Cold Spring Harbor Laboratory

Advanced Sequencing Technologies & Applications

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

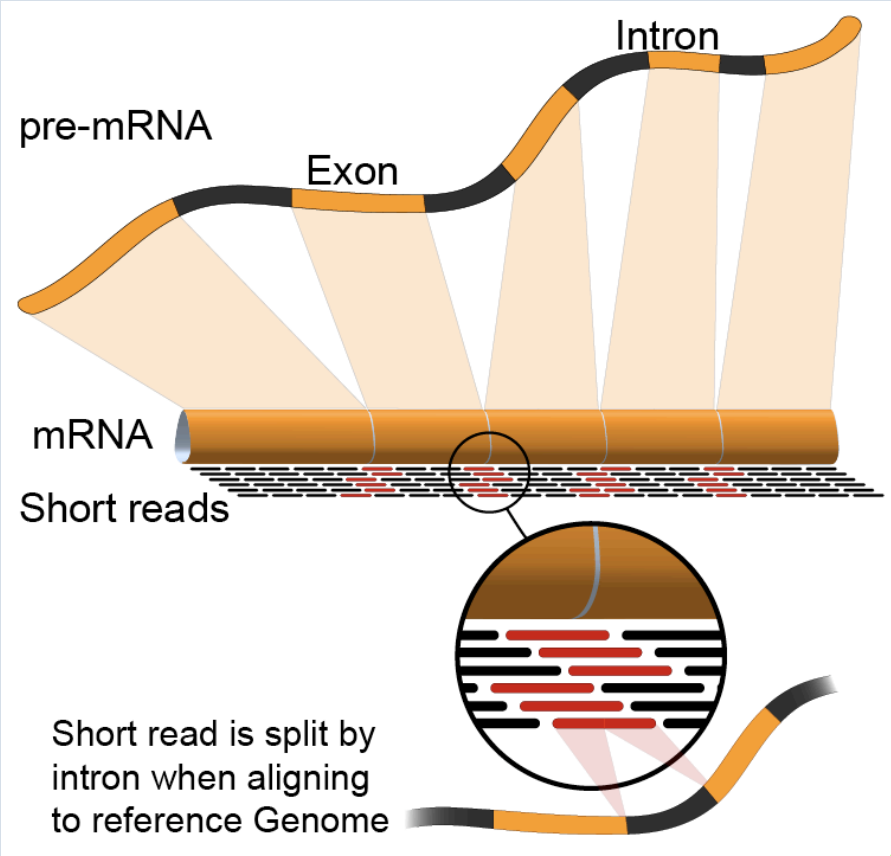
Cold Spring Harbor Laboratory

Introduction to Genome Analysis Platforms

Malachi Griffith, Obi Griffith, Jason Walker, Alex Wagner

Advanced Sequencing Technologies & Applications

November 7 - 20, 2016



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 and workflow languages
- Software development kits (SDKs)
- Application programming interfaces (APIs)
- Distributed storage and processing
- 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 HISAT alignment jobs to a cluster of computers at their institute's data center

What is cloud computing?

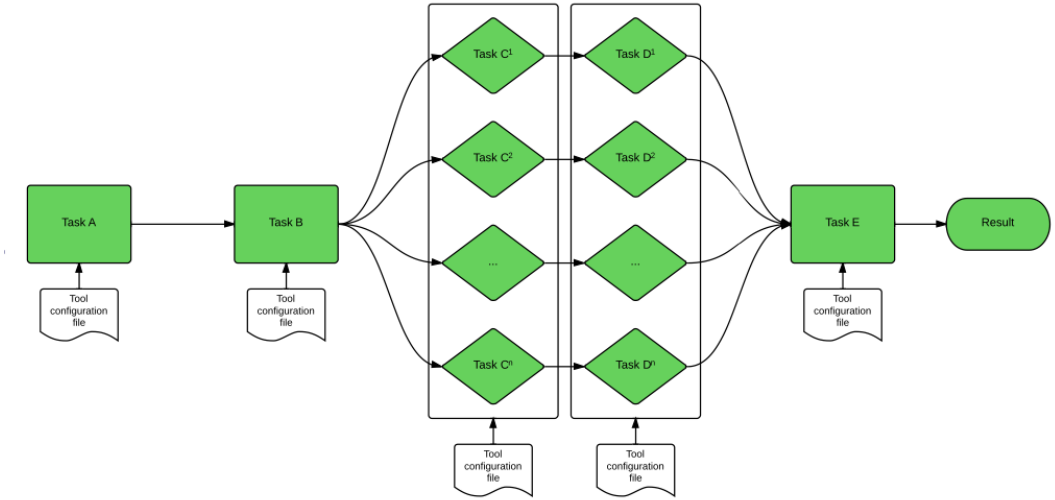
- The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.
- For example, instead of using a local server or buying 25 computers with 8 CPU's each, 70Gg of RAM, etc. for the RNA-seq course we rented 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 workflow language?

