Cold
Spring
Harbor
Laboratory

Advanced Sequencing Technologies & Applications

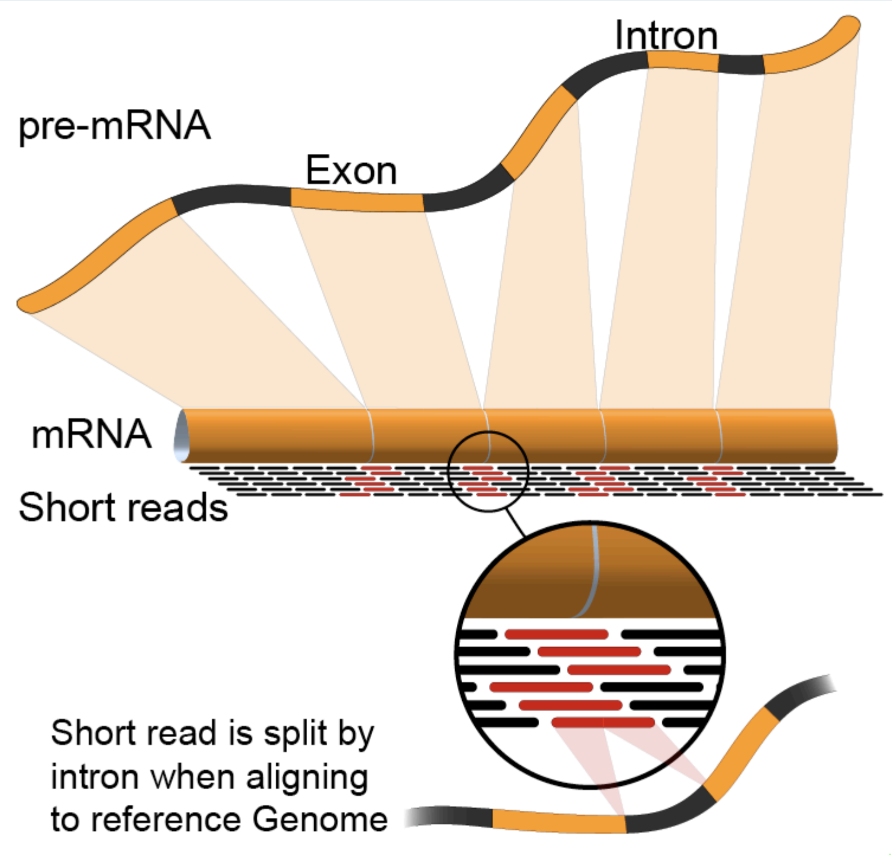
http://meetings.cshl.edu/courses.html

Cold Spring Harbor Laboratory

Introduction to Genome Analysis Platforms

Malachi Griffith, Obi Griffith

Advanced Sequencing Technologies & Applications
November 11-23, 2014


pre-mRNA
Exon
Intron
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

The Genome Institute at Washington University

What is a genome analysis ‘platform’?

- Means different things to different people…
  - Lost of jargon and buzzwords
- Hardware
  - e.g. ‘Dell Genomic Data Analysis Platform’
- Pipelines
- Cloud computing
  - ‘Private clouds’
  - ‘Public clouds’ - Amazon AWS, Google Cloud, digital ocean, etc.
- Virtualization and virtual machines
  - VirtualBox (vagrant), OpenStack, VMWare
- Workflow management systems
- Software development kits (SDKs)
- Application programming interfaces (APIs)
- Distributed storage and processing (e.g. ‘Hadoop’)
- Job schedulers. e.g. pbs, lsf, sge, openlava,

