A Cloud-based Data Analysis
Resource for all CCR Researchers

Peter FitzGerald

Head Genome Analysis Unit
Director of BTEP program

Today Agenda

- Introduction to DNAnexus
- Introduction CCR's Pilot program with DNAnexus
- Highlight and Demo of resources available within DNAnexus
- Highlight of CCR support resources
- Follow on classes:
  - Hands on using DNAnexus - Biologists
    Thursday April 11th, 10:00-11:30 am. - NIH Bldg B37, Rm 4041/4107
  - DNAnexus Development Environment - Bioinformaticists
    Friday April 12th, 10:00-11:30 am. - NIH Bldg 37, Rm 2041/2107

What is DNAnexus ?

DNAnexus is a bioinformatics company that provides a cloud-based data analysis and management platform for DNA sequence data. It was founded in early 2009 as a spin-off from Stanford University

"**DNAnexus** provides a **cloud-based platform** optimized to address the challenges of security, scalability, and collaboration, for organizations that are pursuing genomic-based approaches to health, in the clinic and in the research lab."

DNAnexus provides a simplified, structured and managed access two cloud-based service providers (AWS and Azure).

AWS = Amazon Web Services
Azure = Microsoft Cloud Services

Each environment virtually identical - BUT they are distinct spaces and difficult to move data and apps between the two

What is DNAnexus ?

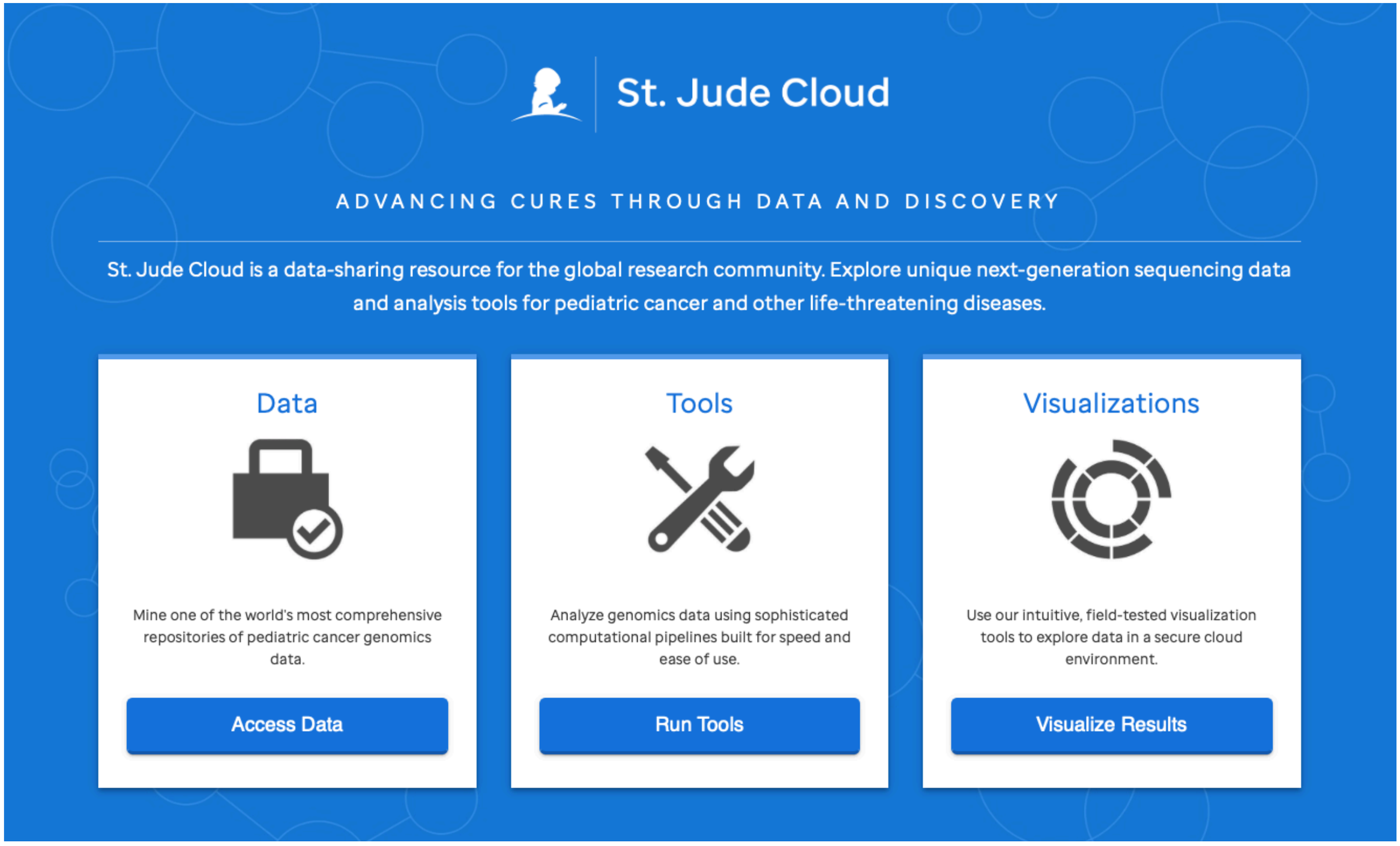
Security has been a major focus during the development of the **DNAnexus** cloud-based platform and this is reflected in it being the only platform of its type that is **Fedramp moderate** certified. Additional, it had the following certifications: ISO 27001 compliance certification supporting **HIPAA**, CAP, **CLIA**, and a Quality Management System supporting GLP and 21 C.F.R. Parts 11, 820, and other relevant regulatory requirements for data stewardship. DNAnexus has a FedRAMP Agency Authority to Operate (ATO) of FISMA-low categorization, with FedRAMP authorization since 10/15/2018

