Sequence analysis

B-SOLANA: an approach for the analysis of two-base encoding bisulfite sequencing data

Benjamin Kreck¹,*, George Marnello², Julia Richter³, Felix Krueger⁴, Reiner Siebert³ and Andre Franke¹

¹Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany, ²Life Technologies, Advanced Sequencing Applications, Carlsbad, CA 92008, USA, ³Institute of Human Genetics, Christian-Albrechts-University, Kiel, Germany and ⁴Bioinformatics Group, The Babraham Institute, CB22 3AT Cambridge, UK

Associate Editor: Alfonso Valencia

ABSTRACT
Summary: Bisulfite sequencing, a combination of bisulfite treatment and high-throughput sequencing, has proved to be a valuable method for measuring DNA methylation at single base resolution. Here, we present B-SOLANA, an approach for the analysis of two-base encoding (colorspace) bisulfite sequencing data on the SOLID platform of Life Technologies. It includes the alignment of bisulfite sequences and the determination of methylation levels in CpG as well as non-CpG sequence contexts. B-SOLANA enables a fast and accurate analysis of large raw sequence datasets.

Availability and implementation: The source code, released under the GNU GPL v3 licence, is freely available at http://code.google.com/p/bsolana/.

Contact: b.kreck@ikmb.uni-kiel.de

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on September 1, 2011; revised on November 17, 2011; accepted on November 24, 2011

1 INTRODUCTION
Methylation at position 5 of cytosines is a major epigenetic modification, which has an important impact on transcriptional and regulatory processes of DNA (Holliday, 1975). It is a stable modification of the genome which can be inherited from one generation to the next, even though it can also be dynamically changed by environmental influences. There are several methods based on high-throughput sequencing, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq), methylated DNA capture by affinity purification (MethylCap-seq) and BS-Seq, which can provide good-quality genome-wide DNA methylation data (Bock, 2010).

Methods that currently provide genome-wide methylation patterns at single base resolution make use of bisulfite conversion and high-throughput sequencing. The treatment of DNA with sodium bisulfite has no effect on methylated cytosines, but it specifically converts unmethylated cytosines to uracils, which are converted to thymines during subsequent polymerase chain reaction amplification. As a result of bisulfite conversion, the Watson and Crick strands of bisulfite-treated DNA are no longer complementary and to each other, they become essentially different genomes. This fact leads to an enlarged alignment reference space. The prevalence of T's that have replaced C's leads to reduced complexity in bisulfite sequences, which increases the bioinformatics challenge of BS-Seq analysis. Bisulfite tools for BS-Seq have generally fallen into two categories: (i) methylation-aware alignment tools, which consider cytosines and thymines as potential matches to genomic cytosine positions and (ii) tools which convert any residual cytosines in bisulfite sequences and all cytosines of the reference genomes into thymines.

2 COLORSPACE BISULFITE SEQUENCING
Due to the two-base encoding of SOLID sequencing, in silico conversions of any residual bisulfite read cytosines into thymines, which can be carried out in basespace data to avoid bisulfite-mismatches during alignment, cannot be performed on bisulfite colorspace sequences, because sequencing errors would lead to the incorrect translation of colorspace to basespace (Supplementary Fig. 1). There are ways to align bisulfite colorspace sequences with methylation-aware alignment approaches, which convert bisulfite colorspace sequences to basespace and index all theoretically possible alignments by creating a hash table. Such an approach is implemented in SOCS-B, which is based on the iterative version of the Rabin–Karp algorithm (Ondov, 2010). Even though SOCS-B turns out to be an accurate tool for the analysis of colorspace BS-Seq datasets, it becomes very computationally intensive for complex genomes such as the human genome (∼150,000 CPU hours for the analysis of 500 Million sequences). Therefore, it is not efficient for huge datasets like those produced in genome-wide methylation analyses with average coverage depths ≥10X and genome size ≥1000MB.

Here, we present B-SOLANA, a tool which performs sequence alignment and methylation calling for colorspace bisulfite sequencing. It is based on the established short-read aligner Bowtie (Langmead, 2009) and SAMtools utilities for manipulating alignments (Li, 2009). B-SOLANA is divided into four individual steps: (i) indexing, (ii) mapping, (iii) determination of best alignment and (iv) methylation calling.

