

Genomics applications in the cloud with the DNAⁿexus Platform



Andrey Kislyuk

BOSC 2013

The DNAexus Platform

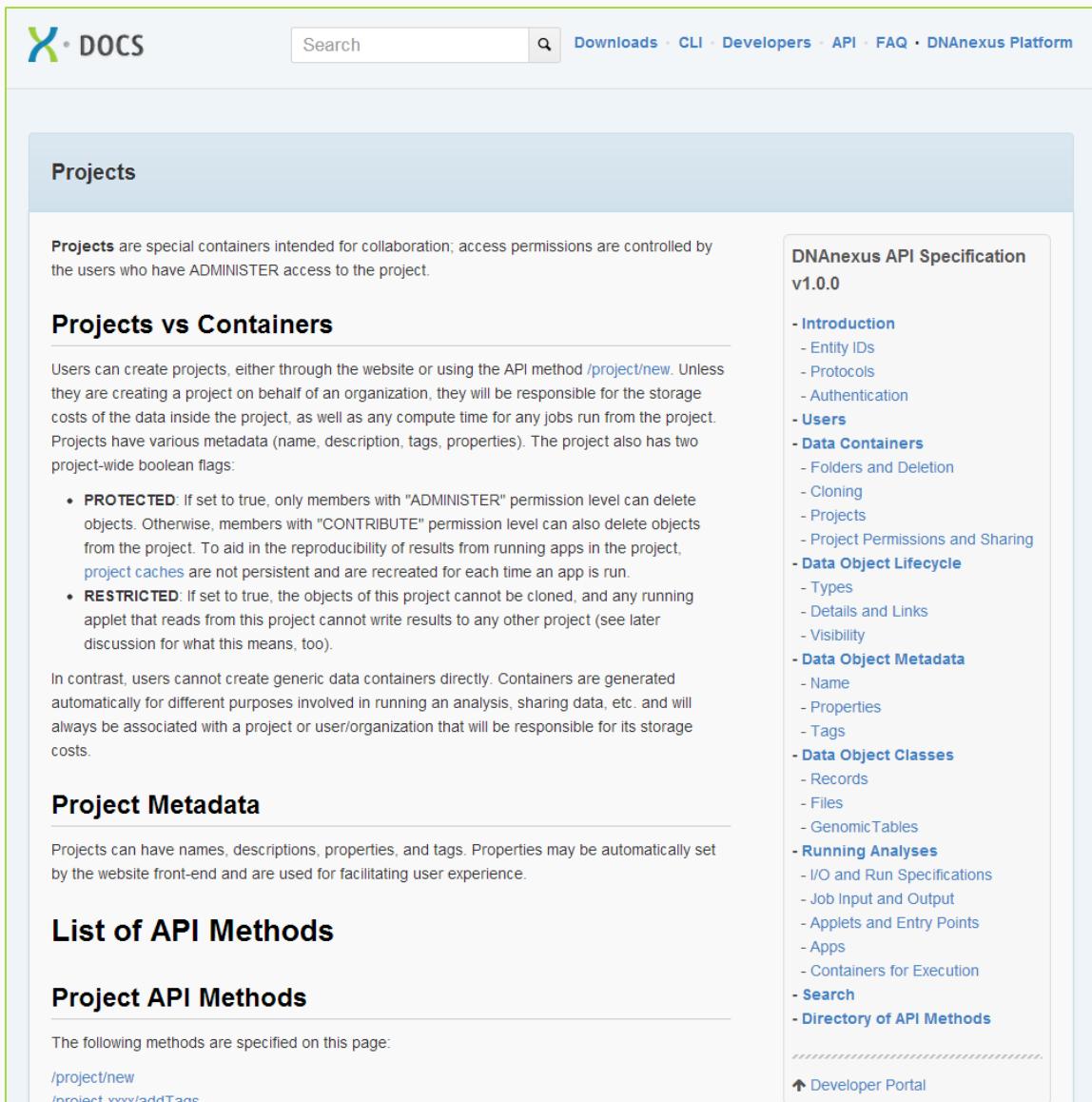
The screenshot displays the DNAexus platform interface. On the left, a sidebar lists various tools and features: All, Alignment, Annotation, Debugging, Export, Import, Reports, RNA-Seq, and ContigSet Validator. The main workspace shows a pipeline workflow starting with the BWA tool, which maps reads to a reference genome. This is followed by a BED Importer tool, which converts a BED file into Spans/Genes. The final step in the main view is a GATK UnifiedGenotyper tool. A circular callout highlights the processing step of the GATK UnifiedGenotyper, showing the command-line script:

```
# Processing
tar zxvf reference.bowtie
if [ "$reads2gz" != "" ]
then
bowtie2 -t -p 2 -x ref
else
bowtie2 -t -p 2 -x ref
fi
samtools view -bS out.bam | samtools sort out.bam
```

Configurable cloud infrastructure for genomics

The DNAnexus Platform

Open and comprehensive API



The screenshot shows the DNAnexus DOCS website with a green header bar. The header includes the 'DOCS' logo, a search bar, and navigation links: Downloads, CLI, Developers, API, FAQ, and DNAnexus Platform.

The main content area has a light blue header titled 'Projects'. Below it, a text block explains that 'Projects' are special containers for collaboration, controlled by users with ADMINISTER access. A section titled 'Projects vs Containers' follows, stating that users can create projects through the website or API, and that they are responsible for storage costs. It also notes that projects have metadata and boolean flags like PROTECTED and RESTRICTED. A note below states that users cannot create generic data containers directly; they are generated automatically for different purposes.

