
as bash arrays.
#
Any code outside of main() (or any entry point you may add) is
ALWAYS executed, followed by running the entry point itself.
#
# See https://wiki.dnanexus.com/Developer-Portal for tutorials on how
# to modify this file.
main() {
echo "Value of forward_reads: '$forward_reads'"
echo "Value of reverse_reads: '$reverse_reads'"
The following line(s) use the dx command-line tool to download your file
inputs to the local file system using variable names for the filenames. To
recover the original filenames, you can use the output of "dx describe
"$variable" --name".
dx download "$forward_reads" -o forward_reads
dx download "$reverse_reads" -o reverse_reads
Fill in your application code here.
#
To report any recognized errors in the correct format in
$HOME/job_error.json and exit this script, you can use the
Line 1, Column 1
Spaces: 4
Shell Script (Bash)

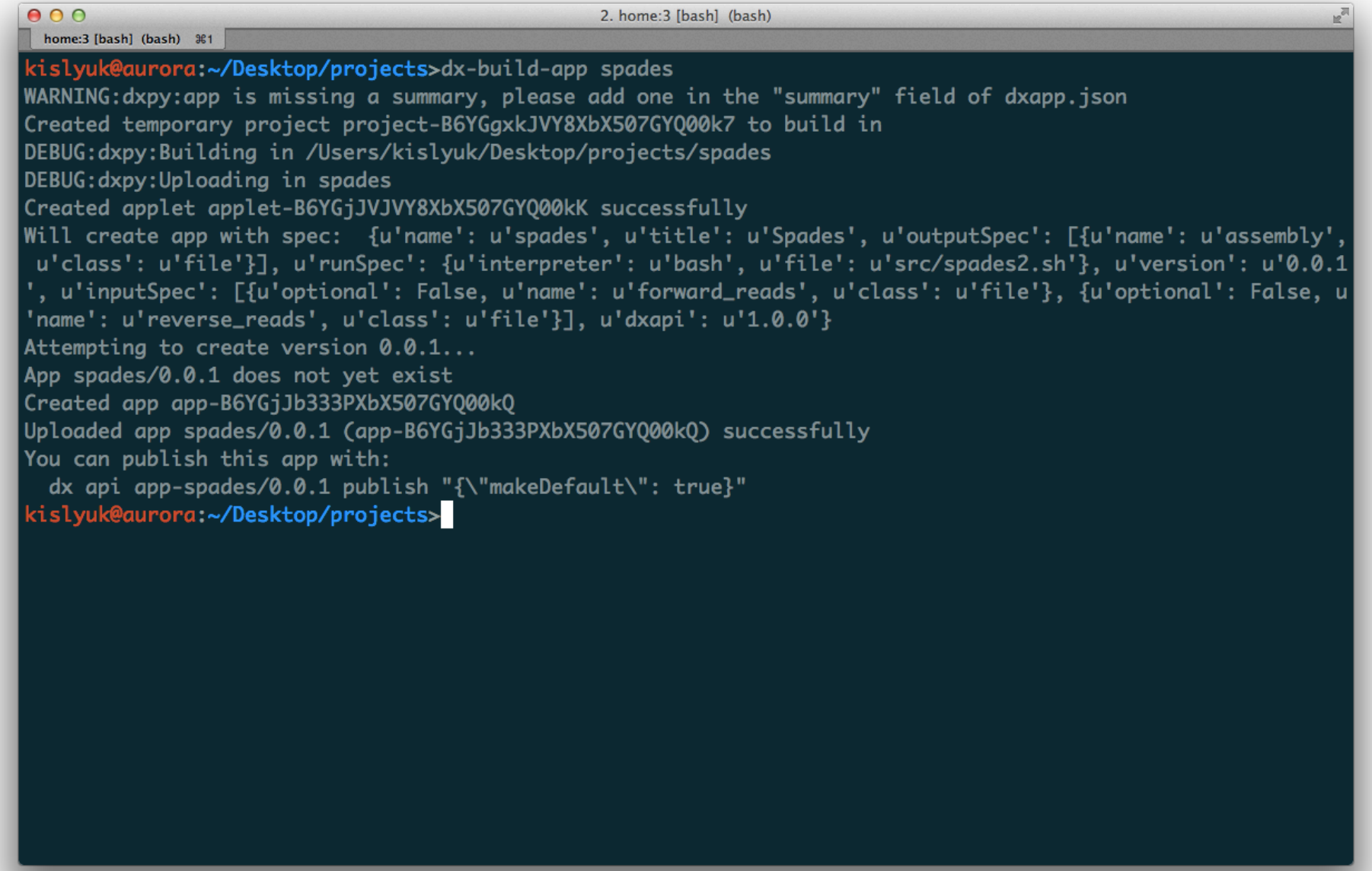
DNAnexus


spades2.sh — spades
OPEN FILES
FOLDERS
spades
resources
src
spades2.sh
test
dxapp.json
Readme.developer.md
Readme.md
spades2.sh
31 # Fill in your application code here.
32 #
33 # To report any recognized errors in the correct format in
34 # $HOME/job_error.json and exit this script, you can use the
35 # dx-jobutil-report-error utility as follows:
36 #
37 # dx-jobutil-report-error "My error message"
38 #
39 # Note however that this entire bash script is executed with -e
40 # when running in the cloud, so any line which returns a nonzero
41 # exit code will prematurely exit the script; if no error was
42 # reported in the job_error.json file, then the failure reason
43 # will be AppInternalError with a generic error message.
44
45 spades.py -1 forward_reads -2 reverse_reads -o assembly
46
47 tar -cf assembly.tar.gz assembly
48
49 # The following line(s) use the dx command-line tool to upload your file
50 # outputs after you have created them on the local file system. It assumes
51 # that you have used the output field name for the filename for each output,
52 # but you can change that behavior to suit your needs. Run "dx upload -h"
53 # to see more options to set metadata.
54
55 assembly=$(dx upload assembly.tar.gz --brief)
56
57 # The following line(s) use the utility dx-jobutil-add-output to format and
58 # add output variables to your job's output as appropriate for the output
59 # class. Run "dx-jobutil-add-output -h" for more information on what it
60 # does.
61
62 dx-jobutil-add-output assembly "$assembly" --class=file
63 }
64
2 lines, 60 characters selected
Spaces: 4
Shell Script (Bash)

