AllMine Documentation

Release 1.0

T.Bersez

Jun 13, 2019

Contents:

1 Quick Start guide **3**

1.1 Getting AllMine code . . . 3
1.2 Input files . . . 3
1.3 Sequencing data . . . 3
1.4 Sample sheet . . . 4
1.5 Reference genome and annotation . . . 4
1.6 Bed file . . . 4
1.7 Making your work space ready . . . 4
1.8 Configuration of an AllMine run . . . 4
1.9 Running AllMine . . . 5

2 AllMine outputs **7**

2.1 Quality control reports and trimmed reads . . . 7
2.2 Mapped reads and de novo splicing junctions . . . 7
2.3 Putative SNPs . . . 7
2.4 Non synonymous variant summary . . . 8
2.5 Run Report . . . 8

3 Modules documentation **9**

3.1 fastp; reads presprocessing . . . 9
3.2 FastQC; quality control . . . 9
3.3 BWA index; genome indexing . . . 10
3.4 STAR index; genome indexing . . . 10
3.5 BWA mem; DNAseq read mapping . . . 10
3.6 STAR; RNAseq reads mapping . . . 10
3.7 Varscan; SNP calling . . . 11
3.8 WhatsHap; SNP Phasing . . . 11
3.9 ANNOVAR; SNP annotation . . . 11

4 Dependancies **13**

5 Licence and contact **15**

AllMine is a flexible allele mining pipeline. AllMine can handle various NGS data types (WGS, RRGS, RNAseq, paired/single end layouts. . .) and process all steps from raw reads to de novo SNPs phasing and annotation. Designed for easy usage, AllMine can be use by non bioinformaticians. AllMine is HPC compatible via Slurm.

AllMine was developed at INRA's PACA GAFL unit. For more informations on our lab, please visit our web page.

Follow AllMine development on GitHub.

CHAPTER 1

Quick Start guide

First check that all requiered *Dependancies* are installed. If you wish to use AllMine on a cluster, contact your admnistrator for instalations.

1.1 Getting AllMine code

AllMine code is GitHub hosted. To get the source code using git CLI, use :

```
git clone https://github.com/tbersez/Allmine.git
```

Then, from within AllMine directory, build **AllMine container** using Singularity (admin rights needed) :

```
sudo singularity build AllMine singularity/AllMine_recipe.sr
```

Here, AllMine is ready to run.

1.2 Input files

To perform allele mining, AllMine needs :

- Sequencing data (DNA or RNA)
- A reference genome and annotation
- A bed file with regions (or genes) of interest

1.3 Sequencing data

AllMine supports **RNA and DNA** sequencing, **paired or single end**, with a maximum read lenght of **350bp** (you can submit longer reads but they will be trimmed to the maximum size). Sequencing data must be in **fastq format** and may be gzip compressed.

1.4 Sample sheet

Configuration maker for AllMine use a **csv** describing samples file as input. The csv file must have the flowing headers :

```
filename R1_ext R2_ext platform date(mm/dd/yy)
```

Here is an example:

```
"filename","R1_ext","R2_ext","platform","date(mm/dd/yy)"
"SRR1538456","_1.fastq.gz","_2.fastq.gz","illumina","03/03/03"
"SRR1538457","_1.fastq.gz","_2.fastq.gz","illumina","04/03/03"
"SRR1538484","_1.fastq.gz","_2.fastq.gz","illumina","05/03/03"
```

You can use a spread sheet editor to create this csv file !

1.5 Reference genome and annotation

Reference genome must be provided in one file, in **fasta format**. Annotation can be provided in **gff** or **gtf** format (recommended). When possible, we advice you to download the reference sequence and annotation from curated sources, such as Ensembl.

1.6 Bed file

Regions of interest must be specified using a **bed** file, here is an example :

```
NC_035163.1 25395963 25398308
NC_035168.1 31042453 31045884
NC_035169.1 25941228 25944616
NC_035175.1 3317633 3320503
NC_035177.1 20184932 20187543
```

Be sure that the first column (contigs) match with the ones used in your reference genome.

1.7 Making your work space ready

Place your reference genome and annotation in a common folder. That one must only contain those both files. Place your bed file with regions of interest in an other folder. **AllMine outputs are created where your start the analysis.** Make sure that you have enought space to store all outputs !