Below this, sections for 'Project Metadata' and 'List of API Methods' are shown. The 'Project API Methods' section lists the '/project/new' and '/project-xxxx/addTags' endpoints. To the right, a sidebar titled 'DNAnexus API Specification v1.0.0' contains a hierarchical table of contents for various API topics.

DNAnexus API Specification v1.0.0

- **Introduction**
 - Entity IDs
 - Protocols
 - Authentication
- **Users**
- **Data Containers**
 - Folders and Deletion
 - Cloning
 - Projects
 - Project Permissions and Sharing
- **Data Object Lifecycle**
 - Types
 - Details and Links
 - Visibility
- **Data Object Metadata**
 - Name
 - Properties
 - Tags
- **Data Object Classes**
 - Records
 - Files
 - GenomicTables
- **Running Analyses**
 - I/O and Run Specifications
 - Job Input and Output
 - Applets and Entry Points
 - Apps
 - Containers for Execution
- **Search**
- **Directory of API Methods**

At the bottom right of the sidebar is a link to the 'Developer Portal'.

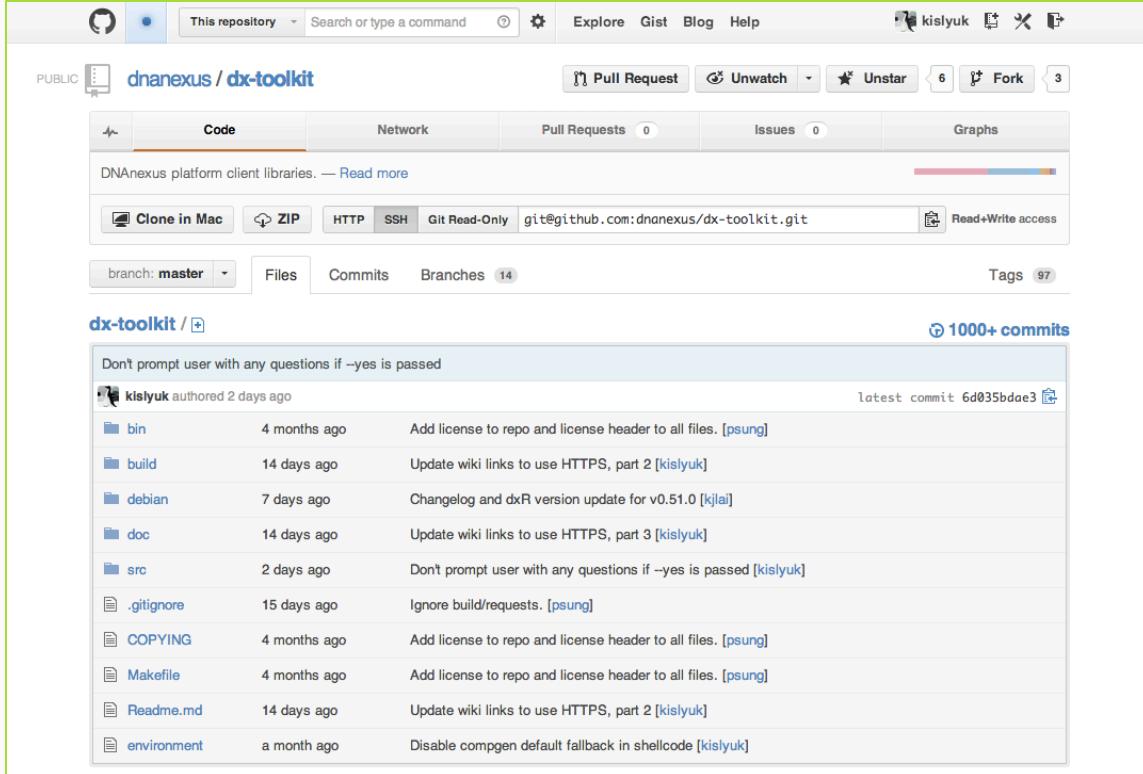
The DNAnexus Platform

Open and
comprehensive
SDK

Linux
OS X
Windows

Upload agents
(Dropbox-like)