DNAnexus

3. Build, test, publish

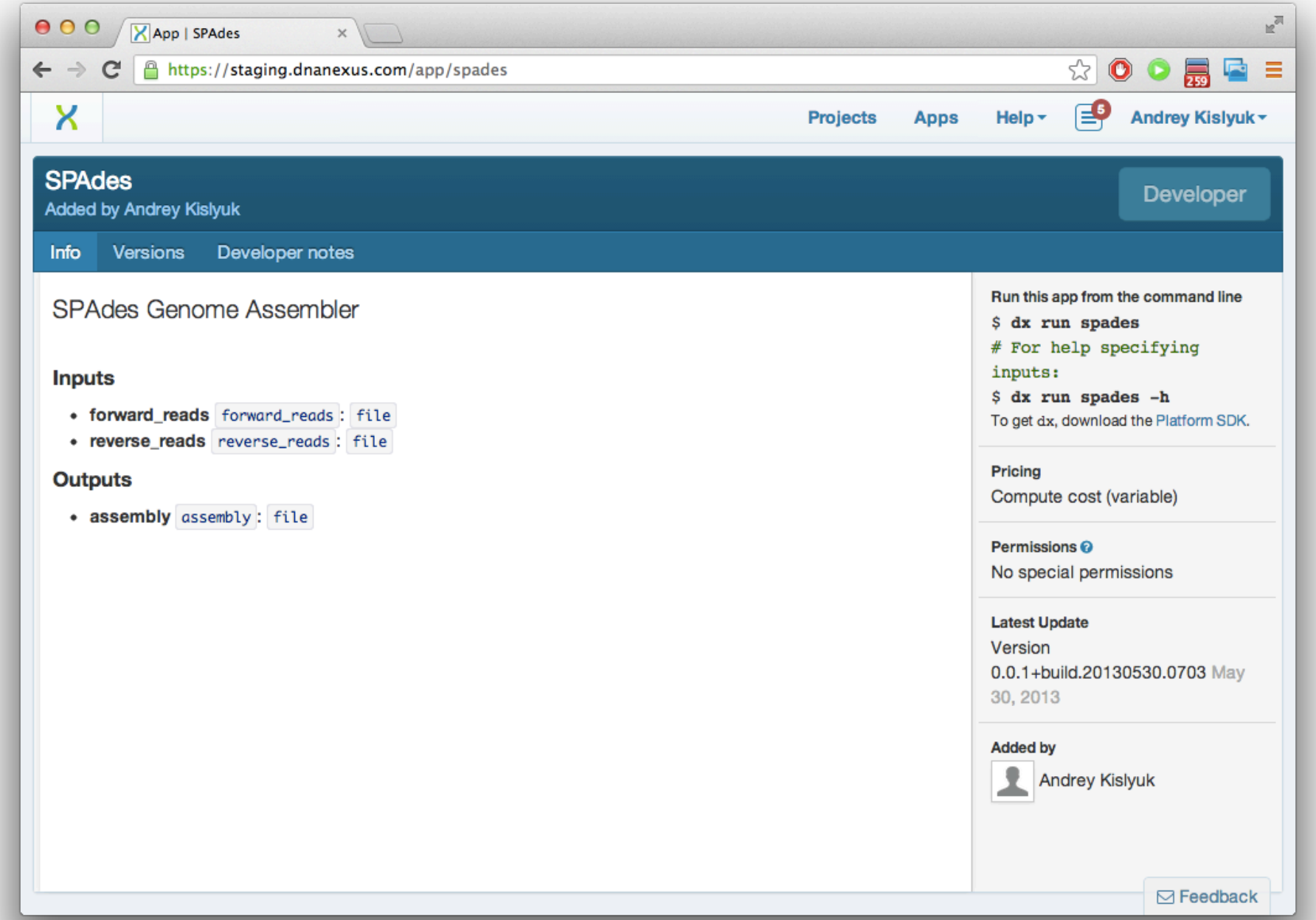
DNAnexus


2. home:3 [bash] (bash)
home:3 [bash] (bash) ⌘1
kislyuk@aurora:~/Desktop/projects>dx-build-app spades
WARNING:dxpy:app is missing a summary, please add one in the "summary" field of dxapp.json
Created temporary project project-B6YGgxkJVY8XbX507GYQ00k7 to build in
DEBUG:dxpy:Building in /Users/kislyuk/Desktop/projects/spades
DEBUG:dxpy:Uploading in spades
Created applet applet-B6YGjJVJVY8XbX507GYQ00kK successfully
Will create app with spec: {u'name': u'spades', u'title': u'Spades', u'outputSpec': [{u'name': u'assembly',
u'class': u'file'}], u'runSpec': {u'interpreter': u'bash', u'file': u'src/spades2.sh'}, u'version': u'0.0.1
', u'inputSpec': [{u'optional': False, u'name': u'forward_reads', u'class': u'file'}, {u'optional': False, u
'name': u'reverse_reads', u'class': u'file'}], u'dxapi': u'1.0.0'}
Attempting to create version 0.0.1...
App spades/0.0.1 does not yet exist
Created app app-B6YGjJb333PXbX507GYQ00kQ
Uploaded app spades/0.0.1 (app-B6YGjJb333PXbX507GYQ00kQ) successfully
You can publish this app with:
dx api app-spades/0.0.1 publish "{\"makeDefault\": true}"
kislyuk@aurora:~/Desktop/projects>