List of existing genome analysis platforms

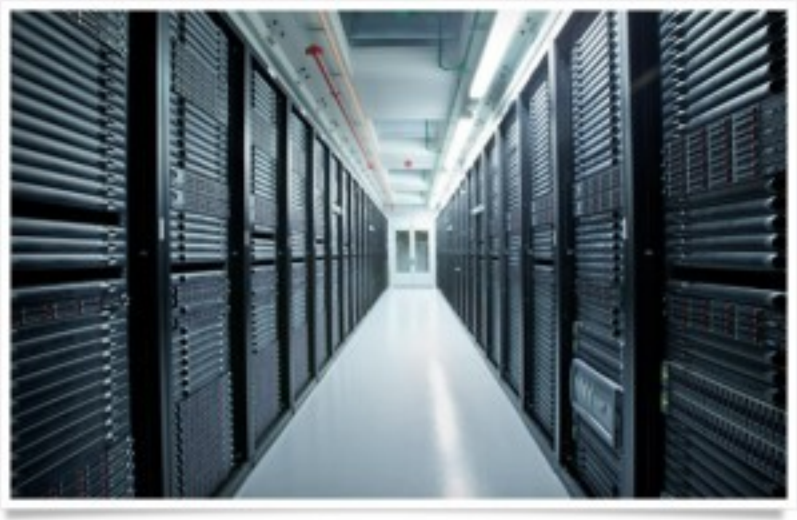
- https://docs.google.com/spreadsheets/d/1o8iYwYUy0V7IECmu21Und3XALwQihioj23WGv-w0itk/pubhtml
- Genome Modeling System (GMS), Galaxy, bcbio-nextgen, Omics Pipe, Illumina BaseSpace, BINA Genomic Analysis System, SeqWare, DNA Nexus Platform, gkno, NGSANE, Appistry's Ayrris, GATK's Queue, Curoverse's Arvados, CGA's Firehose, Seven Bridges Genomics, MIT STAR, GenomOncology, ga4gh, IBM's PowerGene Orchestrator, etc.

What is a job scheduler?

- A **job scheduler** is a computer application for controlling unattended background program execution (commonly called batch processing).
- For example, in genomics data processing, a researcher might use a job scheduler to submit 100 tophat alignment jobs to a cluster of computers at their institute's data center

What is cloud computing?

- The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
- For example, instead of using a local server or buying 25 computers with 8 CPU's each, 70Gg of RAM, etc. for the RNA-seq course we rented these computers on the Amazon 'Cloud'. All analysis for the course actually happened at a massive data center in Northern Virginia

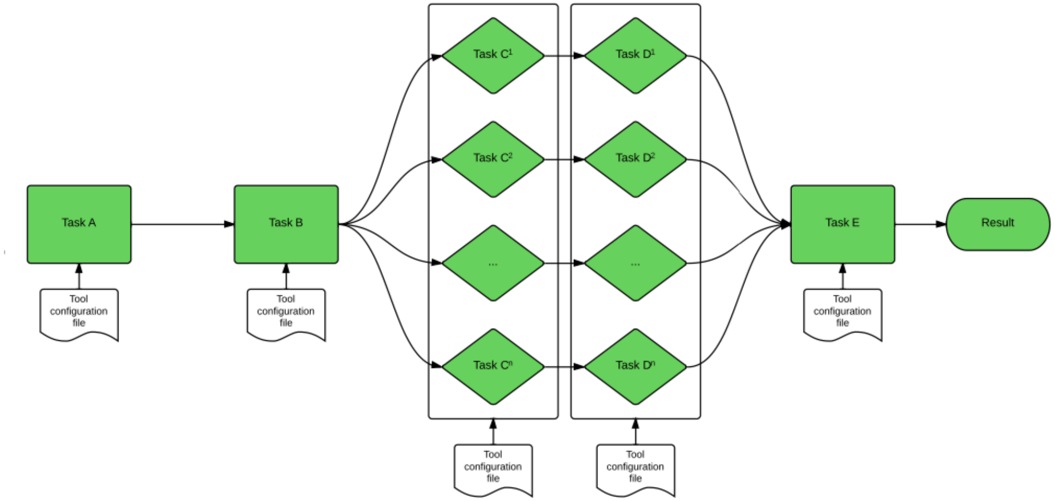


What is a virtual machine?

- A *virtual machine* (*VM*) is an operating system OS or application environment that is installed on software which imitates dedicated hardware. The end user has the same experience on a *virtual machine* as they would have on dedicated hardware.
- In the context of genome analysis pipelines a virtual machine may sometimes be used to allow researchers to share and distribute very complex computing environments (may dependencies) that are difficult to set up.

What is a workflow management system?

- A **workflow management system** (WfMS) is a software **system** for the execution of a defined sequence of tasks, arranged as a **workflow**.
- For example, the RNA-seq analysis has many steps with interconnected dependencies
  - TopHat alignment of several lanes of data needs to happens before they can all be merged into a final BAM file, and merging needs to happen before indexing of the BAM, and so on.
  - Some steps can happen in parallel, other in series. Workflow systems help handle these dependencies



What is a software development kit (SDK)?

