Genomics Calculations and Applications on Beagle2

- Genomics on Beagle2
  - Installed packages
    - Bamtools
      - Loading bamtools
      - Running bamtools
    - Bedtools
      - Loading bedtools
      - Running bedtools
    - Breakdancer
      - Loading breakdancer
      - Running breakdancer
    - BWA - Burrows-Wheeler Alignment Tool
      - Loading bwa
      - Running bwa
    - Bowtie2
      - Loading bowtie2
      - Running bowtie2
    - CUFFLINKS - Transcript assembly, differential expression, and differential regulation for RNA-Seq
      - Loading cufflinks
      - Running cufflinks
    - The Genome Analysis Toolkit (GATK)
      - Loading GATK
      - Running GATK
    - Picard
      - Loading picard
      - Running picard
    - Pindel
      - Loading pindel
      - Running pindel
    - Samtools
      - Loading samtools
      - Running samtools
    - NCBI SRA-SDK
      - Loading NCBI sra-sdk
      - Running NCBI sra-sdk
    - SGA -- String Graph Assembler
      - Loading sga
      - Running sga
    - Tabix (bgzip)
    - Tophat
      - Loading tophat
      - Running tophat
    - Vcftools
      - Loading vcftools

Genomics on Beagle2

Installed packages

Packages and programs can be uploaded simply by using the module command. To see what versions are available of a program type `module avail packagename`, e.g., `module avail bwa`

(remember to add breakseq and another one I installed and forgot to document)

Bamtools

Information about bamtools can be found, for example, at https://github.com/pezmaster31/bamtools/wiki current versions are

> module avail bamtools
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- /soft/modulefiles/applications
bamtools/2.1.1(default)

Where the default version is indicated.

Loading bamtools

- To use the default version of bamtools:
module swap PrgEnv-cray PrgEnv-gnu
module load bamtools

  • To use the specific version of bamtools (<vers>):

module swap PrgEnv-cray PrgEnv-gnu
module load bamtools/<vers>

Please inform us if a specific version you are interested in is not available

Running bamtools

Applications should be run on the compute nodes using the qsub. Bamtools can be used to perform a number of analysis and there are parameter settings that have to date show to be effective when used on Beagle2.

Note that bamtools requires shared libraries to work, so scripts need to include
export CRAY_ROOTFS=DSL

The scripts set up a number of environmental variables that can be used to link and use the various libraries provided by bamtools:

BAMTOOLS_HOME /soft/bamtools/gnu/2.1.1/usr/local
LD_LIBRARY_PATH /soft/bamtools/gnu/2.1.1/usr/local/lib/bamtools
BAMTOOLS_INCLUDE_OPTS -I/soft/bamtools/gnu/2.1.1/usr/local/include
BAMTOOLS_POST_LINK_OPTS -L/soft/bamtools/gnu/2.1.1/usr/local/lib

Other variables will be added when necessary (or requested).

Bedtools

Loading bedtools

  • To use the default version:

module swap PrgEnv-cray PrgEnv-gnu
module load bedtools

Breakdancer

Information about breakdancer can be found, for example, at http://breakdancer.sourceforge.net/ current versions are

module avail breakdancer

breakdancer/1.1(default)

Where the default version is indicated.

Loading breakdancer

  • To use the default version of breakdancer:

module swap PrgEnv-cray PrgEnv-gnu
module load breakdancer

  • To use the specific version of breakdancer:

module swap PrgEnv-cray PrgEnv-gnu
module load breakdancer/<vers>

Running breakdancer

Applications should be run on the compute nodes using the qsub. Breakdancer can used in cpp version (which will be in your path when you load the module and call breakdancer_max) and in the perl version, which will be available also in the path or can be linked at ${BREAKDANCER_HOME}/perl, where one can find:

AlnParser.pm bam2cfg.pl Poisson.pm.
**BWA - Burrows-Wheeler Alignment Tool**

Information about **bwa** can be found, for example, at [http://bio-bwa.sourceforge.net/bwa.shtml](http://bio-bwa.sourceforge.net/bwa.shtml) current versions are

```
module avail bwa
------------------- /soft/modulefiles/applications ---------------------
bwa/0.5.9 bwa/0.6.1(default)
```

Where the default version is indicated.

**Loading bwa**

- To use the default version of **bwa**:

```
module swap PrgEnv-cray PrgEnv-gnu
module load bwa
```

- To use the specific version of **bwa** (in this instance 5.9):

```
module swap PrgEnv-cray PrgEnv-gnu
module load bwa/0.5.9
```

**Running bwa**

Applications should be run on the compute nodes using the `qsub` command. **Bwa** can be used to perform a number of analyses and there are parameters settings that have to date show to be effective when used on Beagle2.