DNA
Nexus



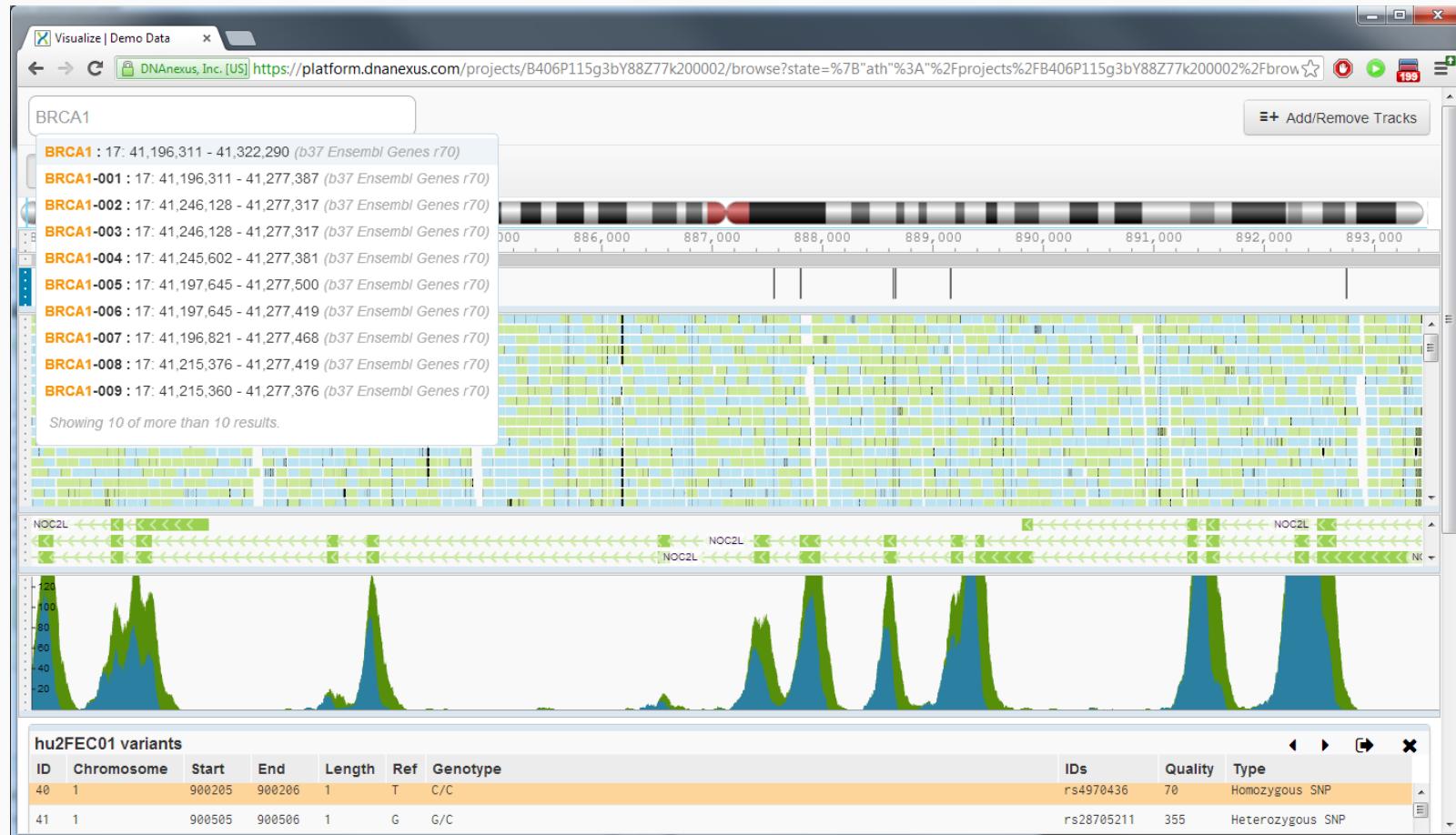
A screenshot of a GitHub repository page for `dnanexus/dx-toolkit`. The repository is public and has over 1000 commits. The code tab is selected, showing a list of recent commits by `kislyuk` and others. The commits are dated from 2 days ago to 4 months ago and involve updates to `bin`, `build`, `debian`, `doc`, `src`, `.gitignore`, `COPYING`, `Makefile`, `Readme.md`, and `environment` files. The `Readme.md` file is also shown, containing the title "DNA^Nexus Platform SDK". Below the file content, there are two bullet points: "To download pre-built packages for your platform, see <https://wiki.dnanexus.com/Downloads>." and "Found a bug? See [Reporting Bugs](#) below."

This repository contains the DNA^Nexus API language bindings and utilities for interacting with the DNA^Nexus platform.

See <https://wiki.dnanexus.com/> and <http://autodoc.dnanexus.com/> for relevant documentation.

