GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Protocol WP2 T2.2
Analysis of RRBS data

Authors: Jani de Vos(WU), Ole Madsen (WU)

Workpackage: WP2

Version: 1.0

Protocol associated with Deliverable(s):	D2.1, D2.2, D2.3
Submission date to FAANG:	30/11/2021
Means of verification:	N/A

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 817998

Table of contents

1 Summary

GENE-SWitCH aims at identifying functional elements located in the genomes of pig and chicken working on seven different tissues at three different developmental stages. A multitude of genomic marks regulate gene expression, these marks work together in a complex manner to facilitate correct gene expression during a whole life cycle. These marks are: DNA methylation, protein modifications and non-coding RNA molecules. DNA methylation is the addition of a methyl group to cytosine which is typically located 5' from a guanine, which creates CpG sites throughout the genome in vertebrates. Bisulphite sequencing, is a method whereby unmethylated cytosine residuals are converted to uracils and methylated cytosines remain unchanged identifying methylation across the genome. Site specific DNA methylation changes are detected through this methodology, however application of this method on a whole genome scale is costly. Reduced representation bisulphite sequencing (RRBS) applies genome wide DNA methylation analysis with reduced sequencing (2% of the genome).

The GSM-pipeline (GENE-SWitCH project methylation analysis) is based on the pre-existing methylseq nf-core pipeline (https://nf-co.re/methylseq/1.6.1) for analysis of methylation data to which the GENE-SWitCH team has added extensions (https://github.com/FAANG/GSM-pipeline).

Processes in this pipeline include:

1. **Quality control**: FastQC (https://github.com/s-andrews/FastQC)
2. **Trimming**: TrimGalore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)
3. **Alignment**: Bismark(https://github.com/FelixKrueger/Bismark) or bwa-meth (https://github.com/brentp/bwa-meth)
4. **Methylation calling**: Bismark methyl caller, CGmaptools (https://github.com/guoweilong/cgmaptools), MethylDackel with bwa-meth aligner (https://github.com/dpryan79/methyldackel)
5. **Downstream analysis**: CGmaptools (https://github.com/guoweilong/cgmaptools), MethylKit (https://bioconductor.org/packages/release/bioc/html/methylKit.html)
6. **Results output**: MultiQC (https://github.com/ewels/MultiQC)

2 Protocol description

2.1 Running the pipeline for analysis of GENE-SWitCH RRBS data

Pipeline can be downloaded from Github (https://github.com/FAANG/GSM-pipeline), and requires installation of newest version of nextflow, together with either singularity or docker. The pipeline (version 1.0) was launched in the following way for analysis of GENE-SWitCH single end raw reads:

```
nextflow run GSM-pipeline/ -profile singularity --single_end --input {raw_reads}.fq.gz --fasta {reference}.fa --aligner bismark --rrbs
```

2.2 Compiling bed output

Methylation calling results can be found in results/cgmaptools/cgmap_methyl_call/ directory and the output {name}.CGmap.gz file contains methylation levels per site. The GENE-SWitCH RRBS data had multiple samples per tissue per development stage which were processed individually. Output CGmap.gz files contain the methylation calls for each sample and these individual samples were then merged into one file per tissue per development stage using CGmaptools (https://github.com/guoweilong/cgmaptools) merge2 cgmap function. Thereafter the file was compiled into a bed file, and filtered on a minimum of 10 reads per site, this parameter excludes non-informative reads.

Output BED files contains the following information:

1. Informative header
2. Columns:
    1. CHROM=Chromosome
    2. POS=Position of RRBS site
    3. POS=Position of RRBS site +1
    4. NUC=Nucleotide on reference genome
    5. CONT=Context
    6. DINUC=Dinucleotide context
    7. METH=Methylation level
    8. MC=Counts of reads that support methylated cytosine
    9. NC=Counts of reads that support all cytosines