Corporate Web Site - *https://www.dnanexus.com*
Web Interface to Research Tools - *https://platform.dnanexus.com/login*
Documentation - *https://wiki.dnanexus.com/Home*

Why DNAnexus ?

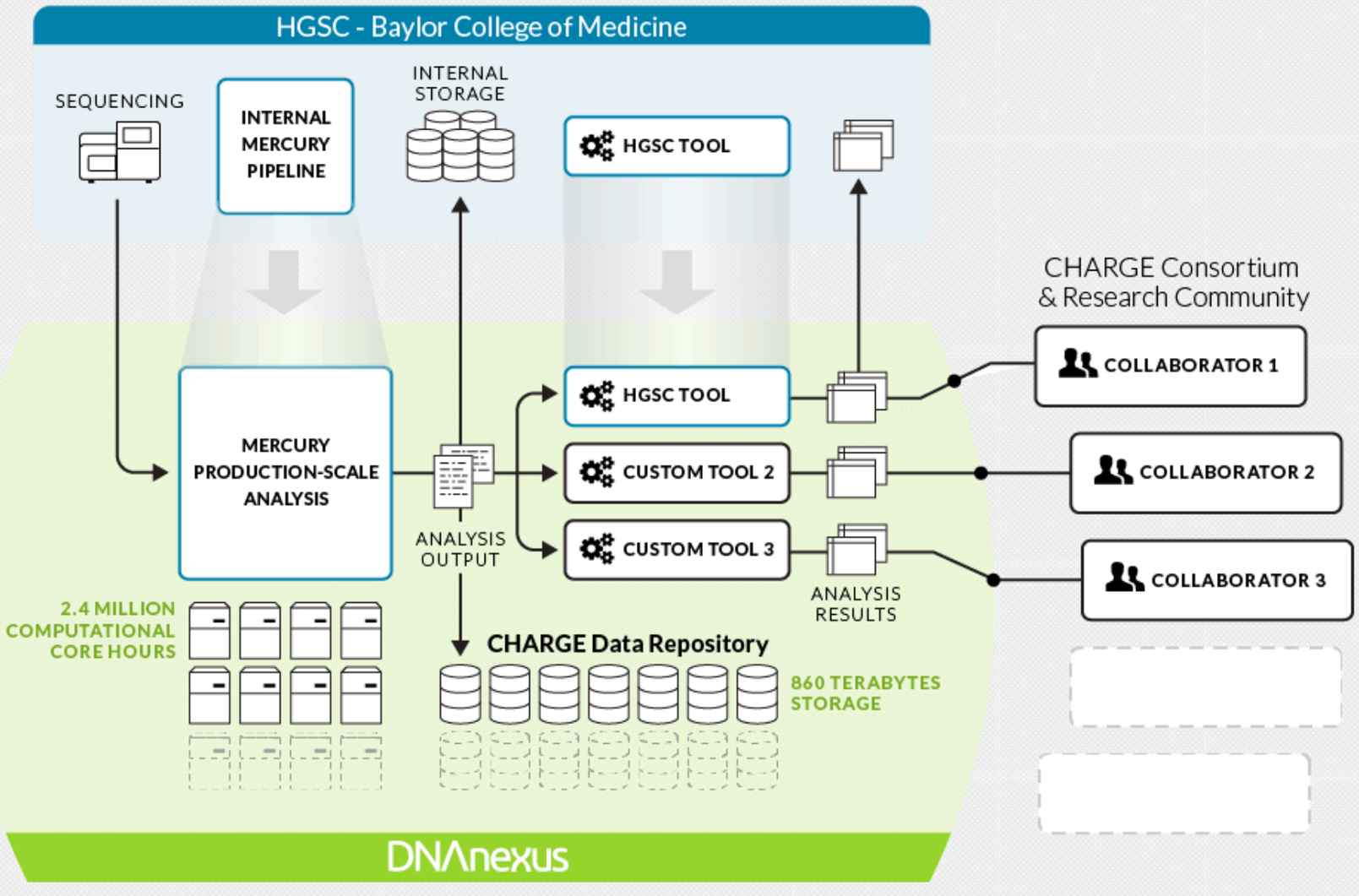
- DNAnexus provides ready access to a variety of prebuilt tools and workflows for the analysis of Genomic Data, within a “User friendly” Web interface.
- The platform is cloud based which means there is no hardware to buy or maintain.
- Their business model, based on resource usage, means that we only pay for what we use.
- Their environment provides an easy method for managing access, managing costs and sharing data internally and externally.
- Command line tools and APIs make the platform accessible from outside the Web interface - ideal for high volume throughput
- A robust development environment.
- Publishing applications within the DNAnexus environment provides a way to make tools available to a wide audience.
- Initial account set up is free

Why DNAnexus ?



CHARGE Project Use Case

Baylor College of Medicine, Amazon Web Services, and DNAnexus team up to run the largest ever cloud-based analysis of genomic data from over 14,000 patients



The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium aims to advance our understanding of human genetics and how it contributes to heart disease and aging. The discoveries that CHARGE makes will be instrumental in understanding disease and aging in mechanistic detail, enabling the development of new medical interventions and analysis tools.

PrecisionFDA is an online, cloud-based, virtual research space where members of the genomics community can experiment, share data and tools, collaborate, and define standards for evaluating and validating analytical pipelines. This open-source community platform, which has become a global reference standard for variant comparison, includes members from academia, industry, healthcare, and government, all working together to further innovation and develop regulatory standards for NGS-based drugs and devices. Launched in December 2015, the precisionFDA community includes nearly 5,000 users across 1,200 organizations, with more than 38 terabytes of genomic data stored.