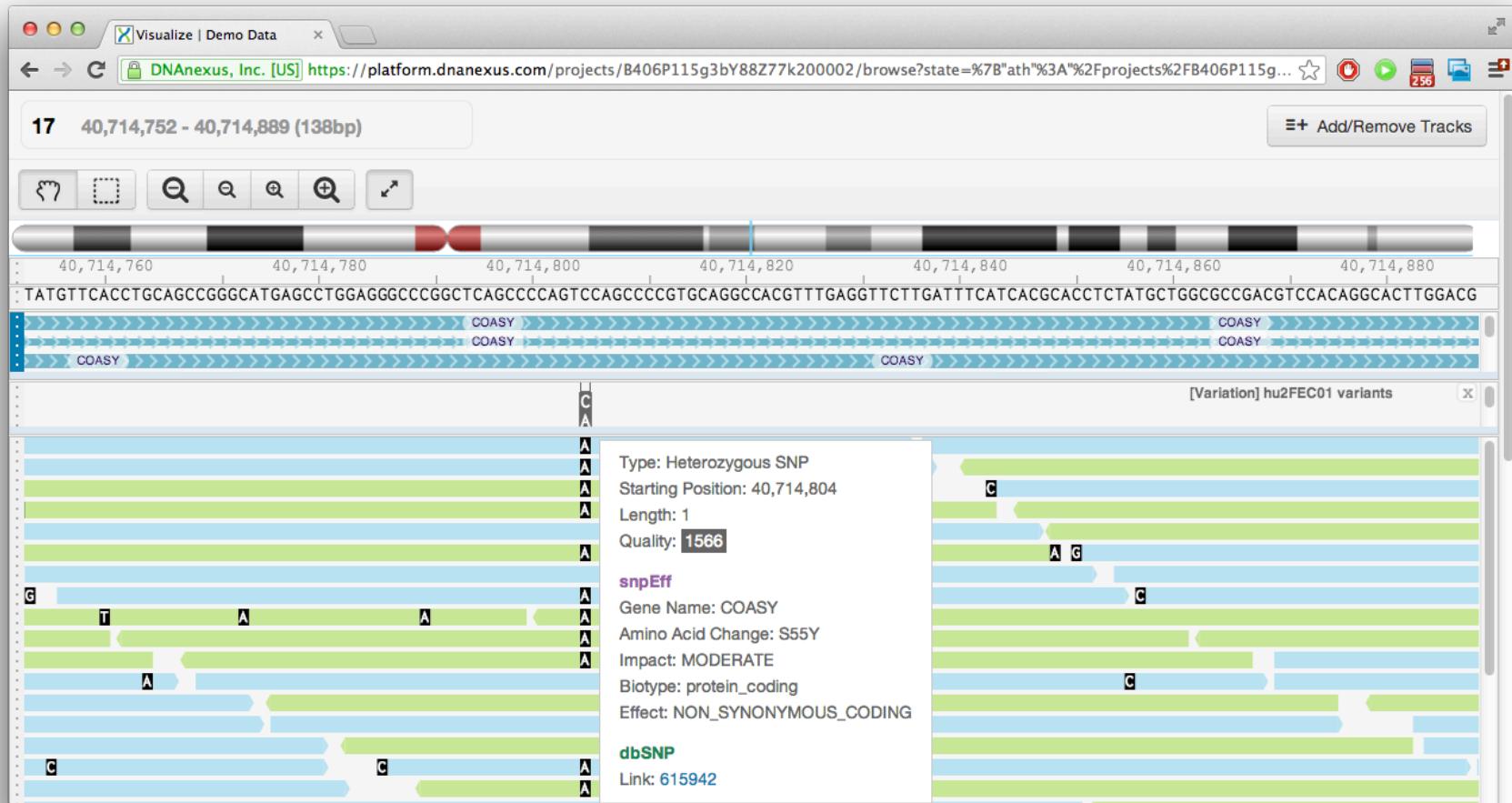
Batteries are included

The DNAexus Platform



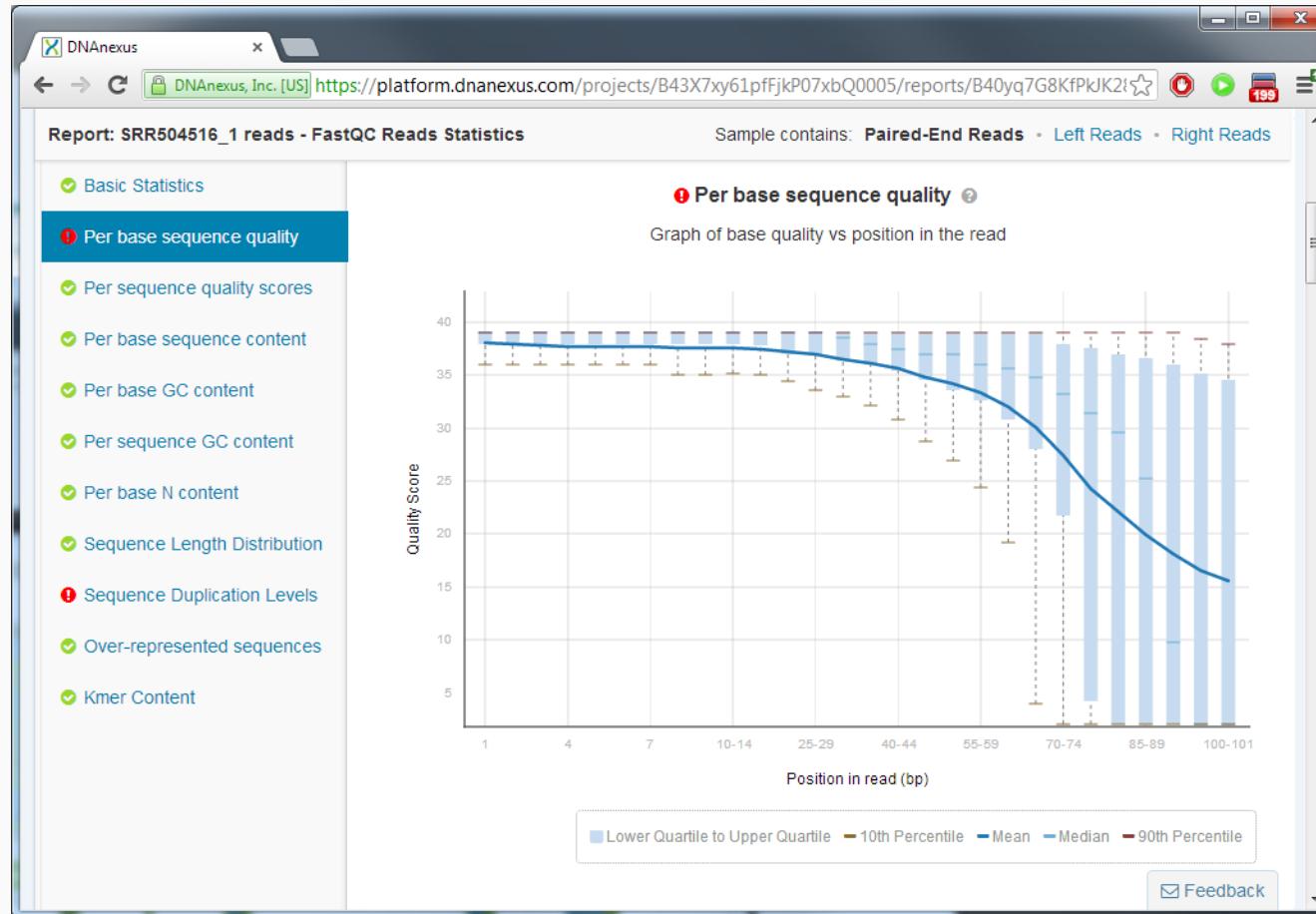
Powerful visualization tools built in

The DNAexus Platform



Powerful visualization tools built in

The DNAexus Platform



Quality control reports

DNAexus

Visualize | *Vibrio cholerae*

DNAnexus, Inc. [US] https://platform.dnanexus.com/projects/B4G58y8jbf41KvV9k0005J/browse

Projects Apps Help Andrei Kislyuk

Vibrio cholerae ChIPseq

7.56 GB Admin Private Share

Select Tracks to Visualize

Reference Genome b37

Special tracks: Reference Ruler

All projects > Demo Data

Filter...