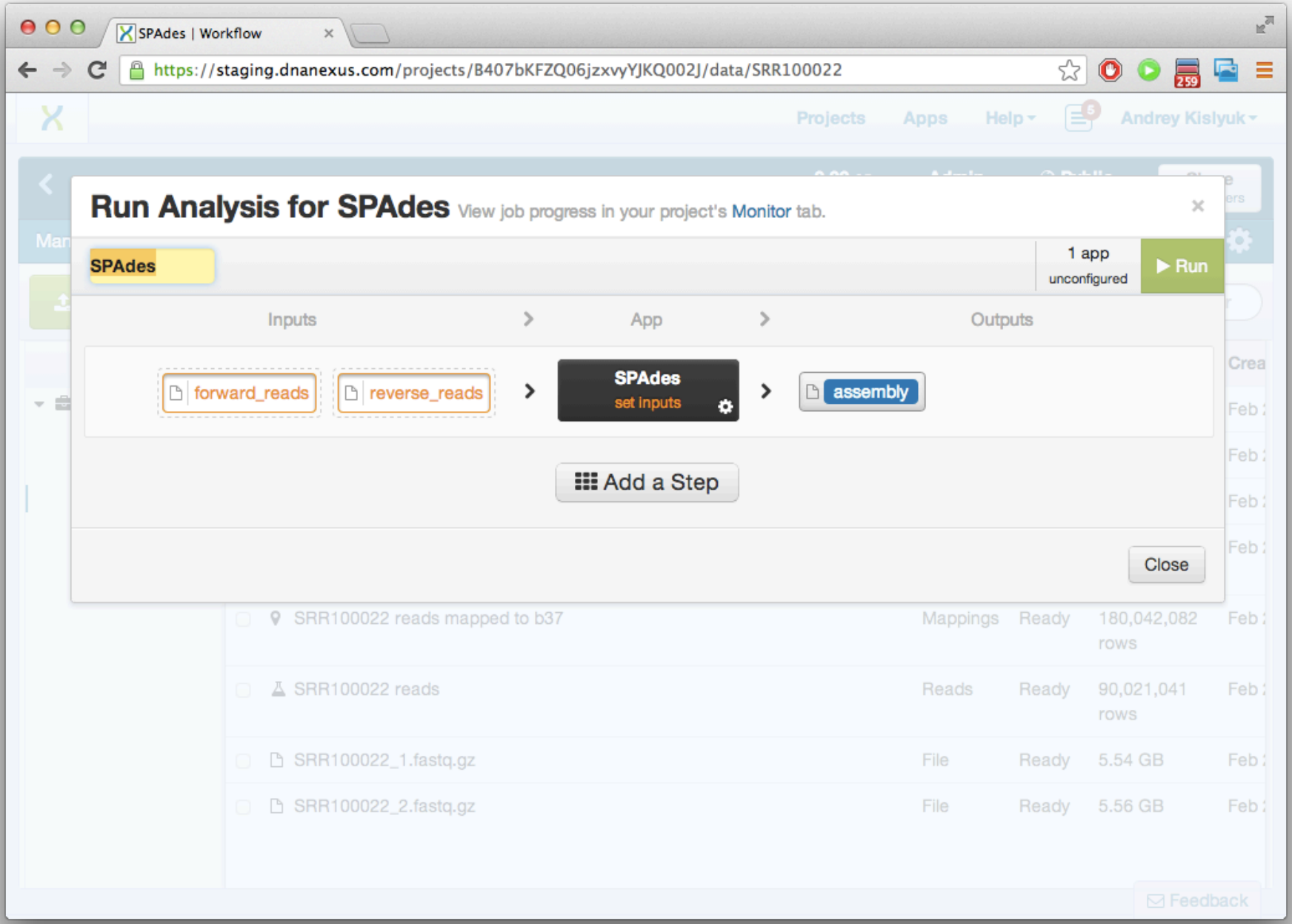
DNAnexus

done

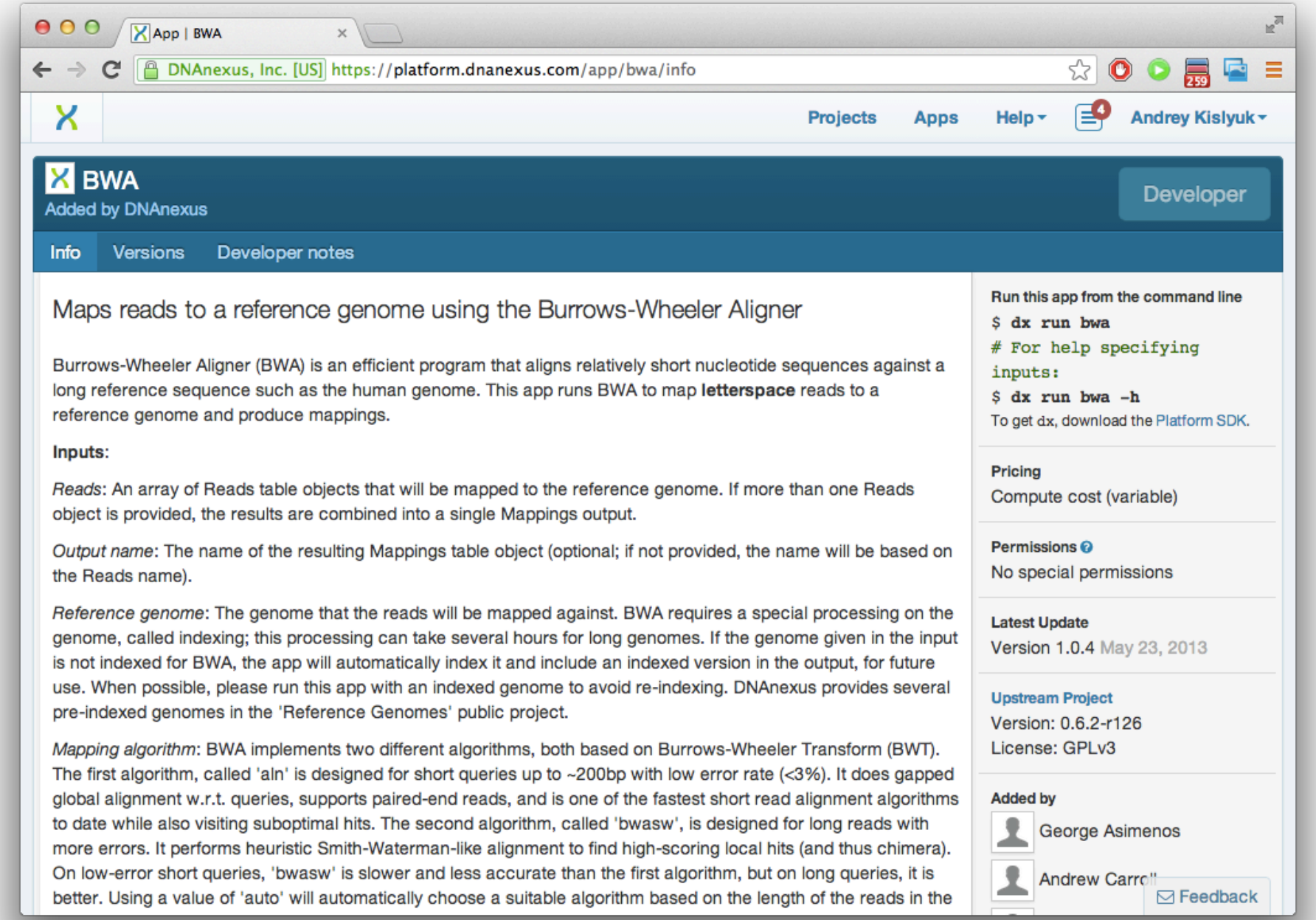
DNAnexus


App | SPAdes
https://staging.dnanexus.com/app/spades
Projects
Apps
Help
5
Andrey Kislyuk
SPAdes
Added by Andrey Kislyuk
Developer
Info
Versions
Developer notes
SPAdes Genome Assembler
Inputs
forward_reads forward_reads : file
reverse_reads reverse_reads : file
Outputs
assembly assembly : file
Run this app from the command line
$ dx run spades
# For help specifying inputs:
$ dx run spades -h
To get dx, download the Platform SDK.
Pricing
Compute cost (variable)
Permissions
No special permissions
Latest Update
Version 0.0.1+build.20130530.0703 May 30, 2013
Added by
Andrey Kislyuk
Feedback