The idea of B-SOLANA is to use Bowtie to uniquely align bisulfite sequences to two different conversions of the reference genome and determine best alignments from the combined set of results. The analysis of whole methylomes of 23 eukaryotic organisms shows a variable percentage of methylation at CpG...
Table 1. The 485,990,920 SOLiD BS-Seq reads (50 bp), taken from SRR204024 (Hansen, 2011), were analyzed with B-SOLANA and MethylCoder (one mismatch allowed) B-SOLANA exhibits a high correlation with the results of Hansen et al.  

|                  | Hansen et al. | B-SOLANA | MethylCoder|
|------------------|--------------|----------|------------|
| Uniquely mapped reads (%) | 37.83 | 49.84 | 19.23 |
| CpG positions: % C | 69.84 | 72.83 | 67.07 |
| CpG positions: % T | 30.03 | 26.97 | 32.93 |
| Non-CpG positions: % C | 0.20 | 0.22 | 0.69 |
| Non-CpG positions: % T | 99.76 | 99.70 | 99.31 |

*Including post-processing quality control.

A further approach for the analysis of colorspace BS-Seq was published with the tool MethylCoder (Pedersen, 2011). MethylCoder applies a conversion of any residual bisulfite read cytosines into thymines, which leads to erroneous alignments, as discussed above. Therefore, we compared B-SOLANA and MethylCoder (one mismatch allowed) by analyzing 485,990,920 SOLiD BS-Seq reads (50 bp), taken from SRR204024 (Hansen, 2011). We found a high concordance between methylation calls of Hansen et al., analyzed by their yet unpublished and unavailable approach, and B-SOLANA. Moreover, B-SOLANA turns out to have a significantly higher mapping efficiency.

As a platform-independent benchmark, we demonstrate that the analysis of colorspace BS-Seq data of the fibroblast cell line IMR90 is comparable to methylome data published by Lister et al. (2009), who used a BS-Seq approach on the Illumina platform (Supplementary Information 1).

3 CONCLUSIONS

We present an efficient tool for the analysis of large colorspace BS-Seq data. B-SOLANA provides a fast and accurate all-in-one approach, including alignment and methylation calling. It is easy to use and generates an intuitive output, which can be used for genome-wide DNA methylation analysis.

ACKNOWLEDGEMENTS

We thank Ole Ammerpohl for helpful discussions (Institute of Human Genetics, Kiel), Gavin Meredith (Life Technologies, Foster City CA, USA) for providing SOLiD BS-Seq data of IMR90 cells and the sequencing facility at IKMB for helpful discussions and support.

Funding: Start-up grant from the Medizinassenschuss Schleswig Holstein, the National Genome Research Network (NGFN) of Germany (BMBF-funded); DFG Cluster of excellence "Inflammation at Interfaces" (infrastructure support); BMBF in the framework of the ICGC MMML-Seq project (to R.S. and J.R.).

Conflict of Interest: none declared.

REFERENCES

Bock,C. et al. (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat. Biotechnol., 28, 1106–1114.

Bormann, Chung,C. et al. (2010) Whole genome methylation analysis by ultra-deep sequencing using two-base encoding. PLoS One, 5, e9520.

Hansen,K. et al. (2011) Increased methylation variation in epigenetic domains across cancer types. Nat. Genet., 43, 768–775.

Holliday,R. et al. (1975) DNA modification mechanisms and gene activity during development. Science, 187, 226–232.

Langmead,B. et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., 10, R25.

Lister,R. et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462, 388–395.

Li,H. et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2578–2579.

Ondre,B.D. et al. (2010) An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. Bioinformatics, 26, 1901–1902.

Pedersen,B. et al. (2011) MethylCoder: Softwaier Pipeline for Bisulfite-Treated Sequences. Bioinformatics, 27, 2435–2436.

Pelizzola,M. et al. (2010) The DNA methylome. FEBS Lett., 585, 1994–2000.