Your visualization engine here (there's an API for that)

23andMe Exome Variants /23andMe Exome Variation Jan 30, 2013 6:30 AM

Developer Quickstart /Developer Quickstart Jan 24, 2013 9:30 AM

ERP001228 (whole-genome) /ERP001228 (whole-genome) Jan 29, 2013 3:30 PM

Illumina BodyMap2 heart tissue RNA-Seq /Illumina BodyMap2 heart tissue RNA-Seq Jan 24, 2013 6:30 AM

SRR100022

SRR100022 mappings to b37 (realigned and recalibrated) /SRR100022 Jan 24, 2013 10:30 AM

SRR100022 variants by GATK (annotated) /SRR100022 Jan 25, 2013 1:30 PM

hu2FEC01 variants /23andMe Exome Variation Jan 7, 2013 12:30 PM

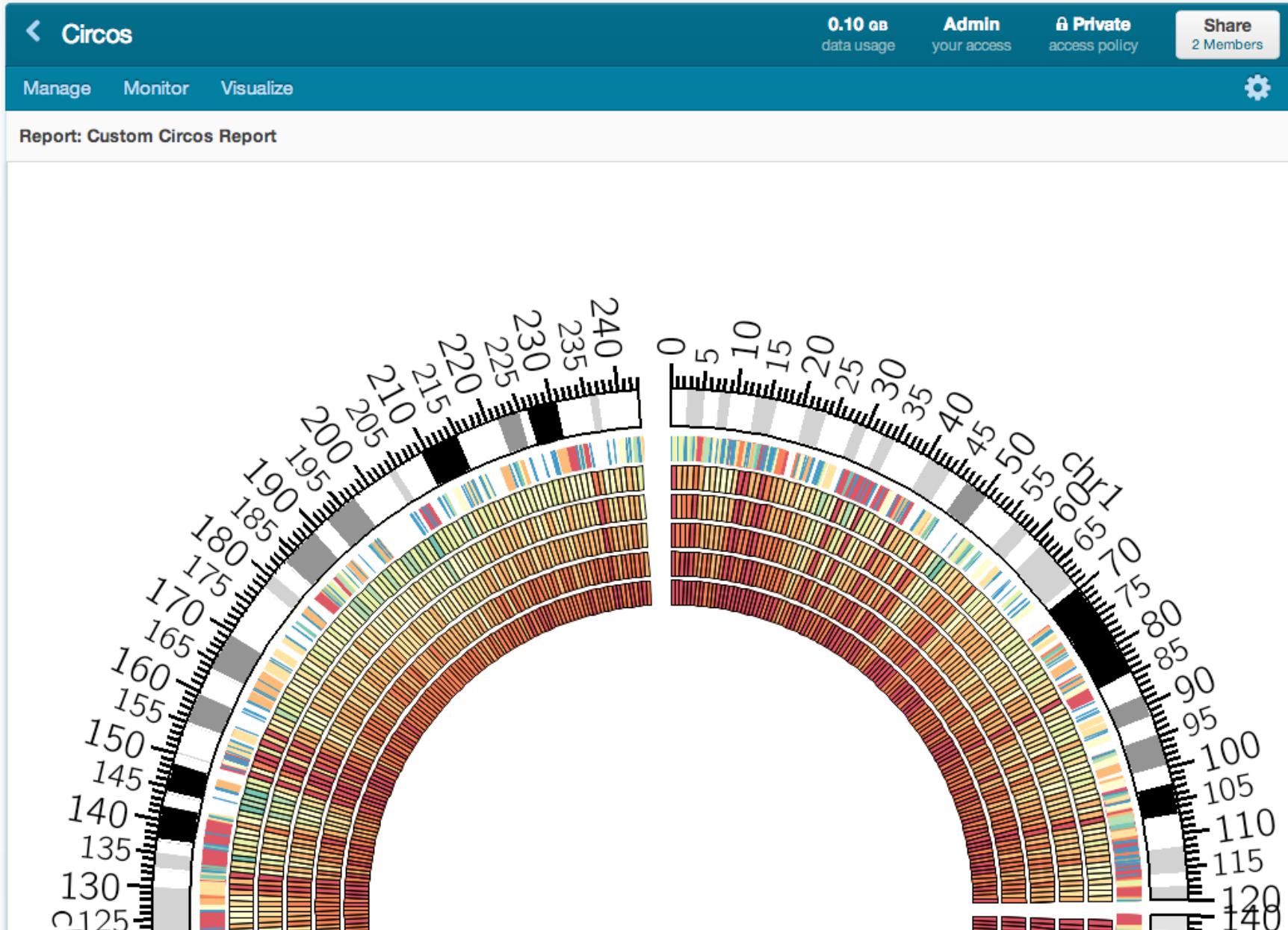
ERP001228 mappings coverage /ERP001228 (whole-genome) Feb 27, 2013 5:30 PM

Ok

DNAnexus Platform · Downloads · Docs · CLI · Developers · API · Answers

Feedback

Yes, we have Circos



DNA**n**exus 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

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

Blazing fast development

Open-source stack

App wizard

walks you through app creation

Learn by example

fork our repos

Collaborate

deploy apps from GitHub

The screenshot shows a GitHub repository page for the DNA Nexus organization. The repository name is 'dnanexus'. The page includes a sidebar with basic metadata fields like name, title, summary, description, input, and output. The main content area lists several projects:

- vcf_importer**: VCF Importer: Converts a VCF file into a Variants GTable object. Last updated 30 minutes ago.
- samtools_mpileup**: SAMtools mpileup: Variant calling for SNPs and Indels using SAMtools mpileup. Last updated 33 minutes ago.
- gtf_importer**: GTF Importer: Converts a GTF File to a Genes object. Last updated 39 minutes ago.
- gff_importer**: GFF Importer: Converts a GFF File to a Spans/Genes Object. Last updated 41 minutes ago.
- dx-toolkit**: DNA Nexus platform client libraries. Last updated 42 minutes ago.
- gatk_variant_annotator**: GATK Variant Annotator: Annotates a Variants GTable object with dbSNP or snpEff, or with overlapping rows of another Variants table. Last updated an hour ago.
- gatk_unifiedgenotyper**: GATK UnifiedGenotyper: Calls SNP and indel variations using GATK's UnifiedGenotyper. Last updated an hour ago.

