NYUAD
Center for Genomics and System Biology

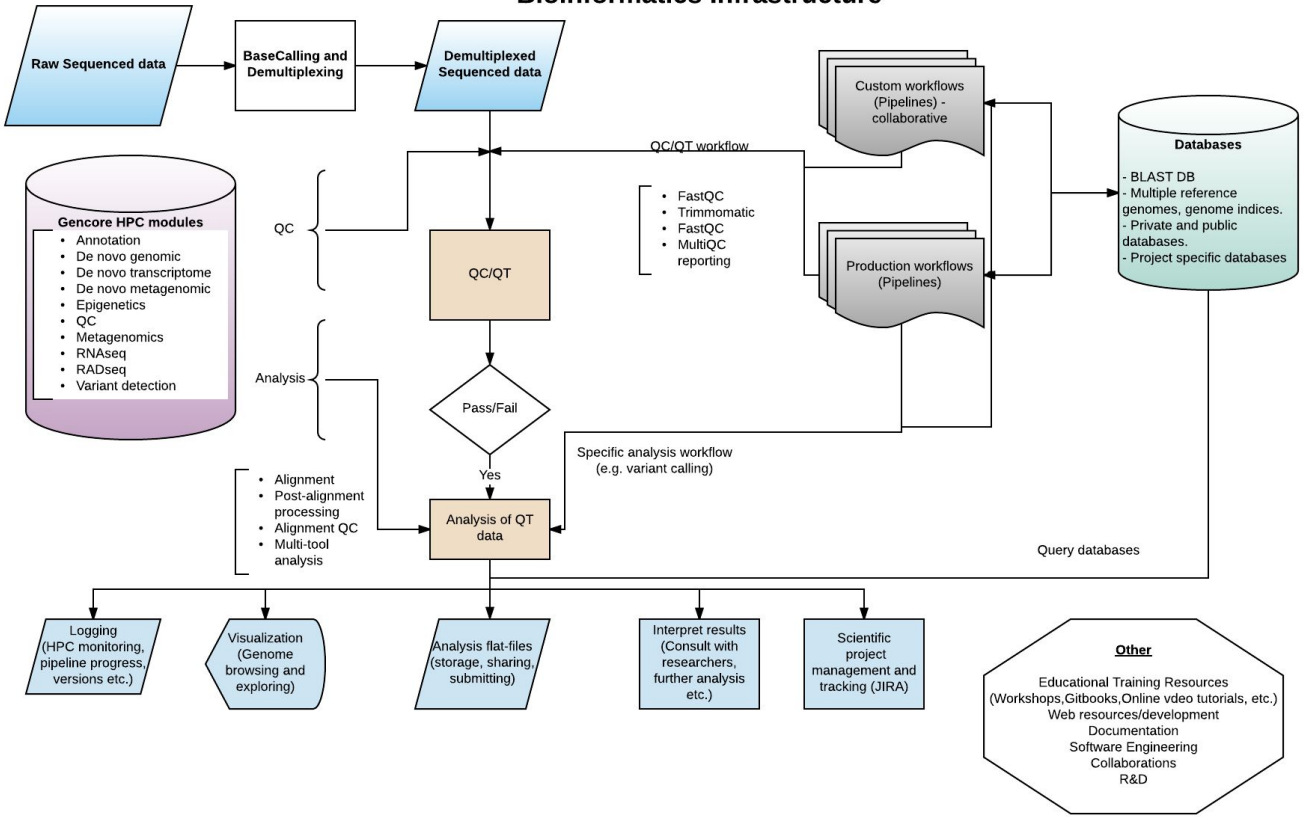
Jillian Rowe, NYU Abu Dhabi Center for Genomics and Systems Biology

- Bioinformatics Core Sequencing
- Infrastructure
- Anaconda
- Gencore Modules (with EasyBuild!)

1.

Core Sequencing

Overview of NYUAD Bioinformatics Infrastructure

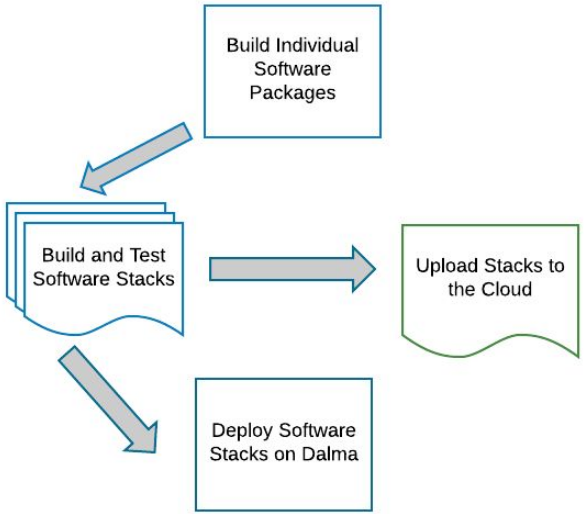


2.

