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

Céline Noirot

INRA - Toulouse - France

April 21, 2016

Summary

1. Pipeline overview

2. Bioinformatics steps
   - Quality control and cleaning
   - Alignment
   - Methylation extraction

3. Biostatistics steps
   - Preprocessing: normalization and filter on coverage
   - Identification of DMCs
   - Identification of DMRs

4. Conclusion and perspective

Pipeline overview

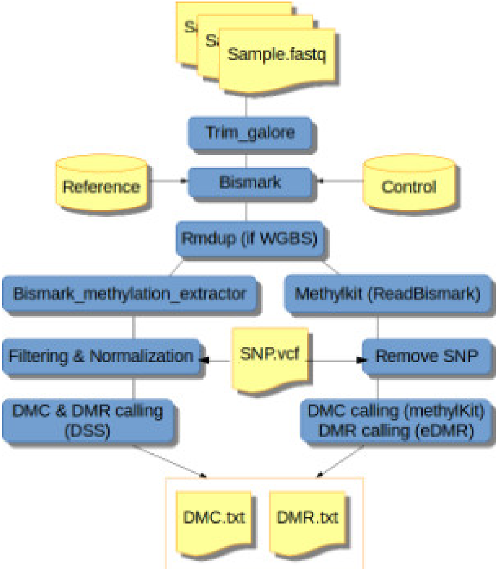
Cleaning

Alignment

Methylation extraction

Preprocessing

DMC and DMR detection



Supported data

- Single or paired reads
- Protocol
  - WGBS: Whole Genome Bisulfite sequencing
  - RRBS: Reduced Representation Bisulfite sequencing
- Input files format :
  - fastq files from illumina sequencing
  - bam files (bismark)
  - methylation calling file (methylkit)

Data from epibird project

- 4 male vs. 4 female chicken embryos
- Sequenced by HiSeq3000
- Whole Genome Bisulfite sequencing

Summary

1 Pipeline overview

2 Bioinformatics steps

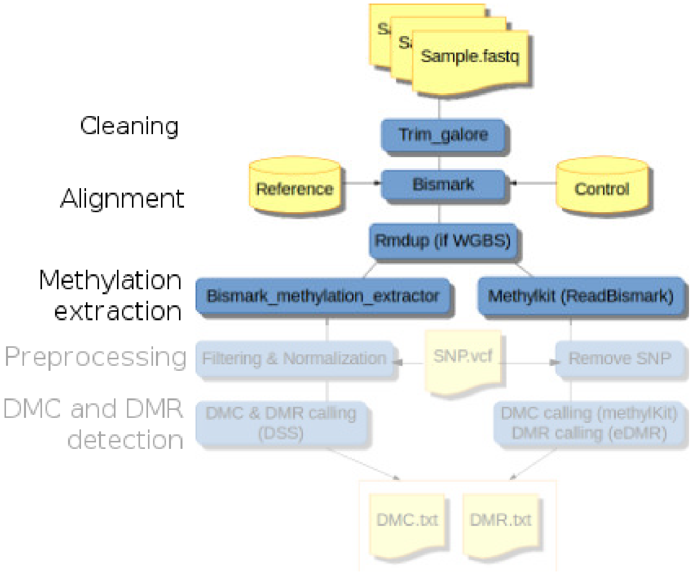
- Quality control and cleaning
- Alignment
- Methylation extraction

3 Biostatistics steps

- Preprocessing: normalization and filter on coverage
- Identification of DMCs
- Identification of DMRs