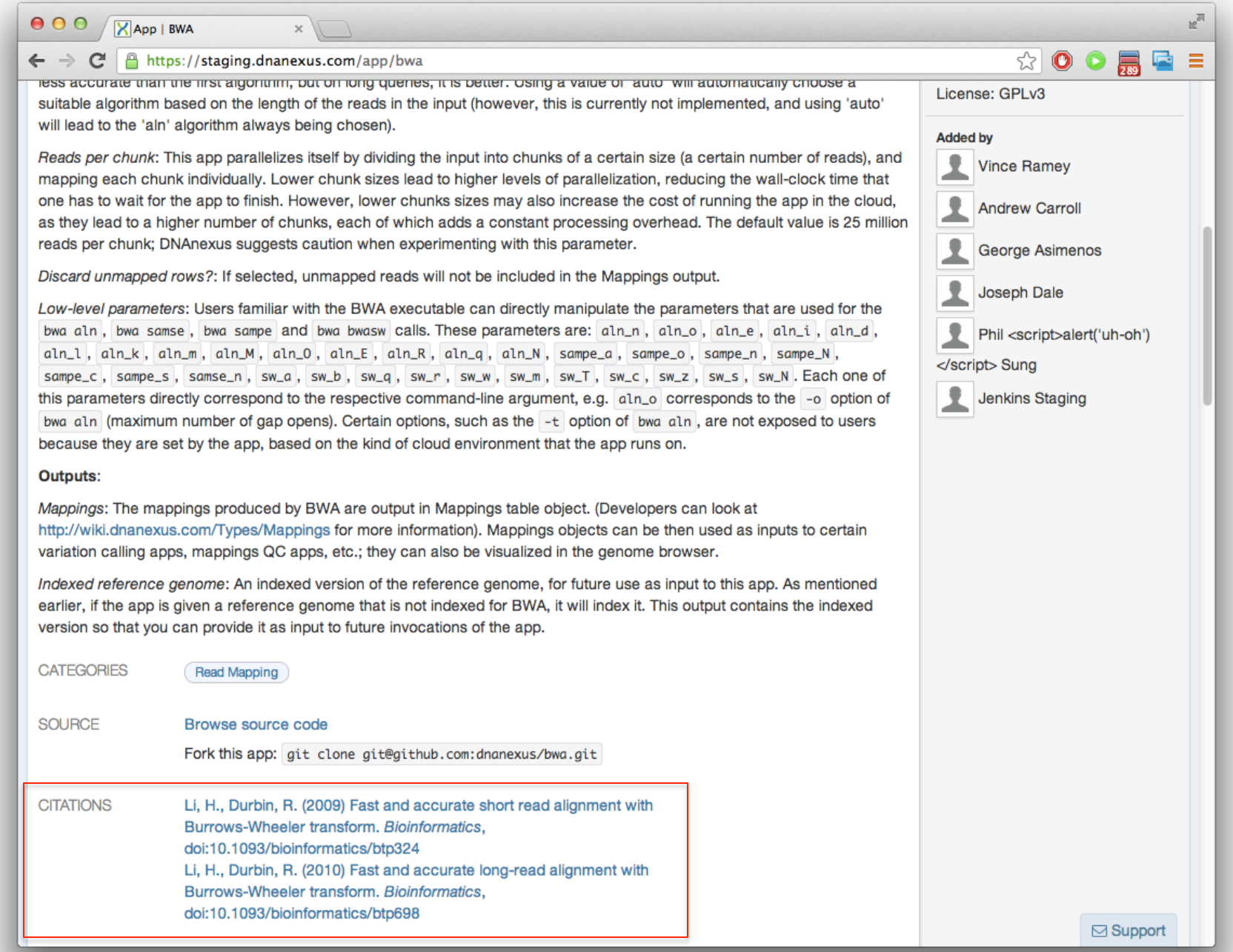
DNAnexus


SPAdes | Workflow
https://staging.dnanexus.com/projects/B407bKFZQ06jzxvyYJKQ002J/data/SRR100022
Projects
Apps
Help
Andrey Kislyuk
Run Analysis for SPAdes
View job progress in your project's Monitor tab.
SPAdes
1 app
unconfigured
Run
Inputs
App
Outputs
forward_reads
reverse_reads
SPAdes
set inputs
assembly
Add a Step
Close
SRR100022 reads mapped to b37
Mappings
Ready
180,042,082 rows
SRR100022 reads
Reads
Ready
90,021,041 rows
SRR100022_1.fastq.gz
File
Ready
5.54 GB
SRR100022_2.fastq.gz
File
Ready
5.56 GB
Feedback

DNAnexus


App | BWA
DNAnexus, Inc. [US] https://platform.dnanexus.com/app/bwa/info
Projects
Apps
Help
Andrey Kislyuk
BWA
Added by DNAnexus
Developer
Info
Versions
Developer notes
Maps reads to a reference genome using the Burrows-Wheeler Aligner
Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. This app runs BWA to map letterspace reads to a reference genome and produce mappings.
Inputs:
Reads: An array of Reads table objects that will be mapped to the reference genome. If more than one Reads object is provided, the results are combined into a single Mappings output.
Output name: The name of the resulting Mappings table object (optional; if not provided, the name will be based on the Reads name).