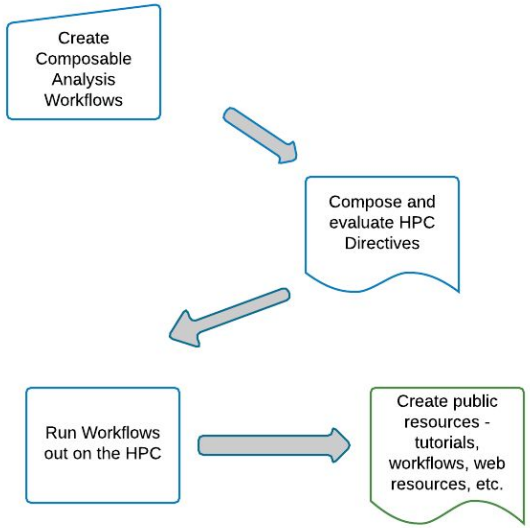
Infrastructure

Our goal is to have a fully tested, semi automated infrastructure from the level of testing and installing individual software packages, to running analyses on the HPC.

Infrastructure



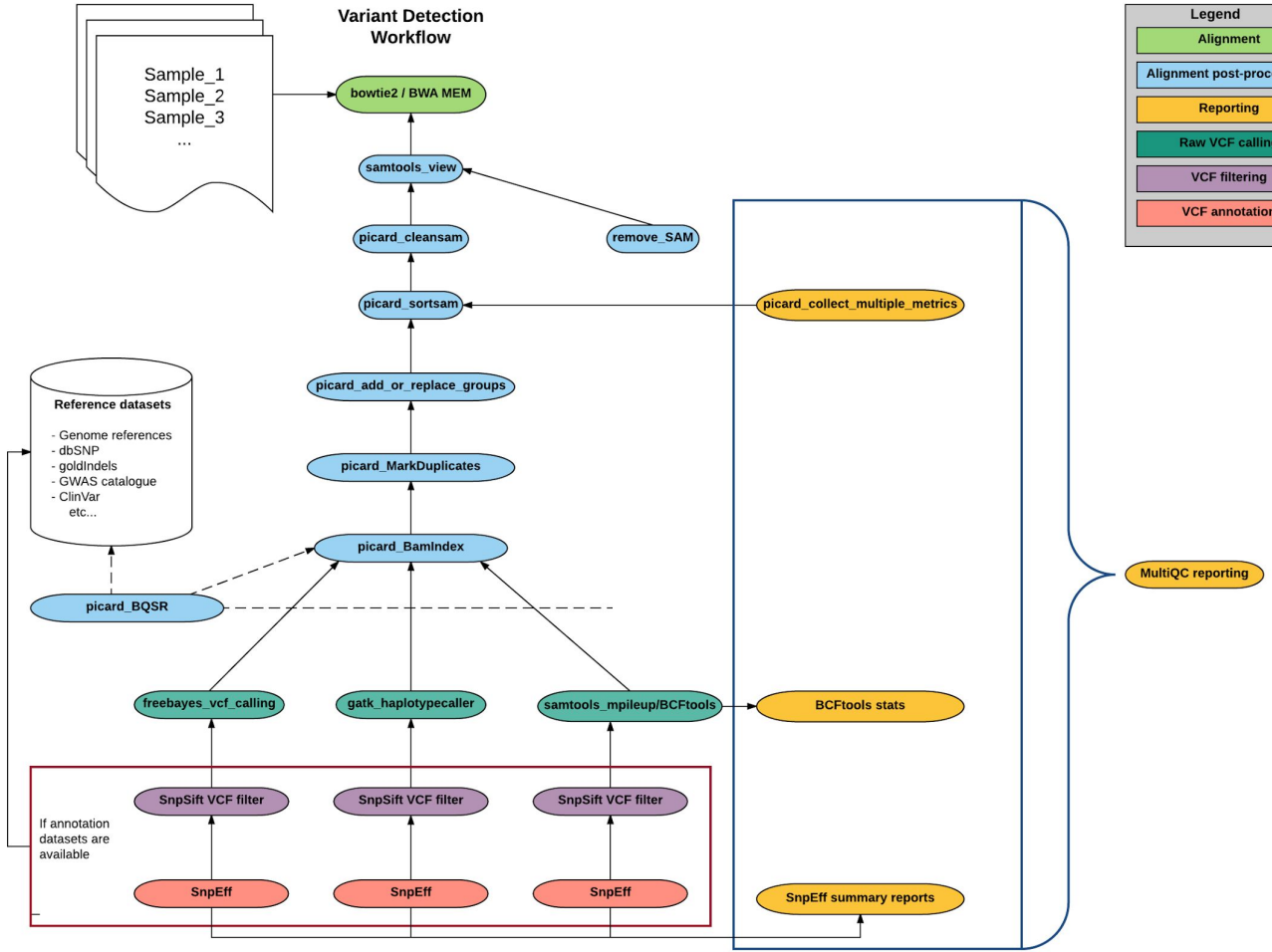
Analysis



Internal Development and Testing

Much of the testing and development of the overall infrastructure is done inhouse by the gencore team.

1. User requested software installs
2. Workflow Software - HPCRunner, BioX-Workflow
3. Scientific development of analyses (researching best practices, scouring papers)


Variant Detection Workflow
Sample_1
Sample_2
Sample_3
...
bowtie2 / BWA MEM
samtools_view
picard_cleansam
remove_SAM
picard_sortsam
picard_collect_multiple_metrics
picard_add_or_replace_groups
picard_MarkDuplicates
picard_BamIndex
Reference datasets
- Genome references
- dbSNP
- goldIndels
- GWAS catalogue
- ClinVar
etc...
picard_BQSR
freebayes_vcf_calling
gatk_haplotypecaller
samtools_mpileup/BCFtools
BCFtools stats
SnpSift VCF filter
SnpSift VCF filter
SnpSift VCF filter
If annotation datasets are available
SnpEff
SnpEff
SnpEff
SnpEff summary reports
MultiQC reporting
Legend
Alignment
Alignment post-processing
Reporting
Raw VCF calling
VCF filtering
VCF annotation

Anaconda

GET SUPERPOWERS WITH ANACONDA

Anaconda is the leading open data science platform powered by Python.

The open source version of Anaconda is a high performance distribution of Python and R and includes over 100 of the most popular Python, R and Scala packages for data science.

Continuum Analytics

Why Anaconda?

1. No system dependencies!
2. Libraries that would normally be system libraries are built in as packages.
3. The package manager conda
4. Anaconda Client exposes an API to the packages.

Bioinformatics software analyses are comprised of one or more packages.

PROS

