GenAP: Genetics and Genomics Analysis Platform

Louis Létourneau
MonBUG
2015/03/04
Data Processing issues

- We have many different projects all needing space and processing.
- We needed to support 200 50x whole genomes when we started, we now need to support ~3000 WGS at 30x. We are not even counting the exomes and other datasets...

- The solution is to use the Compute Canada clusters at our disposal.
- Which bring uniformity problems
  - Different setups Hardware and Software
  - Different configurations
  - Etc.
More than data, technology issues as well.
Our Basic Pipeline Goals

- Easy to maintain/Update
- Leverage sample meta-data in nanuq (sampleSetup)
- Resume on failure
- Have execution reports
- Have pipeline reports
- Reuse, not rewrite
- Track parameters for reproducibility
Other solutions

- Pegasus, taverna and the like require special scheduler configurations we don't control on clusters
- Bpipe seems ok, but is complicated to configure on complicated dependencies
- BigDataScript (written by Pablo, snpEff author at the center), is similar to our approach. Was developed in parallel.
History circa 2009-2011

• Original pipeline was mostly for alignments/snp calling in bash
• The rest was manual
• We didn't use clusters, we used gnu parallel between servers

• Team got bigger
• SGE was installed
• Many ChIP-Seq projects were coming...

This didn’t scale at all
History circa 2011-2012

• With SGE came first pipeline “rewrite”
• Started as daemons ... kinda ... for RNA-Seq and ChIP-Seq
• They use SGE -sync + & (background) + waits
• Seemed to work but submit node died sometimes
• If a job died daemon could be stuck

• DNA-Seq was written in perl with scheduler dependencies
• This was not cluster portable (msub, qsub)
• Not easily configurable between Exome and WGS
• We used BASH env. variables to control software locations and some params
• We started to force a project structure (raw reads, alignments, etc)
• Cluster parameters were still problematic between clusters.
• When steps finished an empty .done was created.
• But this solved only half the problems.

History circa 2012-2013
History circa 2012-2014

- We re-wrote/enhanced that perl pipeline by adding use of ini files for configurations
- We separated the tools/jobs from the pipeline/wrappers
- We checked time of in/out files as well as .done files
- Started using Linux Modules instead of hard coded paths
What didn’t work well

• We relied on scheduler dependencies
• Wrappers were very different between pipelines
• We needed to write too much boiler plate code...there is no excuse for this
• Many pipelines were not well tested
• Documentation separate and was hard to maintain
What worked

- We wrote our pipelines to be easily configurable across clusters.
- Same code, one ini file to customize (we already have templates for 3 sites)
- Track job logs (mem used, walltime vs cpu time, etc)
- We installed and maintain Linux modules readable by all on all these clusters so we know exactly what is available everywhere
- We also deploy common genomes across sites.
- We optimized the performances of many tools on the parallel file systems they run on (IOBUFF, picard params, etc)

https://bitbucket.org/jfstpierre/iobuff
New and improved!

• Joel added a lot of structure
• Rewritten in Python
• Simplified software dependencies installation
• Simplified, documented and improved genome reference installations
• Reworked the documentation
  • The documentation has been moved from the wiki to the source directly
  • It’s simpler and follows code releases
Why Python?

• To be cool...it’s important to be cool.

• Better O-O language than Perl.

• For collaborative work. Less code conventions to enforce
In Production!

The pipeline is used internally by Nanuq
  • PacBio automated bacterial genome assembly
  • Illumina run processing, BAM, metrics, blasts, plots, etc

This has actually been a question we’ve been getting over and over:

Since June 2013 there are no more Fastqs available
The BAMs are lossless
BWA is used for DNA, STAR is used for RNA (since Jan 2015)
TopHat -> STAR

• We weren’t happy with the support by the authors
• We weren’t happy with the bugs that kept creeping in each release
• We weren’t happy with the performance (walltime)

We tested both (thanks Mathieu)
• We get the same results with Diff. Expression analysis
• STAR is faster (albeit more RAM intensive)
• Other labs got better results with STAR
• Fusions we identified with TopHat were still found with STAR
• STAR has Softclips (poly A tail detection)
• STAR seems to generate less false positive fusion events. Still not perfect though
Current pipelines

• DNASeq (WES, WGS)
  • Broad Best Practices + SnpEff (dropping samtools, using HaplotypeCaller, thinking of freebayes...)
  • Cancer/Paired (MuTect, lumpy coming soon, SCONES, SMUFIN(under test))
• RNASeq (stranded or not)
  • Broad Best Practices (STAR)
  • SNP calling coming soon
  • fusions coming soon)
• ChiPSeq
• RNASeq Denovo
  • Trinity + Trinotate
• Pacbio HGap assembly
  • Automatic when genome < 10Mb
  • Used to assemble bigger genomes as well
• 16s Metagenome Detection (separate at the moment)
  • 454 and Illumina
Use across sites

• Groups from Ulaval (on colosse) and sherbrooke (mammouth-mp2) are using our pipelines
• We deployed it on
  • Abacus
  • Guillimin
  • Mammouth
  • Colosse (the Ulaval group deployed it)
• Soon to come
  • Briaree
  • SciNet

These installations include all the software as modules, all the reference genomes as well.