Reference genome: The genome that the reads will be mapped against. BWA requires a special processing on the genome, called indexing; this processing can take several hours for long genomes. If the genome given in the input is not indexed for BWA, the app will automatically index it and include an indexed version in the output, for future use. When possible, please run this app with an indexed genome to avoid re-indexing. DNAnexus provides several pre-indexed genomes in the 'Reference Genomes' public project.
Mapping algorithm: BWA implements two different algorithms, both based on Burrows-Wheeler Transform (BWT). The first algorithm, called 'aln' is designed for short queries up to ~200bp with low error rate (<3%). It does gapped global alignment w.r.t. queries, supports paired-end reads, and is one of the fastest short read alignment algorithms to date while also visiting suboptimal hits. The second algorithm, called 'bwasw', is designed for long reads with more errors. It performs heuristic Smith-Waterman-like alignment to find high-scoring local hits (and thus chimera). On low-error short queries, 'bwasw' is slower and less accurate than the first algorithm, but on long queries, it is better. Using a value of 'auto' will automatically choose a suitable algorithm based on the length of the reads in the
Run this app from the command line
$ dx run bwa
# For help specifying inputs:
$ dx run bwa -h
To get dx, download the Platform SDK.
Pricing
Compute cost (variable)
Permissions
No special permissions
Latest Update
Version 1.0.4 May 23, 2013
Upstream Project
Version: 0.6.2-r126
License: GPLv3
Added by
George Asimenos
Andrew Carro
Feedback