Demo

- Finding an Application
- Simple Application - FASTQC
- Building a Workflow
- Running some Applications - RNA-SEQ, Custom Applet

https://platform.dnanexus.com/

What to Highlight in Demo

FASTQC

- Simple one step application
- How to find app
- Running parameters
- Monitor
- Buttons

RNA-Seq

- Aligner
- Matrix combiner
- Workflow
- Custom output
- Hand off to other environment
- Build workflow

Pausing Peak Finder

- Aligner
- Peak locator
- Custom output
- File Attributes
- Build workflow
- Reanalysis Modular

IGV Session Maker

- Simple App
- Queued delay
- Local viewing of files

CCR DNAnexus Pilot

CCR-ORG

- We have established an Organizational account
- 60% discount on standard rates
- Initial costs subsidized and managed by OSTR
- Support for use and customized development

Questions we hope the Pilot will answer

- Will this resource be adopted by:
  biologists for data analysis and/or
  bioinformaticists for batch analysis and sharing results?
- Will it work for managing and sharing data on a large scale?
- Is the platform effective in disseminating software solutions?
- Is it a solution for patient data analysis (security, speed)?

DNAnexus Costs

Pay-as-you-Go

Compute Costs - Running Jobs

- Rate varies by type of node
- Rates computer per second of allocated resource

Storage Costs - Data/Files

- Fixed rate for standard storage
- Decrease rate for archival storage
- Rates computer per second of allocated resource

Ingress/Egress - Upload/Download

- Ingress (Upload) = no costs
- Egress (Download) = $

Costs, you Say!

An unfortunate reality is that people give little thought to spending $1,000s on NGS sequencing but expect to get the analysis for free; both in terms of compute cycles and manpower.

CCR spends $1,000,000 on Biowulf hardware every year and $100,000s on various commercial NGS sequencing analysis software.

Let’s not even talk about man/woman power costs!

The good news is that the current PILOT program means that you do not have to pay these costs.

While costs of applying Whole Genome Sequencing in a research setting have decreased over time, costs of subsequent **bioinformatics analysis** necessary to interpret sequence data remain substantial, a phenomenon commonly referred to as “the **$1000 genome** and the $100,000 **analysis**” (Mardis 2010; Caulfield et al. 2013; Wetterstrand 2016).

DNAnexus Costs

Practical Terms

Current expenses (during this pilot) absorbed by OSTR (within limits)

Cost dependent on exact compute-time used,
but the following are ball-park estimates

- **RNA-Seq - $1-10/sample**
- **ChIP-Seq - $1-5/sample**
- **Whole-Exome/Whole Genome (see below)**

Sample Type	FASTQ	BAM	VCF	Price/Sample
Panel (fastq.gz ≤ 5 GiB)	----	----	--->	$3.3
WES (5 GiB < fastq.gz ≤ 20 GiB)	----	----	--->	$9.9
WGS (20 GiB < fastq.gz ≤ 90 GiB)	----	----	--->	$47
WGS2 (90 GiB < fastq.gz ≤ 200 GiB)	----	----	--->	$70.00 to ~$100.00

DNAnexus Pilot

OSTR/GAU has established an ORGNAIZATIONAL account
Billing and management is handled centrally
Current expenses (during this pilot) absorbed by OSTR (within limits)

How this works in practical terms

- **Sign Up for Free Account at DNAnexus**
  - Provides access to the platform and starter funds (~ $50)

- **Apply to Join CCR Organizational Account**
  (*https://gau.ccr.cancer.gov/dnanexus*)
  - Provides addition fund for computing
  - Provides access to Shared Resources
  - Provides access to Support Resources

DNAnexus Account Sign up Page
https://gau.ccr.cancer.gov/dnanexus

DNAnexus - CCR Account Application Review Page

Instructions:

Feel free to edit your information and resubmit:

This form is to request access to the CCR DNAnexus organizational account - being part of this account will enable you to easily access CCR developed/highlighted resources (applications, data, and workflows) as well as pay for use costs via the CCR centrally managed account. See (here) for more details about the CCR DNAnexus organizational account.