1. A single scientist can solve a valuable problem
2. The barrier to get software out in the wild is relatively low

CONS

1. Gathering requirements for an analysis is like assembling a ~1M piece jigsaw puzzle.

Collaborations

When possible, we try to collaborate with outside teams who do similar work.

1. Conda and Conda Env
    a. OS-agnostic, system-level binary package manager and ecosystem http://conda.pydata.org
2. Bioconda
    a. Bioconda is a distribution of bioinformatics software realized as a channel for the versatile Conda package manager
3. Easybuild
    a. EasyBuild is a software build and installation framework that allows you to manage (scientific) software on High Performance Computing (HPC) systems in efficiently.

Gencore Module System

1. Name: gencore_variant_detection_1.0
2. Channels
   a. Each channel is a different group contributing software
3. Dependencies
   a. ~25 software packages
   b. Each of these depends upon others, leading to >250 software packages total.

Gencore Variant Detection

Travis Builds

NYUAD CGSB Environments

```
name: gencore_variant_detection_1.0
channels:
- bioconda
- r
- nyuad-cgsb
dependencies:
- perl-hpc-runner-slurm=2.58
- perl-biox-workflow=1.10
- perl-biox-workflow-plugin-fileexists=0.13
- perl-biox-workflow-plugin-filedetails=0.11
- discovar=52488
- discovardenovo=52488
- blast=2.2.31
- bwa=0.7.15
- samtools=1.3.1
- bcftools=1.3.1
- bedtools=2.25.0
- vcftools=0.1.14
- freebayes=1.0.2.29
- bamtools=2.4.0
- seqtk=1.2
- pear=0.9.6
- bowtie2=2.2.8
- tophat=2.1.0
- cufflinks=2.2.1
- circos=0.69.2
- star=2.5.2a
- blat=35
- gatk=3.5
- picard=2.5.0
- prinseq=0.20.4
- snpeff=4.3
- vcflib=1.0.0_rc1
- r
- r-base
- r-essentials
- bioconductor-biobase
- gencore_variant_detection_docs=1.0
```

Bioconda

Bioconda is an open source group that contributes bioinformatics software packages to the conda package manager.