**Bowtie2**

Information about **bowtie2** can be found, for example, at [http://bowtie-bio.sourceforge.net/bowtie2/index.shtml](http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) current versions are

```
module avail bowtie2
----------------------- /soft/modulefiles/applications ---------------------
bowtie2/2.0.0-beta6(default)
```

Where the default version is indicated.

**Loading bowtie2**

- To use the default version of **bowtie2**:

```
module swap PrgEnv-cray PrgEnv-gnu
module load bowtie2
```

- To use the specific version of **bowtie2** (in this instance <vers>):

```
module swap PrgEnv-cray PrgEnv-gnu
module load bowtie2/<vers>
```

**Running bowtie2**

Applications should be run on the compute nodes using the `qsub` command. **Bowtie2**, `bowtie2-align`, `bowtie2-build`, and `bowtie2-inspect` can be used directly because loading the module adds them to the path. The scripts related to `bowtie2` are found under `$(BOWTIE2_HOME)/scripts`, there one can find also the doc directory and the example directory.

**CUFFLINKS - Transcript assembly, differential expression, and differential regulation for RNA-Seq**

Information about cufflinks, cuffdiff cuffmerge and so on can be found, for example, at [http://cufflinks.cbcb.umd.edu/index.html](http://cufflinks.cbcb.umd.edu/index.html) current versions are:

```
module avail cufflinks
------------------- /soft/modulefiles/applications ---------------------
cufflinks/1.3.0(default) cufflinks/2.0.0
```
Loading cufflinks

- To use the default version of cufflinks:
  ```bash
  module swap PrgEnv-cray PrgEnv-gnu
  module load cufflinks
  ```

- To use the specific version of cufflinks (in this instance <vers>):
  ```bash
  module swap PrgEnv-cray PrgEnv-gnu
  module load cufflinks/<vers>
  ```

Running cufflinks

Applications should be run on the compute nodes using the `qsub`. Cufflinks is part of a software package that includes other analysis tools such as `cuffdiff` and `cuffmerge`, they will all be in the PATH variable when the module is loaded.

The Genome Analysis Toolkit (GATK)

Information about GATK can be found, for example, at [http://www.broadinstitute.org/gsa/wiki/index.php/GATK](http://www.broadinstitute.org/gsa/wiki/index.php/GATK)

* Please notice that loading GATK automatically loads java with `module load java`

Current versions are

```bash
module avail gatk
```

```bash
-----------------------------------------------------------------
soft/modulefiles/applications -------------------------------------
gatk/1.5-32 (default)  gatk/2.2-4
```  

Where the default version is indicated.

Loading GATK

- To use the default version of GATK:
  ```bash
  module load gatk
  ```

- To use the specific version of GATK (in this instance <vers>), we recommend using specific versions in order to avoid surprises when defaults are changed:
  ```bash
  module load gatk/<vers>
  ```

Running GATK

Applications should be run on the compute nodes using `qsub`. GATK can be used to perform a number of analysis via various java jar files. For example to run it inside of a PBS script, you can just use

```bash
aprun <aprun-options> java <java-options> -jar ${GATK_DIR}/GenomeAnalysisTK.jar <GATK-options>
```

Thus, every time you change GATK version you do not need to change the directory you look it into because the

```bash
module load GATK/<vers>
```

will set the variable `$GATK_DIR` automatically to the new location.

To find more information about how to run the GATK, one can execute commands such as:

```bash
java -jar $(GATK_DIR)/GenomeAnalysisTK.jar --help | less
```

or look at the Broadinstitute web site at


Picard

Information about picard can be found, for example, at [http://picard.sourceforge.net/index.shtml](http://picard.sourceforge.net/index.shtml)

* Please notice that loading picard automatically loads java with module load java
Current versions are

```bash
module avail picard
-------------------------------------- /soft/modulefiles/applications --------------------------------------
picard/1.57               picard/1.60               picard/1.67(default)
```

Where the default version is indicated.

**Loading picard**

- To use the default version of picard:

  ```bash
  module load picard
  ```

- To use the specific version of picard (in this instance 1.57):

  ```bash
  module load picard/1.57
  ```

**Running picard**

Applications should be run on the compute nodes using the `qsub`. picard can be used to perform a number of analysis and there are parameter settings that have to date show to be effective when used on Beagle. For example to run `SortSam` inside of a pbs script, you can just use `aprun <aprun-options> java <java-runtime-options> -jar $PICARD_DIR/SortSam.jar <picard-options>`

Please note that you should load the java module on the shell script when running on the compute node.

Thus, every time you change picard version you do not need to change the directory you look it into because the `module load picard/<vers>`

will set the variable `$PICARD_DIR` automatically to the new location.