First Name *

Last Name *

Email *

PI First Name *

PI Last Name *

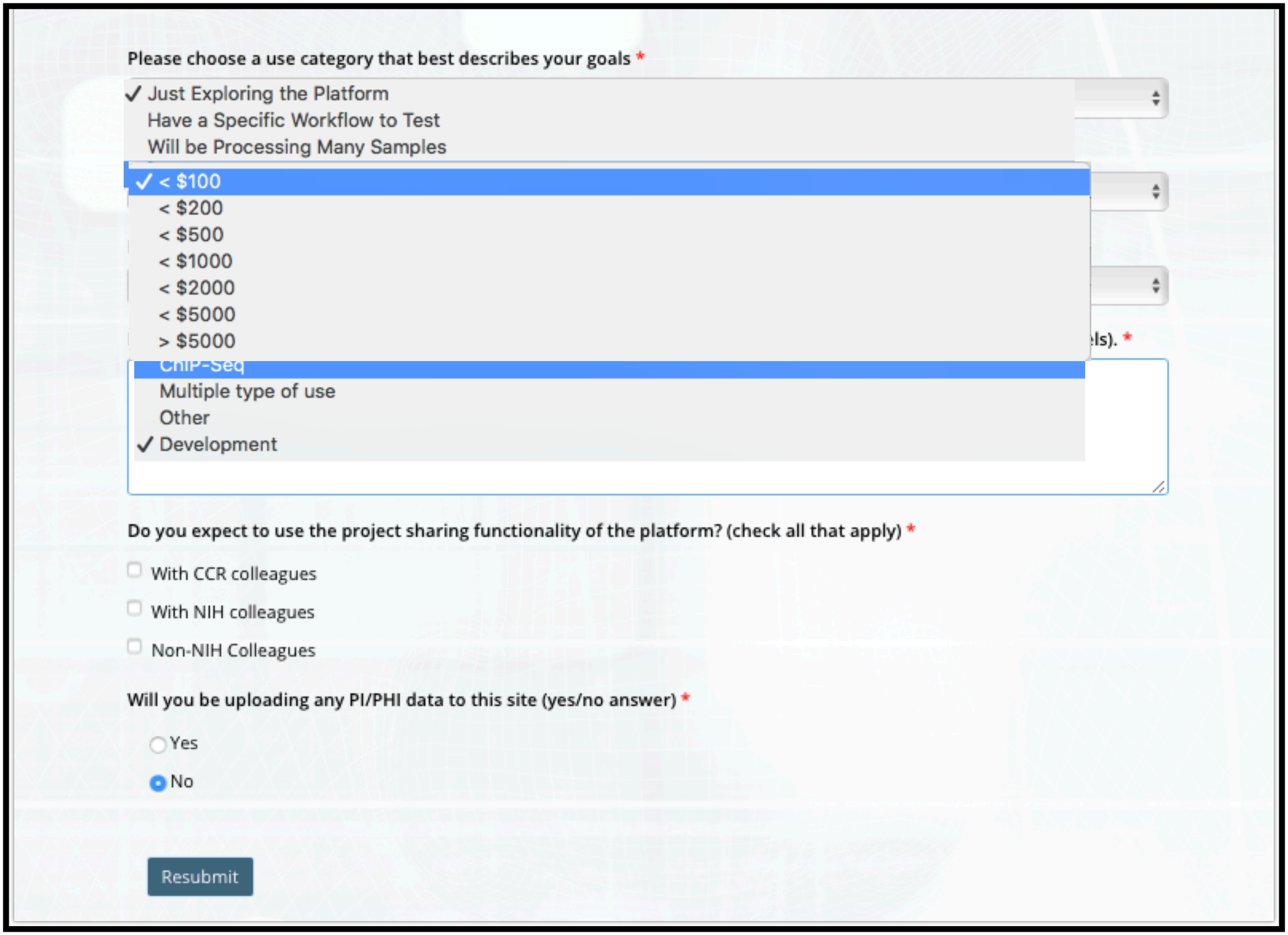
PI Email *

DNAnexus ID *

Slack ID

Lab/Office *

DNAnexus Account Sign up Page
https://gau.ccr.cancer.gov/dnanexus



DNAnexus Accounts

Practical Terms

Current expenses (during this pilot) absorbed by OSTR (within limits)