DNAnexus


App | BWA
https://staging.dnanexus.com/app/bwa
less accurate than the first algorithm, but on long queries, it is better. Using a value of 'auto' will automatically choose a suitable algorithm based on the length of the reads in the input (however, this is currently not implemented, and using 'auto' will lead to the 'aln' algorithm always being chosen).
Reads per chunk: This app parallelizes itself by dividing the input into chunks of a certain size (a certain number of reads), and mapping each chunk individually. Lower chunk sizes lead to higher levels of parallelization, reducing the wall-clock time that one has to wait for the app to finish. However, lower chunks sizes may also increase the cost of running the app in the cloud, as they lead to a higher number of chunks, each of which adds a constant processing overhead. The default value is 25 million reads per chunk; DNAnexus suggests caution when experimenting with this parameter.
Discard unmapped rows?: If selected, unmapped reads will not be included in the Mappings output.
Low-level parameters: Users familiar with the BWA executable can directly manipulate the parameters that are used for the bwa aln, bwa samse, bwa sampe and bwa bwasw calls. These parameters are: aln_n, aln_o, aln_e, aln_i, aln_d, aln_l, aln_k, aln_m, aln_M, aln_O, aln_E, aln_R, aln_q, aln_N, sampe_a, sampe_o, sampe_n, sampe_N, sampe_c, sampe_s, samse_n, sw_a, sw_b, sw_q, sw_r, sw_w, sw_m, sw_T, sw_c, sw_z, sw_s, sw_N. Each one of this parameters directly correspond to the respective command-line argument, e.g. aln_o corresponds to the -o option of bwa aln (maximum number of gap opens). Certain options, such as the -t option of bwa aln, are not exposed to users because they are set by the app, based on the kind of cloud environment that the app runs on.
Outputs:
Mappings: The mappings produced by BWA are output in Mappings table object. (Developers can look at http://wiki.dnanexus.com/Types/Mappings for more information). Mappings objects can be then used as inputs to certain variation calling apps, mappings QC apps, etc.; they can also be visualized in the genome browser.
Indexed reference genome: An indexed version of the reference genome, for future use as input to this app. As mentioned earlier, if the app is given a reference genome that is not indexed for BWA, it will index it. This output contains the indexed version so that you can provide it as input to future invocations of the app.
CATEGORIES
Read Mapping
SOURCE
Browse source code
Fork this app: git clone git@github.com:dnanexus/bwa.git
CITATIONS
Li, H., Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, doi:10.1093/bioinformatics/btp324
Li, H., Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, doi:10.1093/bioinformatics/btp698
License: GPLv3
Added by
Vince Ramey
Andrew Carroll
George Asimenos
Joseph Dale
Phil <script>alert('uh-oh')</script> Sung
Jenkins Staging
Support

DNAnexus

Reproducibility

Ever try to reproduce results from a bioinformatics paper?

How about **CLIA compliance**?