4 Conclusion and perspective

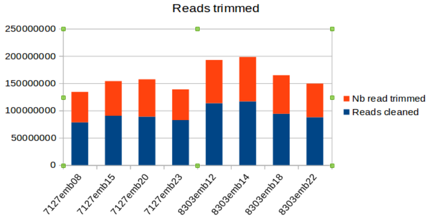
Bioinformatics steps

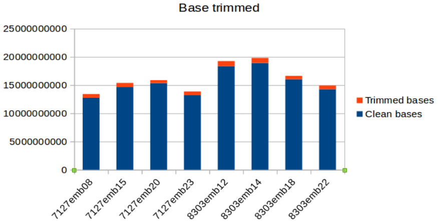


Quality control and cleaning

- Trim adapters
- Trim bad quality

Software: Trim_galore



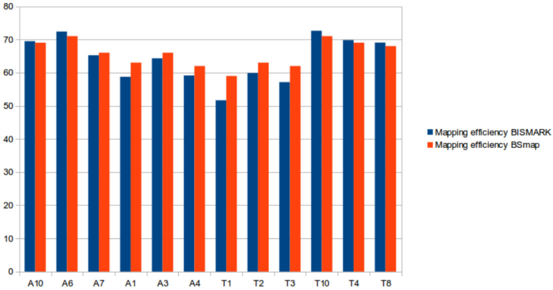


- Reads: about 40% of reads are trimmed
- Bases: about 4% of total bases

Alignment

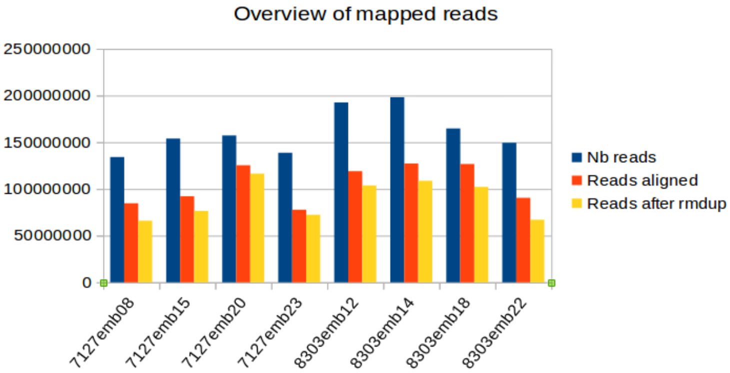
- wild-card alignment
  BSMAP, GSNAP, Last, Pash, RMAP, RRBSMAP, segemehl...
- 3-base encoding
  Bismark, BRAT, BS-Seeker, MethylCoder...

Comparison of 2 strategies (on our data):



M2 Bioinfo 2014 - Julien Plennecassagne - Projet Epiterm

Alignment: Epibird results



- Bismark: 61 to 81% of mapping efficiency
- Rmdup: 73 to 93% of reads kept after rmdup

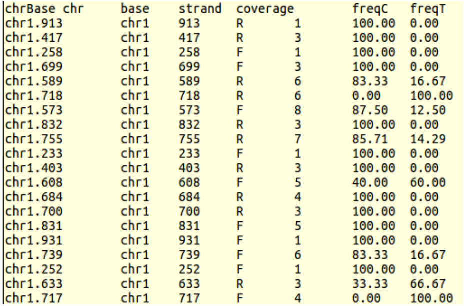
Per base methylation extraction

Per sample extract :

1. C in specific context (CpG, CHG, CHH)

2. choose coverage threshold

Existing software: **methylKit**, bismark_methylation_extraction ...



chrBase	chr	base	strand	coverage	freqC	freqT
chr1.913	chr1	913	R	1	100.00	0.00
chr1.417	chr1	417	R	3	100.00	0.00
chr1.258	chr1	258	F	1	100.00	0.00
chr1.699	chr1	699	F	3	100.00	0.00
chr1.589	chr1	589	R	6	83.33	16.67
chr1.718	chr1	718	R	6	0.00	100.00
chr1.573	chr1	573	F	8	87.50	12.50
chr1.832	chr1	832	R	3	100.00	0.00
chr1.755	chr1	755	R	7	85.71	14.29
chr1.233	chr1	233	F	1	100.00	0.00
chr1.403	chr1	403	R	3	100.00	0.00
chr1.608	chr1	608	F	5	40.00	60.00
chr1.684	chr1	684	R	4	100.00	0.00
chr1.700	chr1	700	R	3	100.00	0.00
chr1.831	chr1	831	F	5	100.00	0.00
chr1.931	chr1	931	F	1	100.00	0.00
chr1.739	chr1	739	F	6	83.33	16.67
chr1.252	chr1	252	F	1	100.00	0.00
chr1.633	chr1	633	R	3	33.33	66.67
chr1.717	chr1	717	F	4	0.00	100.00

Summary

1. Pipeline overview
2. Bioinformatics steps
   - Quality control and cleaning
   - Alignment
   - Methylation extraction
3. Biostatistics steps
   - Preprocessing: normalization and filter on coverage
   - Identification of DMCs
   - Identification of DMRs
4. Conclusion and perspective

Steps

1. Normalization and filter on coverage
2. Identification of differentially methylated cytosines (DMCs)
3. Identification of differentially methylated regions (DMRs)

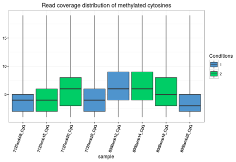
Gaelle Lefort - Biostatistician - 2016

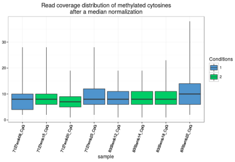
Step 1: Preprocessing on methylation data

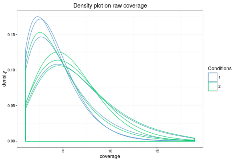
1. Remove known SNPs
2. Remove bases with a very high read coverage
3. Normalize read coverage of each cytosine (5 methods: libsize, median, upper-quartile, RLE and LR)
4. Remove bases with a very low read coverage (a minimum coverage of 5x is recommended)

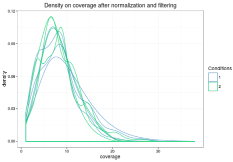
Current version : median normalization of methylKit