- **CCR Organizational Account**
  (Setup will depend on your stated likely-use case)
  - **Mode 1 - Dedicated subspace (Administrator)**
    Those who indicate expected heavy use of the system.
    Funds available will be solely dependent on their use
    In the future may be expected to match funds

  - **Mode 2 - Common subspace**
    This who indicate expected light use of the system
    Several groups will be within the same sub-org, visible
    costs will be based on the group as a whole.

DNAnexus Accounts

- **CCR Sub-Organizational Account**
  - **MemberShip** - Member/Administrator
  - **Billable Activities Access** - Allowed/Not Allowed
  - **Shared Projects Access -**
    Administrator
    Uploader
    Viewer
    Not Allowed
  - **Shared Apps Access** - Allowed/Not Allowed

What it Could Be

DNAnexus “The Platform” overs our community some unique opportunities

For Bench Scientists

- Unlimited compute capacity - the power of Biowulf without the pain.
- Global sharing of data and analyses
- Any Genome ?
- Simple Interface with documentation (Could be even better?)
- Not tied to vendor specifics
- Not dependent on large licensing fees.
- Safe secure Storage

What it Could Be

DNAnexus “The Platform” overs our community some unique opportunities

For Developers/Bioinformaticists

- Unlimited compute capacity - scriptable with easy sharing with collaborators
- Global sharing of data and analyses
- “Easy” development environment
- Distribute/publish software in completely runnable fashion
- Not tied to vendor specifics
- Not dependent on large licensing fees.
- Safe secure Storage

DNAnexus is Cloud Computing (Makes use of AWS or Azure)

Interface is to central Server

- **Web-based**
- **Comman-Line (dx-toolkit)**
- **Jobs submitted to a batch system**

Compute jobs carried out on separate worker nodes

Cloud Computing

Cloud Computing

Cloud Computing

- **End-Users**
  - Everything is essentially a batch job.
    Select the analysis, select the files, start the application, wait for it to run and monitor…. wait for email on completion.
  - Movement of files to and from the worker is automated (not your concern)

- **Developer**
  - Everything is essentially a batch job.
  - No common disk space
  - Your application must explicitly move files to the worker node and move the results back to permanent storage
  - Multiple samples can be analyzed in parallel

DNAnexus Projects

Project-Centric World

Projects are the main unit of control and data management
Data and applications reside within a project
Sub Folders within a project are used to ease to task of data management
(A Structured project with sub-folders is essential for successful data management)

Project Level Controls

- **Viewer -** can **view** and download data
- **Uploader -** can **upload** data, but cannot edit data or run apps
- **Contributer -** Can manage data and **run analyses** (can incur charges)
- **Administrator** - Can manage data, **membership** and run analyses (can incur charges)

DNAnexus Features

All files etc are really known by their IDs
(Non-unique file names are possible)
Meta Data associated with files - can be very powerful
Most Apps have good inline instructions
File History is attached
Presigned URLS
(access data without downloading for IGV)

File search and filtering - simple search or regular expressions - very powerful but a couple of hints go a long way

Prebuilt applications and Workflows
Infinite (?) computing resources

Analysis Tools

- **Applets**
  Applets are lighter-weight executables that can be used as scripts for project-specific analyses or ad hoc data manipulations, proprietary analysis pipelines, or development/testing versions of apps. Unlike apps, they reside inside your projects alongside data
- **Applications**
  Apps represent general-purpose tools, striving for compatibility, ease of use, and robustness. They're published in a dedicated section of the website, and typically include extensive metadata and documentation.
- **Workflows**
  Workflows represent a series of executables (apps or applets) that are linked together by dependencies, e.g. one executable's outputs may be another's inputs. It is easiest to create a workflow in the web interface.