The goal is to use CernVMFS from guillimin to automatically deploy all of these on all/most of the Calcul Quebec or even Compute Canada clusters.
Coming up... Additions to the pipelines

• A lot of work from Robert with Gemini and human variant analysis needs to be put in place. This is very promising.

• Our HLA detection methods will be incorporated (targeted and whole genome)

• WGBS pipeline

• miRNA pipeline

• Put in a place a more universal differential analysis framework that can be used across pipelines

• Many, many others...
Links

• Repo
  https://bitbucket.org/mugqic/mugqic_pipelines

• Mailing List/Group
  https://groups.google.com/d/forum/mugqic_pipelines
  mugqic_pipelines@googlegroups.com

• Docs (mostly deprecated)
  http://biowiki.genome.mcgill.ca/

• Bugs
  https://genomequebec.mcgill.ca/jira/browse/BFXDEV
Main Documentation

• Documentation is part of the source now. Simpler to maintain

• We use python docstrings to generate the MarkDown in the README.md

• The file formats, usage, etc is all there now
<table>
<thead>
<tr>
<th>Deployed software</th>
</tr>
</thead>
<tbody>
<tr>
<td>mugqic/GenomeAnalysisTK/1.6-11</td>
</tr>
<tr>
<td>mugqic/GenomeAnalysisTK/2.5-2</td>
</tr>
<tr>
<td>mugqic/GenomeAnalysisTK/2.6-5</td>
</tr>
<tr>
<td>mugqic/GenomeAnalysisTK/2.7-2</td>
</tr>
<tr>
<td>mugqic/GenomeAnalysisTK/3.1-1</td>
</tr>
<tr>
<td>mugqic/GenomeAnalysisTK/3.2-2</td>
</tr>
<tr>
<td>mugqic/MACS/1.4.2</td>
</tr>
<tr>
<td>mugqic/MACS/2.0.10.09132012</td>
</tr>
<tr>
<td>mugqic/MUUMmer/3.23</td>
</tr>
<tr>
<td>mugqic/MUUMmer/3.23b</td>
</tr>
<tr>
<td>mugqic/MUSCLE/3.8.31</td>
</tr>
<tr>
<td>mugqic/MUSCLE/3.8.31</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.1.2</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.2.1</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.1.1</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.1.1</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.1.1</td>
</tr>
<tr>
<td>mugqic/cufflinks/2.1.1</td>
</tr>
<tr>
<td>mugqic/alignment/2.17.0</td>
</tr>
<tr>
<td>mugqic/fastx/0.0.13.2</td>
</tr>
<tr>
<td>mugqic/bedtools/2.2.25+</td>
</tr>
<tr>
<td>mugqic/ghostscript/8.70</td>
</tr>
<tr>
<td>mugqic/ghostscript/8.70</td>
</tr>
<tr>
<td>mugqic/homer/4.1</td>
</tr>
<tr>
<td>mugqic/homer/4.1</td>
</tr>
<tr>
<td>mugqic/homer/4.7</td>
</tr>
<tr>
<td>mugqic/igvtools/2.3.14</td>
</tr>
<tr>
<td>mugqic/java/openjdk-jdk1.6.0_38</td>
</tr>
<tr>
<td>mugqic/java/openjdk-jdk1.7.0_60</td>
</tr>
<tr>
<td>mugqic/java/oracle-jdk1.6.0_38</td>
</tr>
<tr>
<td>mugqic/java/oracle-jdk1.7.0_15</td>
</tr>
<tr>
<td>mugqic/jellyfish/2.1.3</td>
</tr>
<tr>
<td>mugqic/jellyfish/2.1.3</td>
</tr>
<tr>
<td>mugqic/memtime/1.3</td>
</tr>
<tr>
<td>mugqic/memtime/1.3</td>
</tr>
</tbody>
</table>
Deployed Genomes

https://biowiki.atlassian.net/wiki/display/PS/Reference+genomes+and+annotations

Reference genomes and annotations

Created by Johanna Sandoval, last modified by Joel Fillion on Oct 17, 2014

On every cluster, reference genomes and annotations are located in:

$MUGGIC_INSTALL_HOME/genomes/

All species-related files are in:

$MUGGIC_INSTALL_HOME/genomes/species/<speciesScientificName>./<assembly>/

e.g. for Homo sapiens assembly GRCh37, the directory has the following (incomplete) hierarchy:

$MUGGIC_INSTALL_HOME/genomes/species/Homo_sapiens.GRCh37

- annotations
  - Homo_sapiens.GRCh37.dbSNP141.vcf.gz
  - Homo_sapiens.GRCh37.dbSNP141.vcf.gz.tbi
  - Homo_sapiens.GRCh37.Ensembl75.geneid2Symbol.tsv
  - Homo_sapiens.GRCh37.Ensembl75.genes.length.tsv
  - Homo_sapiens.GRCh37.Ensembl75.genes.tsv
  - Homo_sapiens.GRCh37.Ensembl75.gtf
  - Homo_sapiens.GRCh37.Ensembl75.ncrna.fa
  - Homo_sapiens.GRCh37.Ensembl75.transcript_id.gtf
  - Homo_sapiens.GRCh37.Ensembl75.vcf.gz
  - gtf_to_hummed
  - ncRNA_bwa_index

- genome
  - bowtie2_index
  - bwa_index
  - Homo_sapiens.GRCh37.dict
  - Homo_sapiens.GRCh37.fa
  - Homo_sapiens.GRCh37.fa.fai
  - Homo_sapiens.GRCh37.ini

- log

- source
  - ftp.1000genomes.ebi.ac.uk
  - ftp.ensembl.org
  - ftp.ncbi.nih.gov
Model
Model

.Core
  • Independent Pipeline mechanism

.Pipelines
  • MUGQIC Pipeline: muggqc_log() stats
  • Illumina DNA-Seq, RNA-Seq, RNA-Seq De Novo Assembly, ChIP-Seq
  • PacBio Assembly

.BFX
  • Bioinformatics modules
  • Samples, Readsets (Illumina, PacBio)
Semantic versioning

Try to follow semantic versioning guidelines: http://semver.org/

Given a version number MAJOR.MINOR.PATCH, increment the:

- MAJOR version when you make incompatible API changes,
- MINOR version when you add functionality in a backwards-compatible manner, and
- PATCH version when you make backwards-compatible bug fixes.
Config .ini files

• Can use several config .ini files with –config
• base.ini: all defaults for Homo sapiens/abacus <genome>.ini use path interpolation
  • Homo_sapiens.GRCh37.ini
  • Mus_musculus.GRCm38.ini
• <cluster>.ini: guillimin.ini, mammouth.ini
  • cluster_max_jobs=(30000|3000|2048)
• batch.ini: no cluster_cpu, cluster_queue options
Genomes

- Config files in:
  - muggic_pipelines/resources/genomes/config/

- Description

- Genome file organization
  - Indexes
  - VCF/dbSNP
  - Gene length, Symbol
  - Gene Ontologies (BioMart)
Genome Install

Genome install scripts in: mugqic_pipelines/resources/genomes/
  • GENOME_INSTALL_TEMPLATE.sh
  • install_genome.sh
  • Download only once
  • Create indexes only once
  • Possible to overwrite functions for custom download/install
Module Install

- Module install scripts in:
  mugqic_pipelines/resources/modules/
    - MODULE_INSTALL_TEMPLATE.sh
    - install_module.sh
  - Set Archive URL, build commands, module file
  - Install path, permissions, module deployment automated
- Modules mugqic/R_Bioconductor/3.1.2_3.0 and mugqic/mugqic_R_packages/1.0.1 now separated
CVMFS

- Currently under test
- Modules only installed once on guillimin stm0.hpc.mcgill.ca virtual machine
- Modules visible to all clusters within 1 hour
- In $HOME/.bash_profile:
  - export MUGQIC_INSTALL_HOME=/cvmfs/soft.mugqic/CentOS6
What needs to be done

• Daemon Scheduler
  • Control dependencies internally
  • No more # jobs or dependencies limit
  • No more job submission overload
• “Pipe” pipelines or pipeline plugins
  • Separate DNA-Seq alignment and annotations (Human, cancer, etc)
  • Separate RNA-Seq, RNA-Seq De Novo Assembly, ChIP-Seq and differential expression
GenAp as a whole
Acknowledgements

**IT team**
Terrance Mcquilkin  
Marc-André Labonté  
Genevieve Dancausse  
Andras Frankel  
Alexandru Guja

**Galaxy GenAp team**
David Bujold (McGill)  
Bryan Caron (McGill)  
David Morais (UdeS)  
Carol Gauthier (UdeS)  
Alain Veilleux (UdeS)  
ME Rousseau (McGill)

**Analysis team**
Louis Letourneau  
Mathieu Bourgey  
Maxime Caron  
Gary Lévesque  
Robert Eveleigh  
Francois Lefebvre  
Johanna Sandoval  
Pascale Marquis  
Joel Fillon

**Development team**
Nathalie Émond  
Francois Cantin  
Catherine Côté  
Daniel Guertin  
Louis Dumond Joseph  
Francois Korbuly  
Marc Michaud  
Thuong Ngo  
Francois Massé

guil.bourque@mcgill.ca