- A software development kit (SDK) is a set of software development tools that facilitates the creation of applications for a certain software framework
- E.g. DNA Nexus Platforms provides software development kit with support for several programming languages to help you build pipelines efficiently in their system

What is an application programming interface (API)

- An API is a set of routines, protocols, and tools for building software applications. An API expresses a software component in terms of its operations, inputs, outputs, and underlying types. An API defines functionalities that are independent of their respective implementations, which allows definitions and implementations to vary without compromising each other.

What is distributed storage and processing (e.g. ‘Hadoop’)?

- Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. Its Hadoop Distributed File System (HDFS) splits files into large blocks and distributes the blocks amongst the nodes in the cluster. For processing the data, the Hadoop Map/Reduce moves code (software) to the nodes that have the required data, and the nodes then process the data in parallel. This approach takes advantage of data locality, in contrast to conventional HPC architecture which usually relies on a parallel file system (compute and data separated, but connected with high-speed networking).

Examples

- **Assuming you have some NGS data, how should you analyze it?**
- Depends where you are on the informatics spectrum. Do you want to:
  - Build a completely novel process, a custom pipeline, develop algorithms, write software, etc.
    - Maximum flexibility. Performance and scalability are determined by how well you engineer it.
  - Build on top of someone else genome analysis platform
    - Don't have to start from scratch but still have a lot of flexibility.
    - e.g. GMS, Arvados, DNA Nexus, bcbio-nextgen, Gkno, etc.
  - Upload data in web browser, use graphical user interface
    - Sacrifices flexibility for ease of use
    - Galaxy, Illumina BaseSpace
  - Have someone do the analysis for you and give you the results

Companies whose platforms amount to bioinformatics for hire…

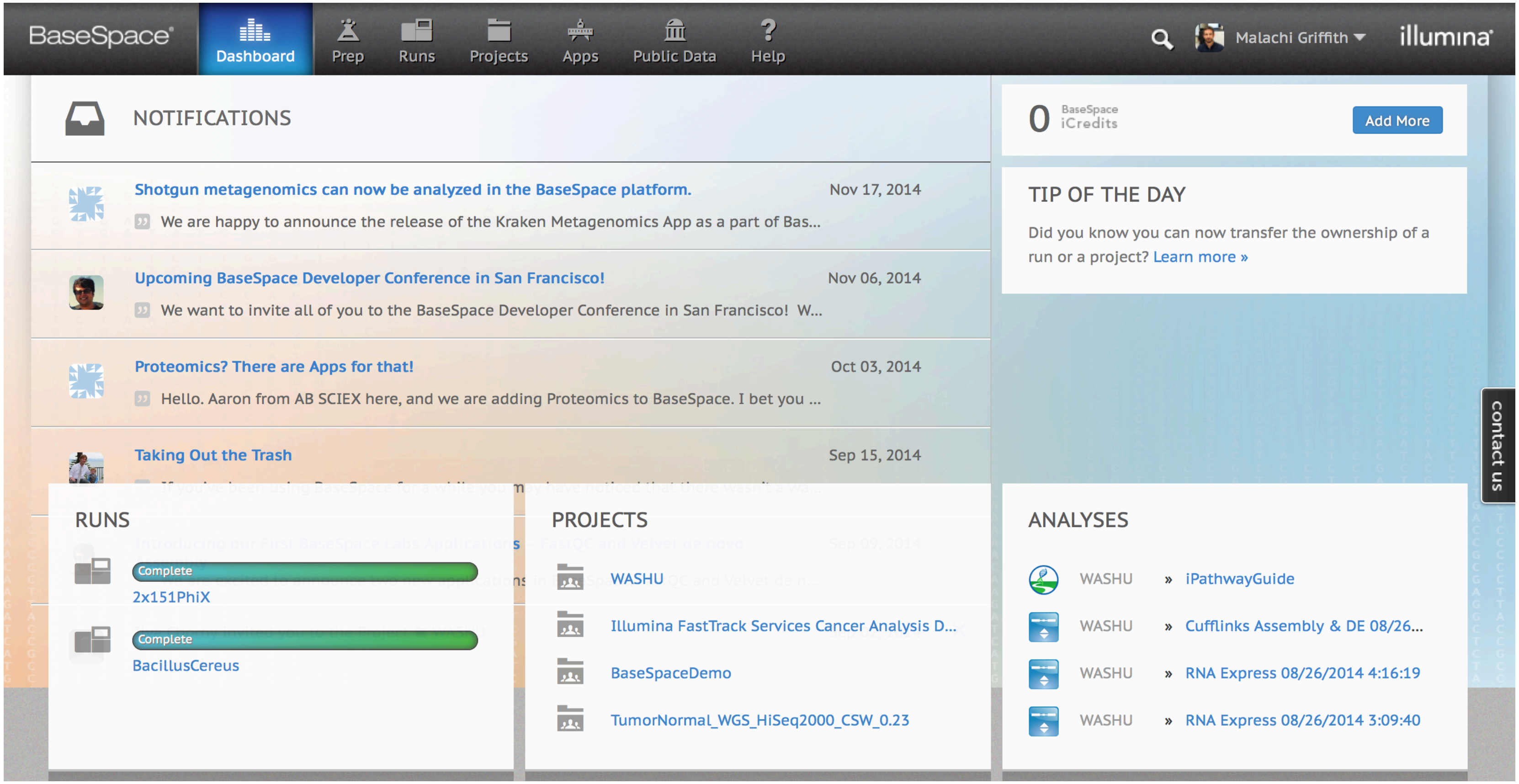
- Appistry's Ayrris
- Seven Bridges Genomics
- GenomOncology
- IBM's PowerGene Orchestrator
- BINA Genomic Analysis System
  - Sort of. They provide a pre-configured hardware + software solution, help you install it and connect it to your in house data production
- Etc.