Debug quickly

```
home:3 [bash] (bash) 361 2. home:3 [bash] (bash)
kislyuk@aurora:~>dx watch job-B6K14F05fyY6XPPfv4g0000P
Watching job job-B6K14F05fyY6XPPfv4g0000P. Press Ctrl+C to stop.
* GATK Pipeline - BWA (re-run) (gatk_pipeline:main) (done) job-B6K14F05fyY6XPPfv4g0000P
  vince 2013-05-21 17:32:44 (runtime 0:02:11)
2013-05-21 17:33:02 GATK Pipeline - BWA (re-run) INFO Logging initialized (priority)
2013-05-21 17:33:02 GATK Pipeline - BWA (re-run) INFO Logging initialized (bulk)
2013-05-21 17:33:09 GATK Pipeline - BWA (re-run) STDOUT Installing apt packages openjdk-6-jre-headless tabix
  pypy dx-toolkit-beta
2013-05-21 17:33:39 GATK Pipeline - BWA (re-run) STDOUT >>> Unpacking resources.tar.gz to /
2013-05-21 17:33:41 GATK Pipeline - BWA (re-run) STDERR python running (job ID job-B6K14F05fyY6XPPfv4g0000P)
2013-05-21 17:33:41 GATK Pipeline - BWA (re-run) STDOUT Recalibrated Table: gtable-B6K14k85fyY9x7fFx13Q00FK
2013-05-21 17:33:42 GATK Pipeline - BWA (re-run) STDERR * Starting dx-contigset-to-fasta...
2013-05-21 17:33:42 GATK Pipeline - BWA (re-run) STDERR - Getting details for ContigSet record-B6Jx5vQ5fyYBz
  8G4X9b00054...
2013-05-21 17:33:42 GATK Pipeline - BWA (re-run) STDERR {"contigs": {"names": ["1"], "offsets": [0], "sizes": [249
  250621]}, "flat_sequence_file": {"$dnanexus_link": "file-B6Jx5Q85fyY9QK07zj10001f"}}
2013-05-21 17:33:42 GATK Pipeline - BWA (re-run) STDERR - Downloading sequence for chromosome 1 (offset = 0,
  length = 249250621)...
2013-05-21 17:33:54 GATK Pipeline - BWA (re-run) STDERR - writing FASTA...
2013-05-21 17:34:21 GATK Pipeline - BWA (re-run) STDERR * Finished dx-contigset-to-fasta.
2013-05-21 17:35:06 GATK Pipeline - BWA (re-run) STDOUT splits: 1
2013-05-21 17:35:06 GATK Pipeline - BWA (re-run) STDOUT [0]
2013-05-21 17:35:06 GATK Pipeline - BWA (re-run) STDOUT -L 1:10000000-11000000
2013-05-21 17:35:06 GATK Pipeline - BWA (re-run) STDOUT [u' -L 1:10000000-11000000']
2013-05-21 17:35:06 GATK Pipeline - BWA (re-run) STDOUT [u' -L 1:10000000-11000000']
* GATK Pipeline - BWA (re-run) (gatk_pipeline:main) (done) job-B6K14F05fyY6XPPfv4g0000P
  vince 2013-05-21 17:32:44 (runtime 0:02:11)
Output: recalibrated_mappings = gtable-B6K14k85fyY9x7fFx13Q00FK
  variants = gtable-B6K1Fbj5fyY0VgyZ8vjQ00XF
kislyuk@aurora:~>
```

The following Types have been defined to facilitate interoperability between applications in the DNAAnexus 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 conventions that are used in the type specifications below.

Foundations for Types

3. python2.7 961

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)

• Report: A specification for how to generate a report that can be viewed in the platform. It may also be one of the following subtypes: [Feedback](#)

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

DNAexus Answers

The screenshot shows the DNAexus Answers Q&A platform. The main content area displays a list of questions:

- How do I connect to an FTP server from a worker? (1 answer, 10 views)
- I can't log-in to the new platform with my classic password. (1 answer, 6 views)
- How do I package a Linux executable into an app? (1 answer, 10 views)
- What's the difference between an app and an applet? (3 answers, 25 views)
- How do I install software requirements for my app? (1 answer, 7 views)
- How do I write my app in my favorite programming language? (1 answer, 17 views)
- What are the default user limits for processes running inside the Linux execution environment? (1 answer, 10 views)
- How do I request more memory/CPU for my app? (1 answer, 9 views)