**Pindel**

Information about pindel can be found, for example, at [https://trac.nbic.nl/pindel/](https://trac.nbic.nl/pindel/) current versions are

```bash
> module avail pindel
-------------------------------------- /soft/modulefiles/applications --------------------------------------
pindel/0.24s(default)
```

Where the default version is indicated.

**Loading pindel**

- To use the default version of pindel:

  ```bash
  module swap PrgEnv-cray PrgEnv-gnu
  module load pindel
  ```

- To use a specific version (`<vers>`):

  ```bash
  module swap PrgEnv-cray PrgEnv-gnu
  module load pindel/<vers>
  ```

Please inform us if a specific version you are interested in is not available.

**Running pindel**

Applications should be run on the compute nodes using the `qsub`. `pindel` `pindel2vcf` `sam2pindel` are set on the path once the module is loaded.

**Samtools**

Information about samtools can be found, for example, at [http://samtools.sourceforge.net/](http://samtools.sourceforge.net/) current versions are
Loading samtools

- To use the default version of samtools:

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load samtools
```

- To use the specific version of samtools (<vers>):

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load samtools/<vers>
```

Please inform us if a specific version you are interested in is not available.

Running samtools

Applications should be run on the compute nodes using the `qsub`. samtools can be used to perform a number of analysis and there are parameter settings that have to date show to be effective when used on Beagle2.

NCBI SRA-SDK

Information about sra-sdk can be found, for example, at http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software current versions are

```
> module avail ncbi-sra-sdk
```

Where the default version is indicated.

Loading NCBI sra-sdk

- To use the default version:

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load ncbi-sra-sdk
```

- To use a specific version (<vers>):

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load ncbi-sra-sdk/<vers>
```

Please inform us if a specific version you are interested in is not available.

Running NCBI sra-sdk

Applications should be run on the compute nodes using the `qsub`. The toolkit can be used to perform a number of operations on various kinds of files, such as sam and fastq files. Libraries are also available (see module show ncbi-sra-sdk, to see some of the options).

SGA -- String Graph Assembler

Information about sga can be found, for example, at https://github.com/jts/sga/

- Please notice that loading sga automatically loads

```bash
module load sparsehash
module load bamtools
module load hoard
```

and requires a GNU environment (if you don't load it it will tell you).

Current versions are
Loading sga

- To use the default version of sga:
  module load sga

- To use the specific version of sga (in this instance <vers>):
  module load sga/<vers>

Running sga

Applications should be run on the compute nodes using the qsub.

Note that sga requires shared libraries to work, so scripts need to include

```bash
export CRAY_ROOTFS=DSL
```

Tabix (bgzip)

Loading Tabix

- To use the default version:
  module swap PrgEnv-cray PrgEnv-gnu
  module load tabix

Tophat

Information about tophat can be found, for example, at http://tophat.cbcb.umd.edu/ current versions are

```
> module avail tophat
----------------------------- /soft/modulefiles/applications -----------------------------
tophat/2.0.1(default)
```

Where the default version is indicated.

Loading tophat

- To use the default version of tophat:

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load tophat
```

- To use the specific version of tophat (in this instance <vers>):

```bash
module swap PrgEnv-cray PrgEnv-gnu
module load tophat/<vers>
```

Running tophat

Applications should be run on the compute nodes using the qsub. Applications related to tophat can be found by typing `ls $TOPHAT_DIR` and can be used directly because loading the module adds them to the path.

Vcftools

Loading vcftools
To use the default version:

```
module swap PrgEnv-cray PrgEnv-gnu
module load vcftools
```

TO DO

Documentation

- Finish vcftools, bedtools, tabix and their related software, add comments about linking and other env variables
- How to use perl modules included
- How to use libraries (e.g., libbam.a with compilers and so on)
- How to run massive calculations based on the various tools and their support software (python, perl, java...)
- How to run complex calculations using swift adding relevant modules that work when packed
- describe the reference genomes available and add material.

Installation

- MACS ([http://liulab.dfci.harvard.edu/MACS/](http://liulab.dfci.harvard.edu/MACS/)).
- Also, MACH ([http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html](http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html))
- As an aside you can use the Broad Institute's GATK [https://www.broadinstitute.org/gatk/guide/tooldocs](https://www.broadinstitute.org/gatk/guide/tooldocs) tool with multi-threading enabled. The differences in the output of the two are likely to be subtle and you will have so many SNPs to look at that you won't have time to worry about the odd FP or FN. Alternatives are running samtools splitting regions (e.g., chromosomes) or parallelizing samtools. When Beagle is up I will (hopefully) make a samtools module and then GATK. We can of course tests them on some files and see how they perform in practice.
- Varscan2