Galaxy

- http://galaxyproject.org/
- Open Source academic project.
- Example RNA-seq workflow
  - https://usegalaxy.org/u/mwolfien/w/rnaseq-wolfien-pipeline

- A web based interface that allows you to run existing workflows or create custom analyses by combining tools in the Galaxy 'toolshed'

Illumina BaseSpace



NOTIFICATIONS

Shotgun metagenomics can now be analyzed in the BaseSpace platform. We are happy to announce the release of the Kraken Metagenomics App as a part of Bas...	Nov 17, 2014
Upcoming BaseSpace Developer Conference in San Francisco! We want to invite all of you to the BaseSpace Developer Conference in San Francisco! W...	Nov 06, 2014
Proteomics? There are Apps for that! Hello. Aaron from AB SCIEX here, and we are adding Proteomics to BaseSpace. I bet you ...	Oct 03, 2014
Taking Out the Trash If you've been using BaseSpace for a while you may have noticed that there wasn't a wa...	Sep 15, 2014

0 BaseSpace iCredits

TIP OF THE DAY

Did you know you can now transfer the ownership of a run or a project? Learn more »

RUNS

Complete
2x151PhiX

Complete
BacillusCereus

PROJECTS

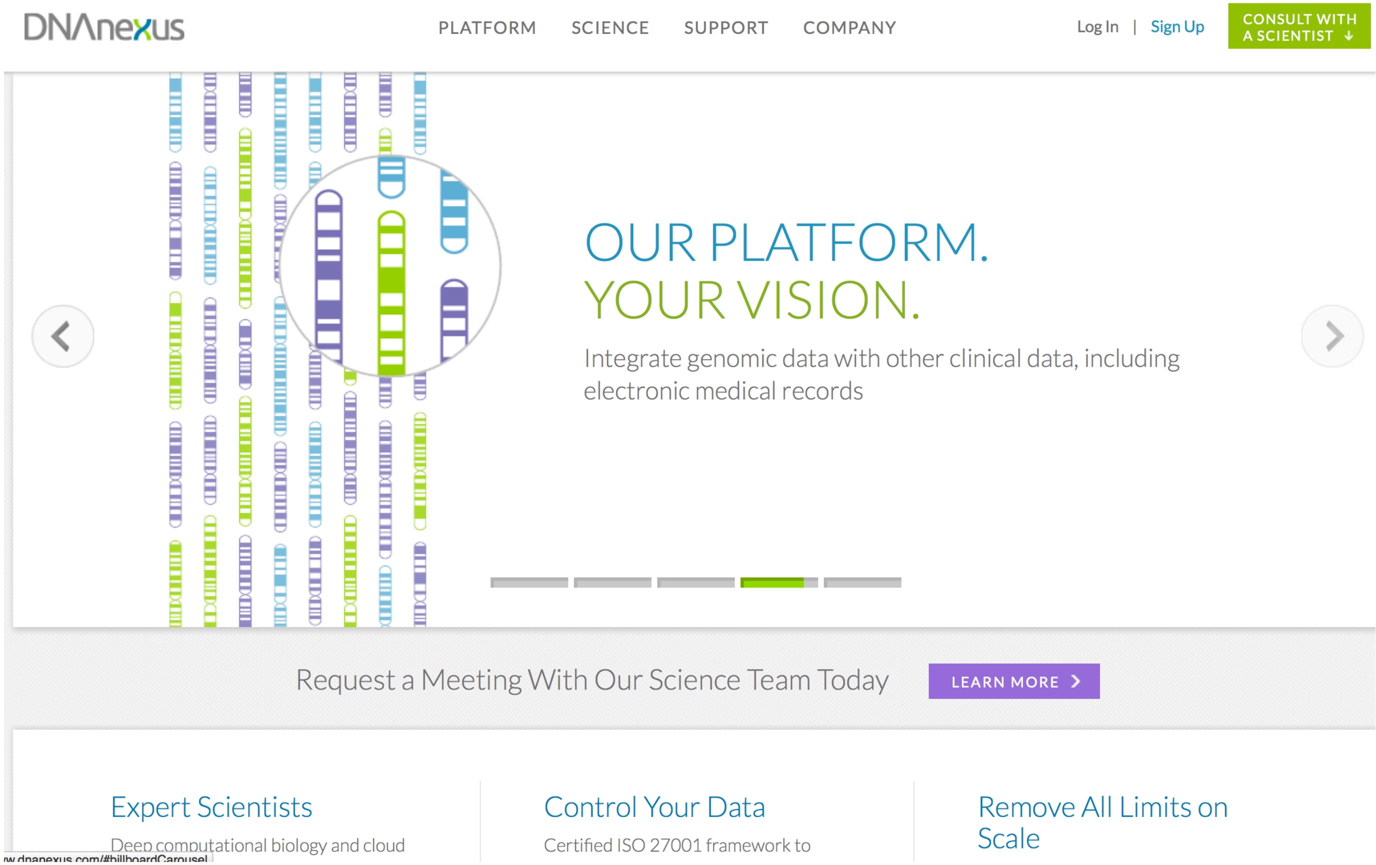
WASHU
Illumina FastTrack Services Cancer Analysis D...
BaseSpaceDemo
TumorNormal_WGS_HiSeq2000_CSW_0.23

ANALYSES

WASHU » iPathwayGuide
WASHU » Cufflinks Assembly & DE 08/26...
WASHU » RNA Express 08/26/2014 4:16:19