1.8 Configuration of an AllMine run

To configure an AllMine run use :

```
./csv_to_yaml.py path/to/sample_sheet.csv
```

Answer the questions the script is asking you to configure your run. Note : the bind path is the path from the **root to your home folder.**

Once done run :

```
./annovar_makebd.py
```

This script build the annotation database of Annovar. It need to done once for each new genome used.

1.9 Running AllMine

We recommend first to do a dry run using the following command. **CORE_NUMBER** must be replaced by the number of cores you wish to use.

```
snakemake -j CORE_NUMBER \
--cluster-config slurm_config.json \
--cluster "sbatch" -n
```

Check the output to ensure that your run is properly configured. If not, return to configuration step to correct errors. If yes, run :

```
snakemake -j CORE_NUMBER \
--cluster-config slurm_config.json \
--cluster "sbatch"
```

AllMine is now running. Depending on how much data you have submited and CORE_NUMBER, the analysis may take from few hours to a few days.

CHAPTER 2

AllMine outputs

AllMine outputs are detailled here. Some of them (such a trimmed reads) are "non-final" outputs and so, should not be conserved after the completion of the run. However if, for any reasons you may need them, you can gather them from the outputs.

2.1 Quality control reports and trimmed reads

During prepossessing, adapters, linker, low quality bases and low complexity regions are trimmed for the raw reads. Trimmed reads are stored in the **trimmed** directory. Thoses reads are then passed to quality checking. Quality control steps are performed by AllMine in order to ensure the quality of the sequencing data submitted. Reports names use the format <sample_name>.html and can be found in the **QC_post_preproc** directory. You can inspect them using your favorite web browser. As the sequencing data quality influence allele mining reliability, quality control report always should be carefully inspected and taken in account during the results interpretation.

2.2 Mapped reads and de novo splicing junctions

Two different strategies of mapping are supported by AllMine. For DNA sequencing inputs, AllMine uses BWA (Burrows-Wheeler Aligner). For RNA sequencing input, STAR (Spliced Transcripts Alignment to a Reference) is used with a two pass strategy. In both cases, mapped and sorted reads are stored in the **mapped** directory. Both full alignments files and parsed around specified regions are conserved, allowing new parsing around other regions without executing the mapping once again. In the case of RNA sequencing data, the sub-directory **STAR_SJ** is created. It contains the de novo splicing junctions discovered by STAR during the first pass of the mapping.

2.3 Putative SNPs

Putative variants called by Annovar can be found in the **variant** directory. Each subfolder correspond to one sample. Annotated, phased and raw variants are displayed.

2.4 Non synonymous variant summary

This is the main AllMine output. **Non_synonymous_variant_summary.tab** is a tabular file displaying all non synonymous SNPs found by AllMine. Base and amino acid reference and variation are indicated as well as genotype (het or hom), location in the gene (exon1, exon2 . . .). If a variant is found in more than one sample, all concerned samples are indicated in the SAMPLE(s) column.

2.5 Run Report

Allmine build an R markdown report to sum up most of the information about the run. This report also include coverage plots for regions of interest.

CHAPTER 3

Modules documentation

AllMine is developed in a **modular fashion**. Current AllMine modules are documented here. Some of them include **tweakable parameters** ! Only key steps modules are documented here.

3.1 fastp; reads presprocessing

AllMine uses fastp v0.20.0 to perform presprocessing on submited reads. Two modules, **fastp_pe** and **fastp_se** have been implemented to handle paired end and single end reads respectively.

Parameters :

- `--correction` : Enable paired end overlaping regions correction.
- `--cut_mean_quality` : Set to 20. Threshold for trimming in **both 5' and 3'**. You can increase this number for a more stringeant quality trimming.
- `--cut_window_size` : Size of the sliding window for sliding window trimming. Increasing this number will relax the trimming. Set to 1 for per base trimming (not recommended).
- `--complexity_threshold` : Set to 30. Complexity is defined as **P(base[i] != base[i+1])**. Reads bellow the threshold are discarded. Usefull to filter polyA tails in RNA seq data !
- `-w` : Threads used.
- `--max_len1 350` : Maximum lenght of submited reads. Longer reads will be trimmed from 3' end to the maximum size.

Please refer to the fastp manual for more informations.

3.2 FastQC; quality control

AllMine uses FastQC v0.11.8 to perform quality control on submited reads. Two modules, **fastqc_pe** and **fastqc_se** have been implemented to handle paired end and single end reads respectively. Numerous indices and statistics are computed by FastQC. You may inspect them to ensure your data quality.

3.3 BWA index; genome indexing

If you have submited DNAseq data, AllMine will index your reference genome using the `bwa index` command from BWA v0.7.17.

