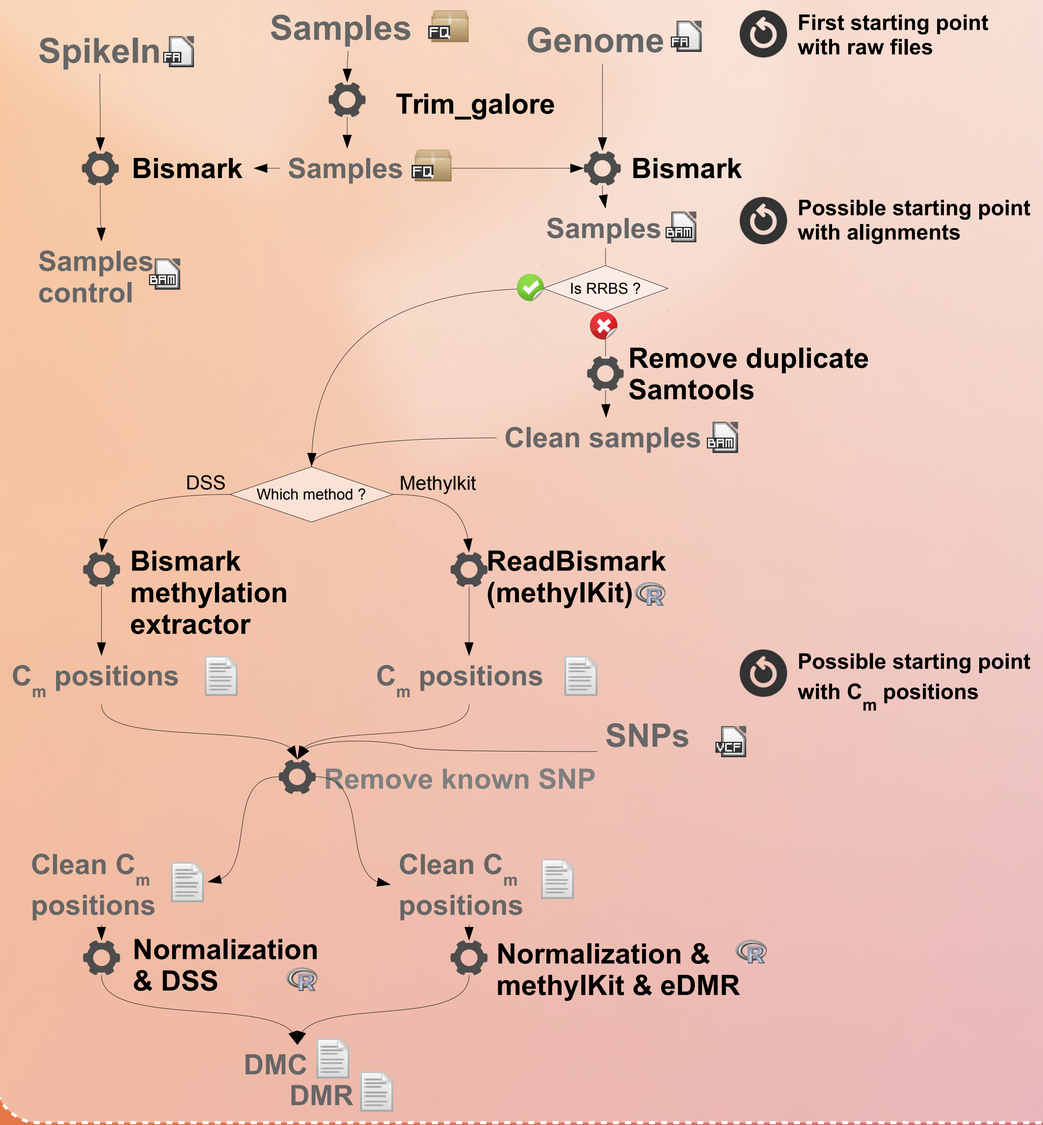
Workflow-BS: an integrative workflow for RRBS and WGBS data.

Gaëlle Lefort, Marjorie Mersch, Sylvain Foissac, Frédérique Pitel, Nathalie Villa-Vialaneix, Céline Noirot

DNA methylation is an epigenetic mark with regulatory roles in a broad range of biological processes and diseases. Sequencing technologies now allow genome-wide methylation studies, at a high resolution and with possibly a large number of samples. Here we present our hands-on workflow that handles single or paired reads from Reduced Representation Bisulfite Sequencing (RRBS) or Whole Genome Bisulfite Sequencing (WGBS) and goes through bioinformatics and biostatistics steps to identify Differentially Methylated Cytosines (DMCs) and Regions (DMRs).

Pipeline description



Pipeline features

The pipeline is based on Jflow [1] and can be launched through the command line with parameters or a config file. It has been designed to ease parallelization on a cluster and can be run on most of core facilities (SGE, Condor, Local …).
The pipeline can process raw fastq reads or be (re-)launched from intermediate results.

Bioinformatics part

Based on the literature and our evaluation of existing tools, we decided to use the main standard packages: Trim_galore for cleaning, Bismark [2] with bowtie 1 or 2 for reads mapping and samtools rmdup to remove PCR duplicates (and singlets reads if paired library). Single or paired libraries can be handled, as well as RRBS or WGBS runs.