- A workflow language is “a specification for describing analysis workflows and tools that are portable and scalable across a variety of software and hardware environments, from workstations to cluster, cloud, and high performance computing (HPC) environments”.
- Two main workflow language being used for NGS:
  - Common Workflow Language (CWL)
  - Workflow Description Language (WDL)

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 a “container” (e.g. Docker container?)

- “Containers wrap up a piece of software in a complete filesystem that contains everything it needs to run: code, runtime, system tools, system libraries – anything you can install on a server. This guarantees that it will always run the same, regardless of the environment it is running in”
  - “Docker provides an additional layer of abstraction and automation of operating-system-level virtualization on Linux. Docker uses the resource isolation features of the Linux kernel to allow independent "containers" to run within a single Linux instance, avoiding the overhead of starting and maintaining virtual machines.”

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.

Examples

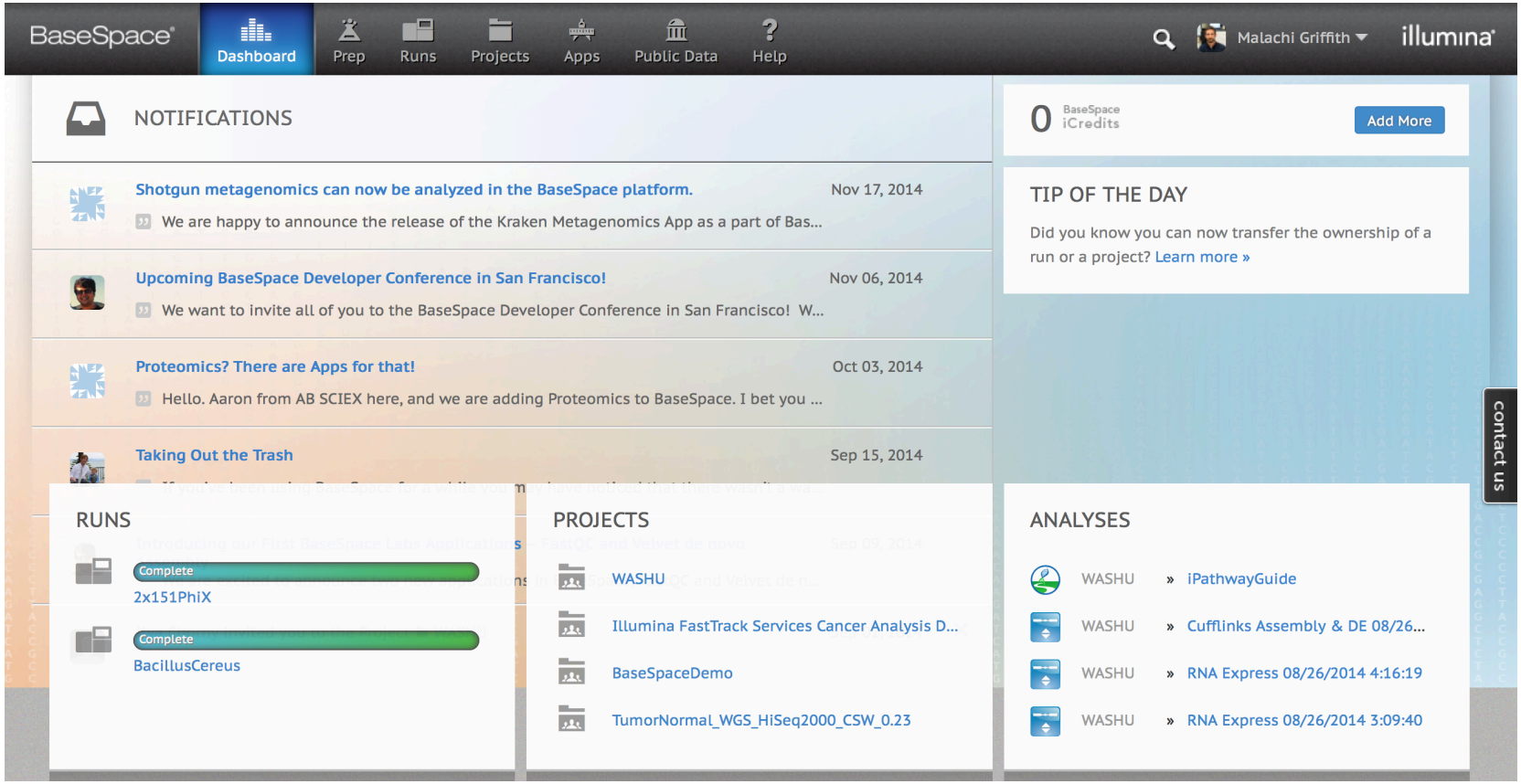
- **Assuming you have some NGS data, how should you analyze it?**
- Depends where you are on the informatics spectrum. Do you want to:
  - Build a completely novel process, a custom pipeline, develop algorithms, write software, etc.
    - Maximum flexibility. Performance and scalability are determined 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