Parameters :

- `-a is` : Algorithm used to index the genome. **IS** does not work for large genome (size > 3Gbp). Switch to `-a bwtsw` if needed.

3.4 STAR index; genome indexing

If you have submited RNAseq data, AllMine will index your reference genome using the `STAR --runMode genomeGenerate` command from STAR v2.7.0f. To perfom a two pass mapping strategy

3.5 BWA mem; DNAseq read mapping

If you have submited DNAseq data, AllMine will map your reads using the `bwa mem` command from BWA v0.7.17. Two modules, **bwa_pe** and **bwa_se** have been implemented to handle paired end and single end reads respectively. This mapping algorithm handle reads from 70bp to 1Mbp.

Parameters :

- `-w` : Set to 100 bp. Band width. Essentially, gaps longer than INT will not be found.
- `-d` : Set to 100. Z-dropoff. Avoids unnecessary extension, but also reduces poor alignments inside a long good alignment.
- `-r` : Set to 1.5. **Larger value yields fewer seeds, which leads to faster alignment speed but lower accuracy**.
- `-B` : Set to 4. Mismatch penalty.
- `-O` : Set to 6. Gap opening penalty.
- `-E` : Set to 1. Gap extention penalty.
- `-L` : Set to 5. When performing SW extension, BWA-MEM keeps track of the best score reaching the end of query.

Note : In most of the cases BWA parameters are well suited for balance between accuracy and computational cost.

3.6 STAR; RNAseq reads mapping

If you have submited RNAseq data, AllMine will map your reads using `STAR` command from STAR v2.7.0f. AllMine apply a two pass mapping strategy. Two both handle paired and single end data four modules where developped **star_pe_FP**, **star_pe_SP**, **star_pe_FP** and **star_pe_SP**.

Parameters :

- `--scoreGap` : Set to 0. Gap penalty.
- `--scoreGapNoncan` : Set to -8. Non-canonical junction penalty.
- `--scoreGapGCAG` : Set to -4. GC/AG and CT/GC junction penalty.
- `--scoreGapATAC` : Set to -8. AT/AC and GT/AT junction penalty.

- `--scoreGenomicLengthLog2scale` : Set to -0.25. Extra score logarithmically scaled with genomic length of the alignment : scoreGenomicLengthLog2scale*log2(genomicLength).
- `--scoreDelOpen` : Set to -2. Deletion open penalty.
- `--scoreDelBase` : Set to -2. Deletion extension penalty per base (in addition to scoreDelOpen).
- `--scoreInsOpen` : Set to -2. Insertion open penalty.
- `--scoreInsBase` : Set to -2. Insertion extension penalty per base (in addition to scoreInsOpen).
- `--scoreStitchSJshift` : Set to 1. Maximum score reduction while searching for SJ boundaries in the stitching step.
- `--runThreadN` : Set to 10. Number of threads used per jobs.

Note : STAR include numerous parameters, please read the manual for more informations.

3.7 Varscan; SNP calling

AllMine use Varscan v2.4.3 to perform SNP calling on aligned reads.

Parameters :

- `--p-value` : Set to 0.99. We did not choosed to use p value as filter for SNP calling. However you can tune this parameter if wanted.
- `--min-coverage` : Set to 8. Minimal deep at SNP loci for calling.
- `--min-var-freq` : Set to 0.15. Minimal variant frequency at SNP loci for calling.
- `--min-avg-qual` : Set to 20. Minimal sequencing quality at SNP loci for calling.

Note : The tunning of Varscan parameters is a trade-off between sensitivity and specificity. It should be keeped in mind when analysing your results!

3.8 WhatsHap; SNP Phasing

AllMine use WhatsHap v0.18 to perform phasing of called SNPs.

Note : 330 Bp long paired end reads at 15X depth are required for haplotypes chunks of 300 Kb in average.

3.9 ANNOVAR; SNP annotation

AllMine use SnpEff v2018-04-16 00:43:31 to annotate called SNPs. A gene base annotation is done by default.

CHAPTER 4

Dependancies

- Python3 >= V3.5.3
  - including modules csv and yaml
- Snakemake >= V4.8.0
- Slurm >= V0.4.3
- Singularity >= V2.5.1

Admin rights are needed to build the AllMine's Singularity container.

CHAPTER 5

Licence and contact

AllMine is distributed under the MIT License

Questions? Please mail to thomasbersez@gmail.com with AllMine as object.