All objects are versioned

Analysis I/O is read-only

Jobs enter into project's *permanent record*

DNAnexus

Publishing

Authors who publish their software

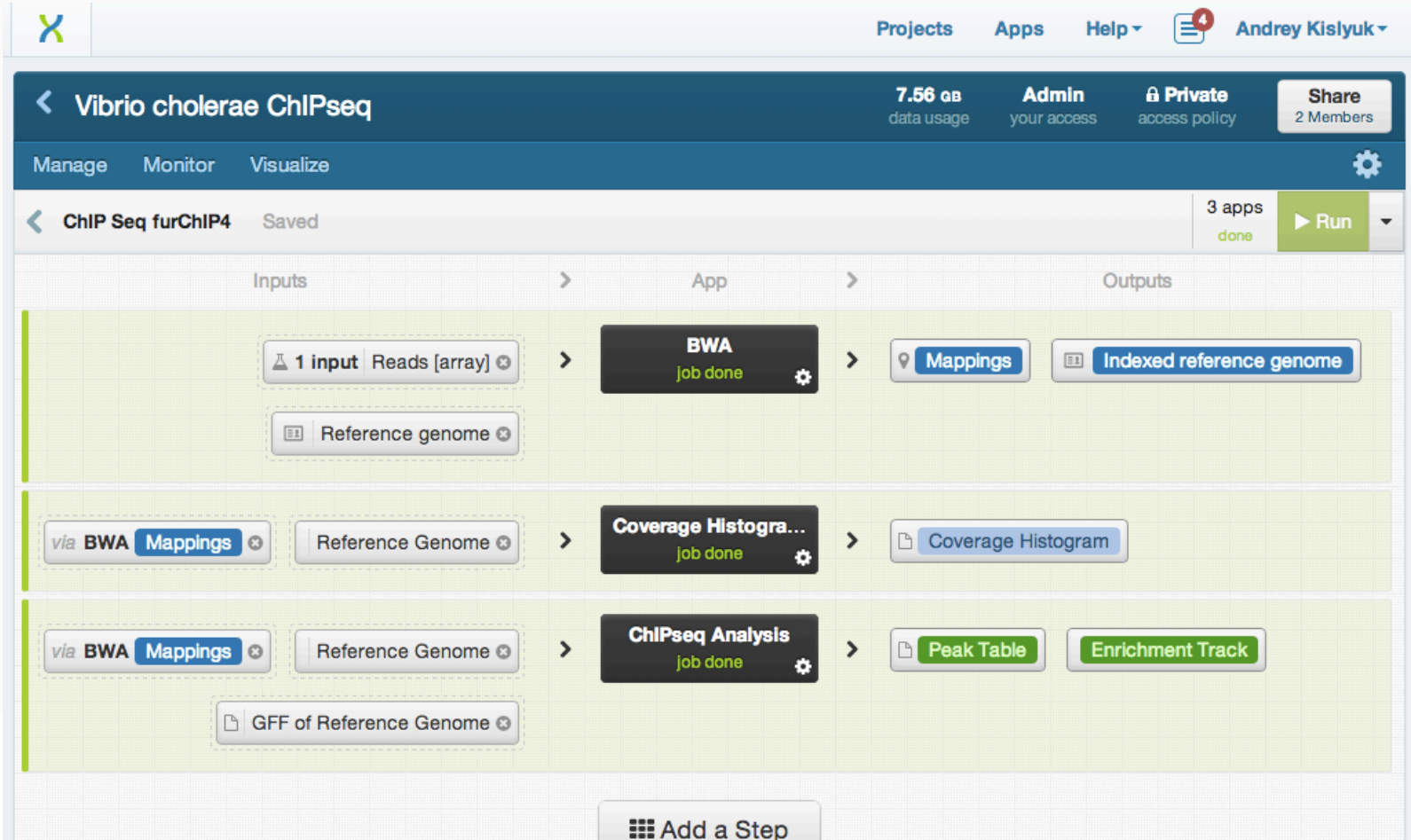
- Don't worry about supporting diverse installs
  - You installed my package on WHAT?
- Leverage all Platform features
  - Accessible UI
- Compose with other apps
  - It's an ecosystem
  - Publish your workflows as apps, too

DNAnexus

Recognition

Users are encouraged to cite app authors

One-click bibliographies coming soon



We care about developers

DNAnexus developer program

- $1000 credit
- App Bounties
- Featured Apps

Vision

DNAnexus roadmap

- AWS Glacier – *cheap data archival*
- Projects as publications

  Supplementary materials!

And of course…

- More apps

DNAnexus roadmap

Always improving…

…And so are our cloud providers

We pass on the savings to you

DNAnexus is the platform for ***publishing*** your algorithms

DNAnexus is the platform for ***delivering*** genomics results to users

Acknowledgments

DNAnexus is
Andreas Sundquist
Arend Sidow
Serafim Batzoglou

Major Investors

Google ventures
TPG BIOTECH
First Round CAPITAL

Thank you

dnanexus.com