BS-seq Methods

Pablo Cingolani

1 Methods

1.1 Sequencing Analysis

All sequencing was performed using Illumina Genome Analyzer GAIIx with a Paired End Cluster Generation Kit. Image analysis, base calling and sequence extraction was performed using standard Illumina Pipeline v1.6 software.

1.2 Bisulphite Sequencing Analysis: MoreBs

We performed bisulphite sequenceing (BS-Seq) on two different types of bees: Africanized honey bees (AHB) and European honey bees (EHB). The number of lanes sequenced was 11 (pair end) for AHB and 8 pair-end for EHB. Obtaining 240 millon reasd for AHB and 317 millon reads EHB. Reads were 76 base-pair long yielding a total of 18.2 giga bases and 24.1 giga bases for AHB abd EHB respectively.

The reference genome was Amel2, which is 228M bases long. Read mapping and downstream analysis was done using our in-house system (MoreBs version 1.0) which uses either BWA[7][8] or Bowtie[5] for read alignment. Both alignment programs are based on Burrows-Wheeler transform[3, 10] and create *SAM*[9] output format. There are other tools and methods based on similar approaches [12, 14, 13]. In this case BWA was selected as main mapping method in order to have better alignment near insertions and deletions [1].

MoreBs perform methylation calls by invoking *Samtools* to create *BAM* files, *sorted BAM* files *mpileup* and finally *VCF* files. In this last step methylation calls are produced using an MAQ[11] probabilistic model. It must be noted that BAQ[6] model is explicitly disabled by MoreBs, since some of the its assumptions do not apply for methylation calls. Finally *BcfTools* package is invoked to produce methylation calls in *VCF4.1*[4] format.

MoreBs was set to filter out low quality ($Q < 20$) methylation calls. After all mapping and filtering steps, the mean coverage was 20.8 and 27.4 for AHB and EHB respectively.

As a final step, MoreBs performs several statistics on methylation as well as ranking of hypo-methylated and hyper-methylated genes by means of Fisher exact test. Multiple testing is corrected using False Discovery Rate methodology[2]. Some additional statistics were carried out using custom programs in R programming language (www.r-project.org).

1.3 mDip Sequencing Analysis

Methylated DNA immunoprecipitation followed by sequenceing (mDip-Seq or MeDip-Seq) was performed for a total of 41.3 millon 76 bases long reads. Reads were aligned using BWA and SamTools. Peak calling was performed using MACS [15] 1.4 beta version.

2 References

References

[1] S. Bao, R. Jiang, W.K. Kwan, B.B. Wang, X. Ma, and Y.Q. Song. Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 2011.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] M. Burrows and D.J. Wheeler. A block-sorting lossless data compression algorithm. 1994.

[4] Broad Institute. VCF (Variant Call Format) version 4.0. *http://www.1000genomes.org/wiki/Analysis/vcf4.0*, 2010.

[5] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[6] H. Li. Mathematical Notes on SAMtools Algorithms. *http://lh3lh3.users.sourceforge.net/download/samtools.pdf*, 2010.

[7] H. Li and R. Durbin. Fast and accurate short-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(5), 2009.

[8] H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589, 2010.

[9] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078, 2009.

[10] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473, 2010.

[11] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11):1851, 2008.

[12] C. Pao-Yang, C. Shawn, and P. Matteo. BS Seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11.

[13] P.S. Samarakoon. Epigenomics and genome wide methylation profiling. *Sri Lanka Journal of Bio-Medical Informatics*, 1(1):53–62, 2009.

[14] Y. Xi and W. Li. Bsmap: whole genome bisulfite sequence mapping program. *BMC bioinformatics*, 10(1):232, 2009.

[15] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoute, D.S. Johnson, B.E. Bernstein, C. Nussbaum, R.M. Myers, M. Brown, W. Li, et al. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.