Computational Analysis of RNA-Seq Data - Flemish Super Computer (VSC)

RNA-Seq Pipelines and Computational Analysis – 2020 Workshop

GENOMICS CORELEUVEN

Computer Cluster

NGS data means big data...means big computing power

- Whole Human Genome:
- 30Gb, gzip file

- Exome data:
- 6GB

- RNA-Seq
- 1GB

- NGS data is usually analyzed on a supercomputer or cluster.
  - UZ Leuven: Hydra, Google genomics
  - KU Leuven: VSC Flemish Super Computer

GENOMICS CORELEUVEN

Before we start…

Some resources:

- How to get access?
  - https://www.vscentrum.be/getaccess
- Training:
  - https://www.vscentrum.be/training
    - Linux, Linux for HPC (1st, 13th Oct)
- VSC Documentation:
  - https://vlaams-supercomputing-centrum-vscdocumentation.readthedocs-hosted.com/en/latest/
- High Performance Computing for Genomics
  - https://github.com/GenomicsCoreLeuven/vsc_ngs_workshop/tree/master/presentations

GENOMICS CORELEUVEN

VSC Data Analysis Process

1. Obtain and transfer to VSC raw data:
   - Fastq files

2. Determine pipeline
   - FastQC, Hisat2, MappingQC, HTSeq-count, DEseq2

3. Log into VSC:
   - ssh vsc31420@login1-tier2.hpc.kuleuven.be

4. Are tools available at the VSC?
   - Module av
     - Yes -> PBS Scripts
     - No? -> Contact VSC, Create own Conda environment
5. Create PBS scripts for each step of the pipeline

6. Submit jobs
   - qsub alignment_Hisat2.pbs

7. Download results

VSC Storage System

- Personal Storage:
  - Home Directory:
    - $VSC_HOME
    - /user/leuven/3XX/vsc3XXXX
    - 25GB
    - Personal data
    - Configuration files
  - Data Directory
    - $VSC_DATA
    - /data/leuven/3XX/vsc3XXXX
    - 75GB
    - Input data
  - Scratch Space
    - $VSC_SCRATCH
    - /scratch/leuven/3XX/vsc3XXXX
    - Fast processing temporary data, deleted periodically

PBS System

- At the heart of each HPC computation there is PBS script
  - Specification of the requested resources
  - Input and output data
  - Package and parameters

PBS System: Header

- Header: Specification of the requested resources

```
#PBS -l walltime=20:00:00
#PBS -l mem=30gb
#PBS -l nodes=1:ppn=20:ivybridge
#PBS -M alvaro.cortes@uzleuven.be
#PBS -m eab
#PBS -N fastqc
#PBS -A lp_biogenomics
```

- Walltime: the maximum time the job can run
- Mem: Maximum requested memory
- Number of nodes, processors per node (ppn) and the processor type

PBS System: Input and output Data

```
PROJECT_DIR="/staging/leuven/stg_00019/RNASeq";

SAMPLE_DIR="$PROJECT_DIR/fastq";

SCRATCH_DIR="$VSC_SCRATCH";

OUTPUT_DIR="$PROJECT_DIR/fastqc";
```

PBS System: Defining Pipeline Step

Running FastQC on all fastq input files:

- module load FastQC/0.11.5-Java-1.8.0_77
- fastqc -o $OUTPUT_DIR -t 20 –d $TMP_DIR $files;
  - $OUTPUT_DIR: Results
  - -t 20: number of cores
  - -d $TMP_DIR: Input directory
  - $files: List of fastq input files

PBS System: Submitting PBS job

- Job submission:
  - qsub fastqc.pbs
- Checking job status:
  - qstat –u vsc3xxxx
- Job estimated start:
  - showstart 2321214
- Job overview:
  - checkjob 2321214
- Job stop:
  - qdel 2321214

PBS System: Differential Expression

ONLINE DEMO

Thanks!

GENOMICS CORELEUVEN