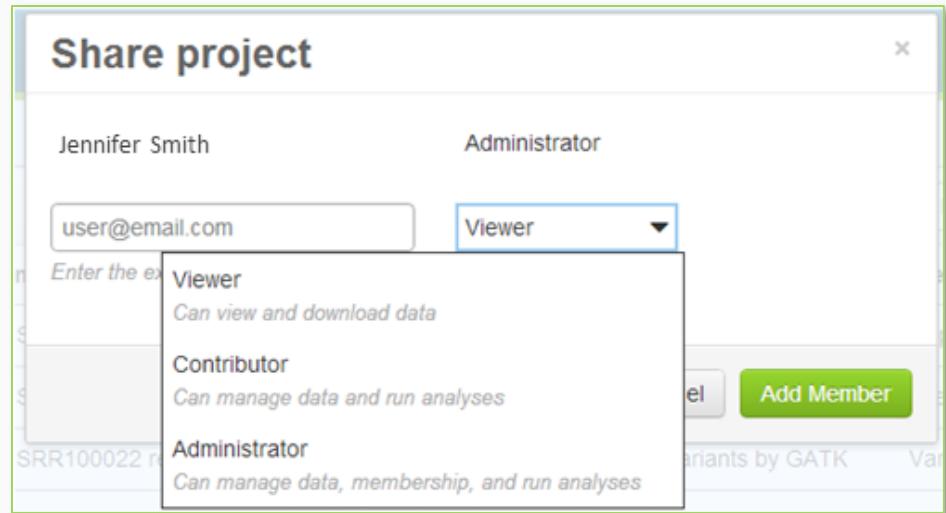
The sidebar on the right includes:

- News**: dx-toolkit 0.48.2 released, dx-toolkit 0.39.1 released [closed]
- Recent Tags**: See All (fastq, bash, array, json, dx-jobutil-parse-link, linux, developers, apps, errors, data-files, data-upload, log-in, sign-up)
- Important Tags**: See All (dx-toolkit, dxr)

A community working together

Instant collaboration

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



Manage | **Vibrio cholerae** X

DNAnexus, Inc. [US] https://platform.dnanexus.com/projects/B4G58y8jbfpf41KvV9k0005J/data

Projects Apps Help Andrey Kis

Vibrio cholerae ChIPSeq 7.56 GB Admin Private Sh

1 of 5 tips · Add Click the Add reads from the

Manage Monitor

Add Data

Vibrio cholerae .logs ChIP Seq ChIP Seq fu... Genome Im... GFF Importer Reads

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

Folder

Folder

Feedback

STA/FASTQ) file or copy

Filter contents of current folder

Size

Workflow | Trio Exome

DNAexus, Inc. [US] https://platform.dnanexus.com/projects/B5Xv0vx96fk1vP8jj0V0008X/workflows/B... 254

Projects Apps Help Andrey Kislyuk 4

Andrey's demos 39.20 GB Admin Private Share 1 Member

Manage Monitor Visualize Run

Trio Exome Saved 2 apps configured

Inputs App Outputs

1 input Reads [array] > BWA configured Mappings > 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?

1. Run App Wizard

```
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 :

app spec

The screenshot shows a Mac OS X application window with a dark theme. The title bar reads "dxapp.json — spades". The left sidebar lists "OPEN FILES" and "FOLDERS". Under "OPEN FILES", "dxapp.json" is listed. Under "FOLDERS", "spades" is expanded, showing "resources", "src" (containing "spades2.sh"), and "test". The main pane displays the JSON code for "dxapp.json".

```
1 {  
2   "name": "spades2",  
3   "title": "Spades",  
4   "dxapi": "1.0.0",  
5   "version": "0.0.1",  
6   "inputSpec": [  
7     {  
8       "name": "forward_reads",  
9       "class": "file",  
10      "optional": false  
11    },  
12    {  
13      "name": "reverse_reads",  
14      "class": "file",  
15      "optional": false  
16    }  
17  ],  
18  "outputSpec": [  
19    {  
20      "name": "assembly",  
21      "class": "file"  
22    }  
23  ],  
24  "runSpec": {  
25    "interpreter": "bash",  
26    "file": "src/spades2.sh"  
27  }  
28}  
29
```

Line 1, Column 1 Spaces: 2 JSON

2. Add entry point code

spades2.sh — spades

```
1 #!/bin/bash
2 # spades2 0.0.1
3 # Generated by dx-app-wizard.
4 #
5 # Basic execution pattern: Your app will run on a single machine from
6 # beginning to end.
7 #
8 # Your job's input variables (if any) will be loaded as environment
9 # variables before this script runs. Any array inputs will be loaded
10 # as bash arrays.
11 #
12 # Any code outside of main() (or any entry point you may add) is
13 # ALWAYS executed, followed by running the entry point itself.
14 #
15 # See https://wiki.dnanexus.com/Developer-Portal for tutorials on how
16 # to modify this file.
17
18 main() {
19
20     echo "Value of forward_reads: '$forward_reads'"
21     echo "Value of reverse_reads: '$reverse_reads'"
22
23     # The following line(s) use the dx command-line tool to download your file
24     # inputs to the local file system using variable names for the filenames. To
25     # recover the original filenames, you can use the output of "dx describe
26     # "$variable" --name".
27
28     dx download "$forward_reads" -o forward_reads
29     dx download "$reverse_reads" -o reverse_reads
30
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
```

Line 1, Column 1

Spaces: 4

Shell Script (Bash)

```
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)

3. Build, test, publish

2. home:3 [bash] (bash)

```
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-B6YGjJVJFY8XbX507GYQ00kK 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>
```

done

App | SPAdes

https://staging.dnanexus.com/app/spades

Projects Apps Help  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

SPAdes | Workflow x

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

	Mappings	Ready	180,042,082	Feb :
SRR100022 reads mapped to b37			rows	
SRR100022 reads	Reads	Ready	90,021,041	Feb :
SRR100022_1.fastq.gz	File	Ready	5.54 GB	Feb :
SRR100022_2.fastq.gz	File	Ready	5.56 GB	Feb :

Feedback

The screenshot shows a DNA Nexus workflow interface for running SPAdes. The main window title is "SPAdes | Workflow". The URL in the address bar is "https://staging.dnanexus.com/projects/B407bKFZQ06jzxvyYJKQ002J/data/SRR100022". The top navigation bar includes "Projects", "Apps", "Help", and a user account section for "Andrey Kislyuk". A central modal dialog titled "Run Analysis for SPAdes" displays the workflow steps: "forward_reads" and "reverse_reads" as inputs, followed by the "SPAdes set inputs" step, which has produced the output "assembly". Below the modal, a table lists four data files: "SRR100022 reads mapped to b37", "SRR100022 reads", "SRR100022_1.fastq.gz", and "SRR100022_2.fastq.gz", each with their respective file type, status, size, and timestamp. A "Feedback" button is located at the bottom right of the modal.

