Uni.lu HPC School 2019
PS12: Bioinformatics workflows with Snakemake and Conda

Uni.lu High Performance Computing (HPC) Team
S. Peter

University of Luxembourg (UL), Luxembourg
http://hpc.uni.lu
Latest versions available on Github:

UL HPC tutorials:
https://github.com/ULHPC/tutorials

UL HPC School:
http://hpc.uni.lu/hpc-school/

PS12 tutorial sources:
ulhpctraining.rtfd.io/en/latest/bio/snakemake/
1 Introduction

2 Workflow
In this tutorial you will learn how to run a ChIP-seq analysis with the conda package manager and the snakemake workflow engine on the cluster.
**ChIP-seq**

*ChIP-sequencing*, also known as *ChIP-seq*, is a method used to analyze protein interactions with DNA. ChIP-seq combines *chromatin immunoprecipitation* (ChIP) with massively parallel *DNA sequencing* to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest. — *Wikipedia*
ChIP-seq

1. Cross-link protein to DNA
2. Shear DNA strands by sonication
3. Cell lysate
ChIP-seq

Add bead-attached antibodies to immunoprecipitate target protein

precipitate

unlink protein; purify DNA

sequencing

map to genome

By Jkwchui - Cell diagram adapted from LadyOfHats’ Animal Cell diagram. Information based on Illumina data sheet, as well as ChIP and immunoprecipitation articles & references., CC BY-SA 3.0, Link
Conda

- open source package and environment management system
- runs on Windows, macOS and Linux
- quickly installs, runs and updates packages and their dependencies
- easily creates, saves, loads and switches between environments on your local computer
Introduction

Snakemake

- create **reproducible and scalable** data analyses
- workflows described via human readable, Python based language
- seamless scaling to server, cluster, grid and cloud environments, without the need to modify workflow definition
- workflows can entail description of required software, which will be automatically deployed to any execution environment
Summary

1 Introduction

2 Workflow
Workflow Overview

1. Setup the environment
2. Create snakemake workflow
   (a) Mapping
   (b) Peak calling
   (c) Generate bigWig files for visualisation
   (d) Summary rule
3. Cluster configuration for snakemake
   (a) Adjust mapping step to run on multiple threads
   (b) Configure job parameters with cluster.yaml
   (c) Run snakemake with cluster configuration
4. Inspect results in IGV
5. (Optional) Immediately submit all jobs
Snakemake workflow

- **mapping**
  - sample: INPUT-TC1-ST2-D0.12
  - sample: H3K4-TC1-ST2-D0.12

- **peak_calling**
  - sample: TC1-ST2-D0.12

- **bigwig**
  - sample: TC1-ST2-D0.12_control_lambda
  - sample: TC1-ST2-D0.12_treat_pileup

- **all**
Thank you for your attention...

Questions?

http://hpc.uni.lu

High Performance Computing @ uni.lu

Prof. Pascal Bouvry
Dr. Sebastien Varrette
Valentin Plugaru
Sarah Peter
Hyacinthe Cartiaux
Clement Parisot
Dr. Frederic Pinel
Dr. Emmanuel Kieffer

University of Luxembourg, Belval Campus
Maison du Nombre, 4th floor
2, avenue de l’Université
L-4365 Esch-sur-Alzette
mail: hpc@uni.lu

1 Introduction

2 Workflow