WASHU » RNA Express 08/26/2014 3:09:40

- Use integrated ‘apps’ and automated pipelines.
- Graphical interface
- https://basespace.illumina.com

DNA Nexus Platform



OUR PLATFORM.
YOUR VISION.

Integrate genomic data with other clinical data, including electronic medical records

Request a Meeting With Our Science Team Today

Expert Scientists

Deep computational biology and cloud

Control Your Data

Certified ISO 27001 framework to

Remove All Limits on Scale

- Build your own pipeline or use an existing one
- DNA Nexus handles cloud deployment, etc. for you
- https://www.dnanexus.com/

Other pipeline development platforms to build on top of

- Gkno
  - http://gkno.me/
- Genome Modeling System (GMS)
  - https://github.com/genome/gms
- Arvados
  - https://arvados.org/
- Bcbio-nextgen
  - https://bcbio-nextgen.readthedocs.org/en/latest/
- OmicsPipe
  - http://sulab.org/tools/omics-pipe/
- NGSANE
  - https://github.com/BauerLab/ngsane

The Global Alliance for Genomics Health (ga4gh)

- An international coalition, formed to enable the sharing of genomic and clinical data.
- Work on data models and APIs for Genomic data.
- Not yet entirely clear what is available to be used by end users beyond the ‘beacon’ project:
- http://genomicsandhealth.org/
- http://ga4gh.org/#/
- https://github.com/ga4gh
- http://ga4gh.org/#/beacon