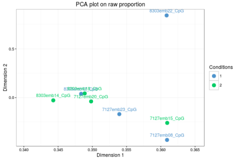
Normalization diagnostics plots

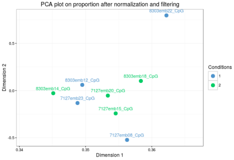

Read coverage distribution of methylated cytosines

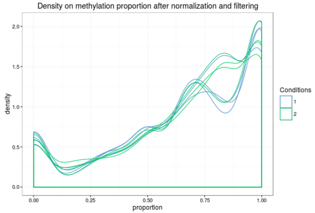

Read coverage distribution of methylated cytosines after a median normalization


Density plot on raw coverage


Density on coverage after normalization and filtering


PCA plot on raw proportion


PCA plot on proportion after normalization and filtering


Density on methylation proportion after normalization and filtering

Step 2: identification of DMCs

Used methods

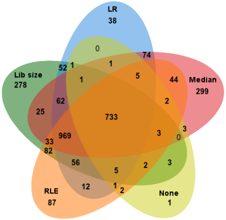
🙁 Fisher exact test: all replicates are pooled (methylKit without replicates)

😐 Logistic regression: the hypothesis is that all data is from the same distribution (methylKit)

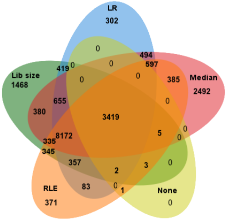
🙂 Beta-binomial model: take into account of the biological variability between samples (DSS)
Case without replicats: nearby cytosines can be used to estimate variability (DSS)

🙂 *Hidden Markov model: take into account of the spatial correlation between nearby cytosines*

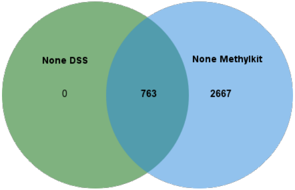
Comparison of normalization methods



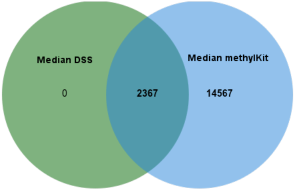
DSS DMCs depending on the normalization method



methylKit DMCs depending on the normalization method



methylKit and DSS DMCs without normalization



methylKit and DSS DMCs with median normalization

- Normalization enable to detect more DMC
- All DMC found by DSS are found by methylkit

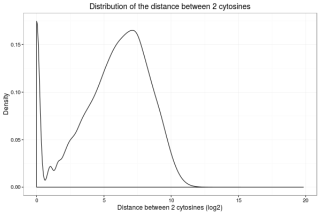
Step 3: identification of DMRs

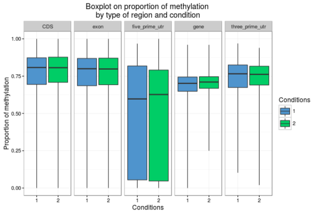
Used methods

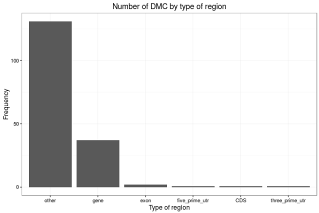
- Sliding windows or predefined regions (MethylKit, DMRcaller. . .)
- From results on DMCs (DSS, eDMR. . .)
- With a hidden Markov model (Bisulfighter)

Current version : eDMR

Differential analysis results and other plots







If annotation file is provided, the pipeline plot DMC categorization. If TSS file is provided, the pipeline plot methylation level around TSS per sample.

Summary

1 Pipeline overview

2 Bioinformatics steps
- Quality control and cleaning
- Alignment
- Methylation extraction

3 Biostatistics steps
- Preprocessing: normalization and filter on coverage
- Identification of DMCs
- Identification of DMRs

4 Conclusion and perspective

Pipeline conclusion

- based on Jlow: error recovery, use most HPC (SGE, Condor, ...), extensible
- configuration with one config file
- include all steps (bioinfo and biostats) within a single command line
- re-runable after main step : alignment and methylation extraction

Available :

- github FAANG consortium `https://github.com/FAANG/faang-methylation/tree/master/workflowbs`
- mulcyber: `https://mulcyber.toulouse.inra.fr/plugins/mediawiki/wiki/jflow-toolshed/index.php/Accueil`

Coming soon ...

- New aligners ?
- Several normalizations method
- DMC and DMR detection with DSS
- A web server

Acknowledgement

- Nathalie Villa-Vialaneix (MIAT)
- Frédérique Pitel, Gérald Salin, Sylvain Foissac, Marjorie Merch, Julien Plennecassagne, Diane Esquerre (GenPhySE)
- Erika Sallet, Pascal Gamas (LIPM)
- Monique Falières (Genotoul)