Biostatistics part

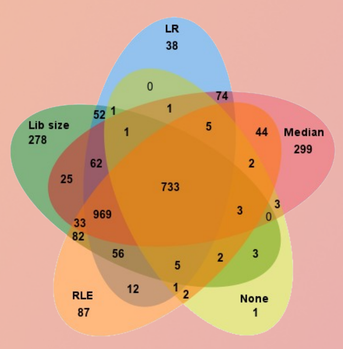
The user can analyze data from all contexts and set the coverage threshold. Two kinds of analyses are proposed, depending on which method the user wants to use for identifying DMCs:

- ✔ a version based on the R packages methylKit [3] and eDMR: standard steps available in this package have been implemented (median normalization → filter on coverage → merge of both strands → identification of DMCs based on a logistic regression → identification of DMRs. Most of the steps are not mandatory.
- ✔ a version based on the R package DSS [4]: two R scripts have been developed, one for normalizing raw coverage data thanks to one of the 5 available methods, the second for identifying DMCs and DMRs with or without replicates (beta-binomial model).

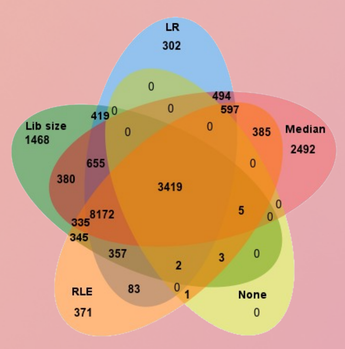
Results

Differential analysis of 4 male vs. 4 female chicken embryos sequenced by HiSeq3000 and analyzed with our pipeline.

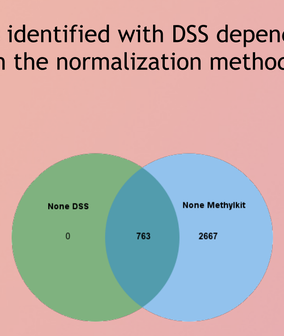
Normalization and DMC comparison



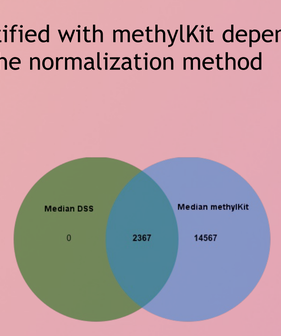
DMCs identified with DSS depending on the normalization method



DMCs identified with methylKit depending on the normalization method



DMCs identified with methylKit and DSS without normalization

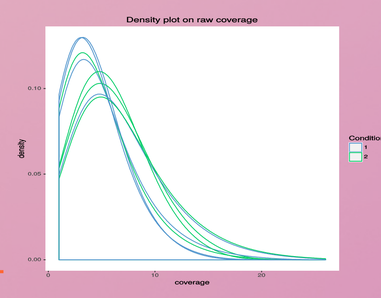


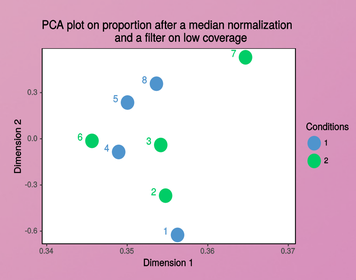
DMCs identified with methylKit and DSS with median normalization

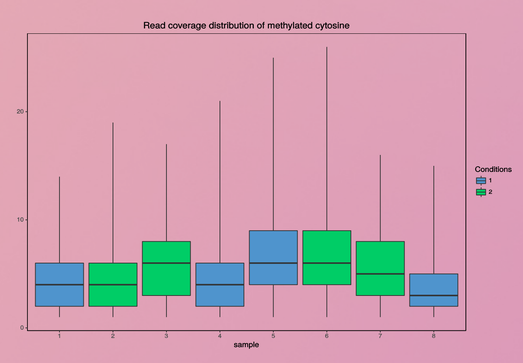
Output files

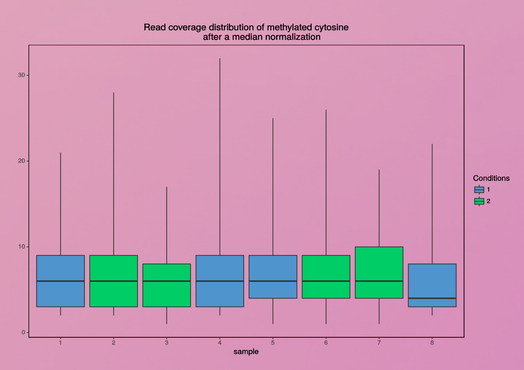
For each step, user has access to result files and log files with all performed command lines.

Graphics









Availability

http://mulcyber.toulouse.inra.fr/projects/jflow-toolshed/

Upcoming ...

- ✔ Web server
- ✔ New aligners
- ✔ DSS statistics
- ✔ More graphics

References

[1] Jérôme Mariette, Frédéric Escudié, Philippe Bardou, Ibouniyamine Nabihoudine, Céline Noirot, Marie-Stéphane Trotard, Christine Gaspin, Christophe Klopp (2015) Jflow: a workflow management system for web applications. *Bioinformatics*. DOI: 10.1093/bioinformatics/btv589
[2] Krueger F, Andrews SR. Bismark (2011) a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. Jun 1;27(11):1571-2
[3] Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick,A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13, R87.
[4] Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic acids research, 42(8), e69-e69.

INRA SCIENCE & IMPACT
Plate-forme Bio-info - Tel 33(0)5 61 28 52 82
INRA Auzeville - Chemin de Borde Rouge BP 52627 31326 Castanet.
anim.bioinfo@toulouse.inra.fr
http://bioinfo.genotoul.fr