It has a very robust build and test system, as well as just having sheer man power thrown at the software problems we all face.

Travis Builds

NYUAD Gencore App

Each environment is deployed to anaconda cloud using an application that was developed in house.

Additionally, each stack builds documentation and an EasyBuild config.

Easybuild

jerowe commented on Jun 2, 2016

It works! Thanks!

boegel commented on Jun 2, 2016 Member

@jerowe Please consider contributing back the easyblock/easyconfig, see also https://github.com/hpcugent/easybuild/wiki/Contributing-back .

If that's too much for now, please provide them via a gist (cfr. https://gist.github.com).

Also, please close this issue if you consider your problems answered.

Gencore Module System

Gencore_* modules were all deployed with EasyBuild!

```
[gencore@login-0-4 ~]$ :
[gencore@login-0-4 ~]$ :
[gencore@login-0-4 ~]$ :module avail

--------------------------- /scratch/gencore/.local/easybuild/modules/all ----------------------------
EasyBuild/2.8.2                  gencore_de_novo_metagenomic/1.0   gencore_qc/1.0
gencore_anaconda/2-4.0.0         gencore_de_novo_transcriptome/1.0 gencore_rad/1.0
gencore_anaconda/3-4.0.0         gencore_dev/1.0                   gencore_rad_ddocent/1.0
gencore_annotation/1.0           gencore_epigenetics/1.0           gencore_rnaseq/1.0
gencore_base/1.0                 gencore_malaria/1.0               gencore_soapdenovo2/1.0
gencore_build/1.0                gencore_metagenomics/1.0          gencore_tuxedo/1.0
gencore_de_novo_genomic/1.0      gencore_metagenomics_dev/1.0      gencore_variant_detection/1.0

--------------------------------- /share/apps/NYUAD/modules/SOFTWARE ---------------------------------
admirable/2015.01       discovardenovo/52488    libpng/1.6.24           python/2.7.11
allinea/5.0             eigen/3.2.8             libtool/2.4.6           qchem/4.4
allpathslg/52488        expat/2.1.0             libxml2/2.9.2           R/3.1.2
```

EasyBlocks PR

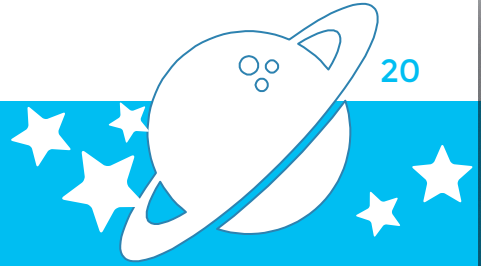
Future Work

1. Add features to the Conda Easyblock
2. **Use sanity check test features from EasyBuild to add more robust testing the modules themselves.**
3. Open collaborations with other sequencing institutes.
4. Create infrastructure for modules as a service - those that wish to create their own software stacks can easily do so.

Collaborate

Why should we collaborate ?

Many bioinformatics teams operate only in house, with all tools and platforms being developed by the time. Genomics institutes everywhere are trying to solve big problems on a massive scale. This can't happen if we all work in isolation. Collaborate!



BIOCONDA

CONDA

easybuild

docker

Publically Available Resources

1. Gencore Easybuild Configs - On Github
2. Variant Detection POC - Variant Detection
3. NYUAD CGSB Website
4. NYUAD CGSB GitHub Site
5. NYUAD Virtual Machines Hosting
6. NYUAD Software Stacks - In the cloud!

Acknowledgements

NYUAD Core Bioinformatics

Nizar Drou, Kristin Gunsalas

Ayman Yousif

Alan Twaddle

NYUAD HPC

Muataz Barwani

Benoit Marchand

Jorge Naranjo, Guowei Hei

NYU Core Bioinformatics

David Gresham

Mohammed Khalfan

Tatiana Polunina