App | BWA

DNAexus, Inc. [US] https://platform.dnanexus.com/app/bwa/info

Projects Apps Help 4 Andrey Kislyuk

BWA
Added by DNAexus

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. DNAexus 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 'alm' 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 Carroll

Feedback

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 chunk 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_0`, `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 these 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

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*

Publishing

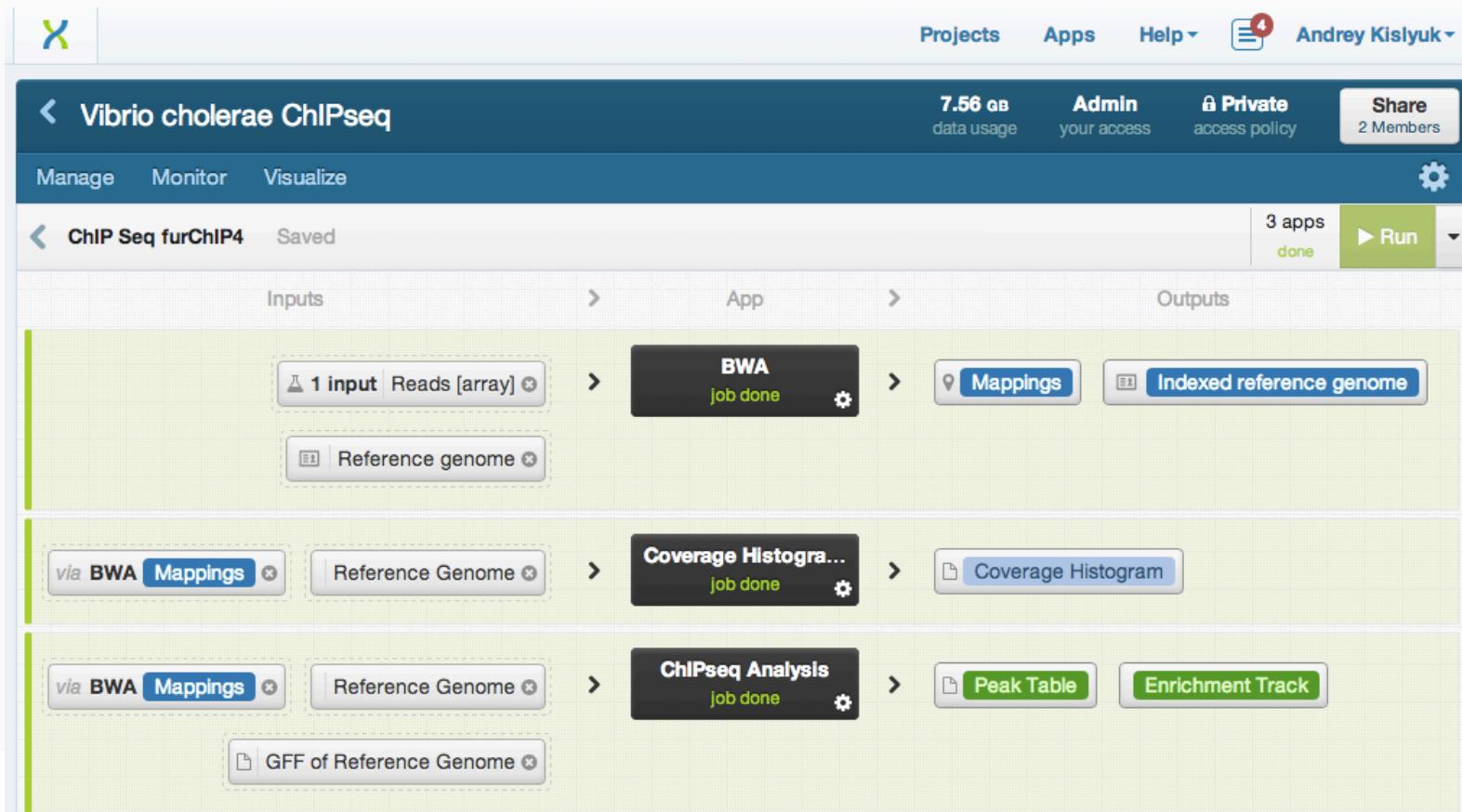
Authors who publish their software

- 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

Recognition

Users are encouraged to cite app authors

One-click bibliographies coming soon



We care about
developers

DNAexus developer program

- \$1000 credit
- App Bounties
- Featured Apps

Vision

DNAexus roadmap

- AWS Glacier – *cheap data archival*
- Projects as publications
Supplementary materials!

And of course...

- More apps

DNAexus roadmap

Always improving...

...And so are our cloud providers

We pass on the savings to you

DNAexus is the platform for
publishing your algorithms

DNAexus is the platform for
delivering genomics results to users

Acknowledgments

*DNA*nexus is
Andreas Sundquist
Arend Sidow
Serafim Batzoglou

Major Investors



Thank you

dnanexus.com