- **Web-based**
  Drag and Drop files for upload and download
- **Command-line upload tool**

DNAnexus Help

- On line help
- List of tools
- Common Resources
- DNAnexus support
- DX-toolkit available on Helix/Biowulf (module load)
- CCR support
- CCR community support (through the Slack channel)

  Slack is an cloud-based set of proprietary team collaboration tools and services

St. Jude Cloud

- **Portal to data and Applications built on top of DNAnexus**
  - Unique user-friendly interface
  - Access to St. Jude curated data
  - Billed to your (CCR) account
  - Runs on Azure while the rest is mostly AWS

AWS = Amazon Web Services
Azure = Microsofts Cloud Services

Each environment virtually identical - BUT they are distinct spaces and difficult to move data and apps between the two

- **Potential model for future CCR resources ?**

Sentieon - When Speed is important

Sentieon DNAseq

- Identical mathematics as Broad Institute's BWA-GATK Best Practice Workflow, but over 10X faster FASTQ-to-VCF, 20X-50X faster BAM-to-VCF, measured in core-hours
- No run-to-run difference, no down-sampling in high coverage regions 100K samples joint-calling without intermediate file merging
- Pure software solution running on any generic-CPU-based system

Sentieon TNseq

- Identical mathematics as Broad Institute's Mutect and MuTect2, but over 10X faster FASTQ-to-VCF, measured in core-hours
- No run-to-run difference, no down-sampling in high coverage regions
- Pure software solution running on any generic-CPU-based system

Sentieon TNscope

- Complete tumor-normal somatic variant detection suite, calling SNV, INDEL, and SV
- No run-to-run difference, no down-sampling in high coverage regions
- Pure software solution running on any generic-CPU-based system
- Leads ICGC-TCGA DREAM Mutation Calling Challenge 6

Demos

Its a **platform** not just a program, therefore not limited by prebuilt tools or interfaces, or preconceive notions about appropriate data analysis routines

Demo What is possible

- **RNA-SEQ Workflow**
  Rapid analysis of RNA-Seq data, converting fastq files to count matrices for further analysis.
- **Pausing Peak aligner/finder**
  Powerful applet - simple workflow construction, novel use of metadata, dynamic results
- **IGV Session Maker**
  Access the data without having to wait on large downloads - reproducible views
- **Senteion**
  Rapid optimized identification of SNP's in Whole Genome/Exome Sequencing.

Applications Support

Hands on using DNAnexus - Biologists

Thursday April 11th, 10:00-11:30 am. -
NIH Bldg B37, Rm 4041/4107

Hands-on Session

- **RNA-Seq - Expression, Variants**
  Several applications and Workflows
- **DNA-Seq - Variants, Structural Variants, Copy number**
  Several applications and Workflows including Senteion and Parliment2
- **ChIP-Seq - Binding Sites**
  Several applications and Workflows

Development/Batch Support

DNAnexus Development Environment - Bioinformaticists

Friday April 12th, 10:00-11:30 am. - NIH Bldg 37, Rm 2041/2107

- **dx-toolkit - command line access**
- **Development languages (python, bash, docker)**
- **Applet development**
- **Cloud workstation application**
- **Batch processing**
- **Resource selection and optimization**

The Next Steps

- **Get an account**
- **Joint the CCR ORG**
- **Attend the follow up User Session and/or developer session**
  - Hands on using DNAnexus - Biologists
    Thursday April 11th, 10:00-11:30 am. - NIH Bldg B37, Rm 4041/4107
  - DNAnexus Development Environment - Bioinformaticists
    Friday April 12th, 10:00-11:30 am. - NIH Bldg 37, Rm 2041/2107
- **Use the system, interact with the community, encourage others.**
- **Request help with application use or special case development**