Galaxy

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

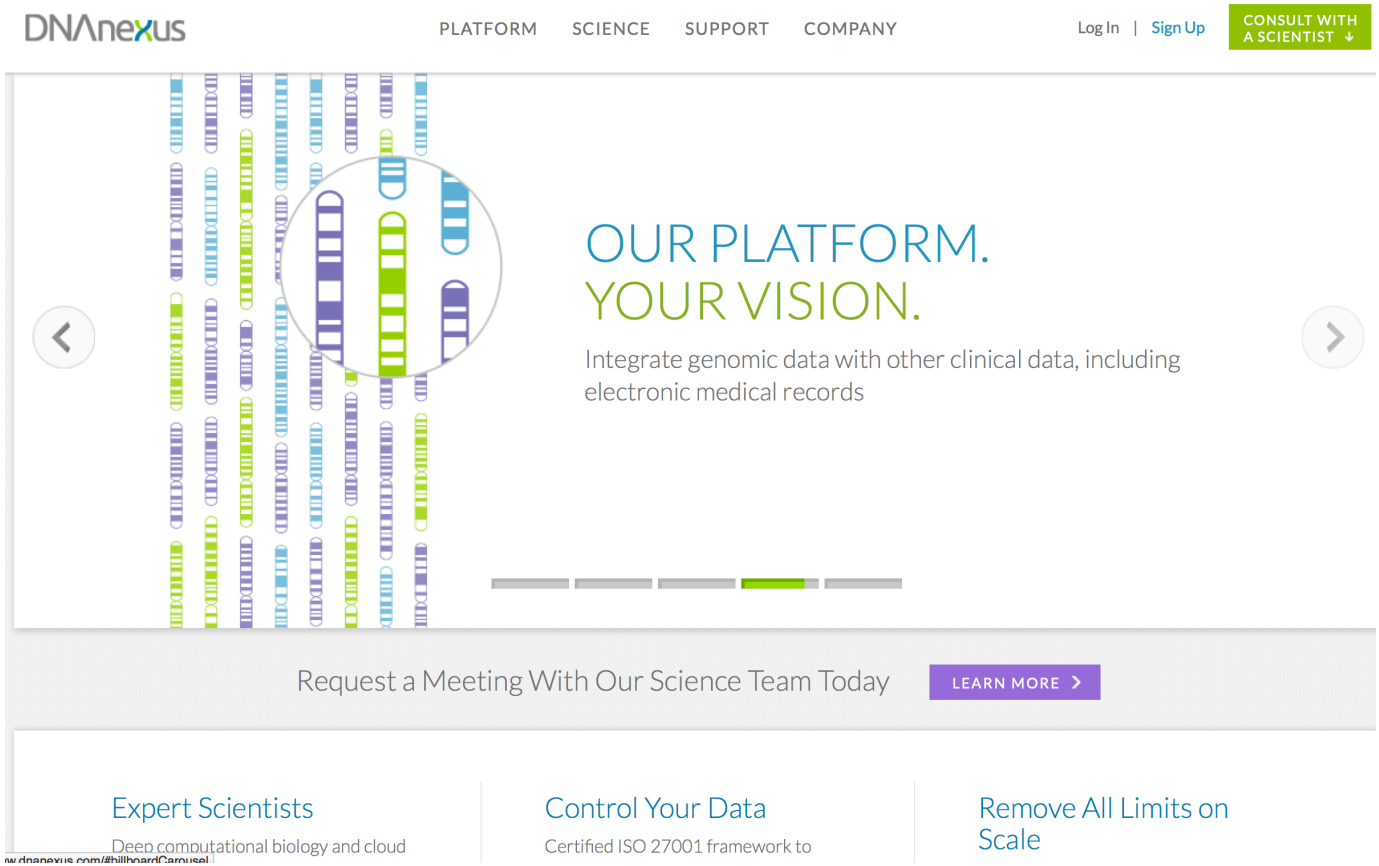
- A web based interface that allows you to run existing workflows or create custom analyses by combining tools in the Galaxy 'toolshed'

Illumina BaseSpace



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

DNA Nexus Platform



- 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

The NCI Cancer Genomics Cloud (CGC) Program

- "Bringing data and computation together to create knowledge that accelerates cancer research and enables precision medicine". NCI Cloud Initiative.
- Make large cancer data sets available (e.g. TCGA and TARGET) and allow you to analyze these data on the cloud without downloading yourself
- Credits to perform analysis are available: Cancer Cloud Credits
- Three CGC implementations available
  - ISB Cancer Genomics Cloud
  - Broad FireCloud
  - Seven Bridges Cancer Genomics Cloud

The *NIH Commons*

- The NIH Commons
  - "The Commons is a shared virtual space where scientists can work with the digital objects of biomedical research, i.e. it is a system that will allow investigators to find, manage, share, use and reuse data, software, metadata and workflows."
- The Commons Credit Portal
  - Used to apply for compute credits
    - For example, to perform analysis on approved vendors such as Amazon, IBM, Microsoft or Google clouds
    - Need to provide scientific proposal (analysis plan) and detailed budget for compute costs

Break