Agilent SureSelect^{XT} Human Methyl-Seq for the Quantitative Analysis of DNA Methylation with Single-Base Resolution

Technical Overview

Abstract

Agilent SureSelect^{XT} Human Methyl-Seq is a unique solution-based tool for analyzing under and over-methylated cytosine sites in the human genome. The assay combines SureSelect, the leading target enrichment platform, with bisulfite sequencing, the gold standard for DNA methylation research and the first comprehensive discovery system. This enables unprecedented sequence coverage of only the most relevant regions for epigenetic studies, including those associated with a wide range of diseases such as cancer, imprinting disorders, behavioral and mental disorders, and many others.

Agilent SureSelect^{XT} Human Methyl-Seq allows researchers to analyze over 3.7 million individual CpG dinucleotide sequences for their methylation state. The system targets promoters, canonical CpG islands, and the more recently described shores and shelves found up to 4000 base pairs on either side of CpG islands. Studies have indicated that many methylation alterations are not in promoters or CpG islands, but most are within the 2kb CpG island shore. The SureSelect^{XT} Human Methyl-Seq kit also targets known differentially methylated regions (DMRs).

Agilent Technologies

Introduction

Epigenetic signatures that regulate gene expression are currently of great interest in the genomics field, especially for cancer research. While several methods have been developed to investigate further, existing methods are limited when it comes to resolution and susceptibility to bias.

- Methylation microarrays use bisulfite conversion and hybridization, rather than sequencing. This method has low resolution and cannot indicate the status of individual CpG sites.
- Whole genome bisulfite sequencing (WGBS) is based on bisulfite treatment followed by sequencing. It can be extremely costly and time consuming for large studies and is not practical for studying large sample numbers in profiling and biomarker research.
- MeDip-Seq is used for the enrichment of methylated regions followed by sequencing. This methods uses an antibody to immunoprecipitate methylated single-stranded DNA fragments prior to sequencing. While this method does enrich for methylated targets, it is biased towards repeat sequences and CpG-rich sequences.
- Reduced representation bisulfite sequencing (RRBS) is used to reduce the portion of the genome analyzed for methylation. DNA is first digested using a restriction enzyme, which cuts DNA at CpG recognition sites. This enrichment method cannot selectively target specific methylated regions and is biased towards repeats and CpG-rich sequences.

SureSelect^{XT} Human Methyl-Seq offers an alternative that allows for the quantitative analysis of DNA methylation with single-base resolution. When using the SureSelect Target Enrichment System, only the genomic areas of interest are sequenced, creating process efficiencies that reduce costs and allow more samples to be analyzed per study. As a result, investigators can more easily perform the experiments with a statistically relevant numbers of samples. Since the probes used in SureSelect are not methylation-state dependent, all targeted regions are captured regardless of their methylation states. SureSelect^{XT} Human Methyl-Seq's comprehensive design includes CpG islands, Gencode promoters, and known tissue- and tumor-specific DMRs in CPG island shores/shelves, DNase I hypersensitive sites, Refseq genes, and ensemble regulatory features.

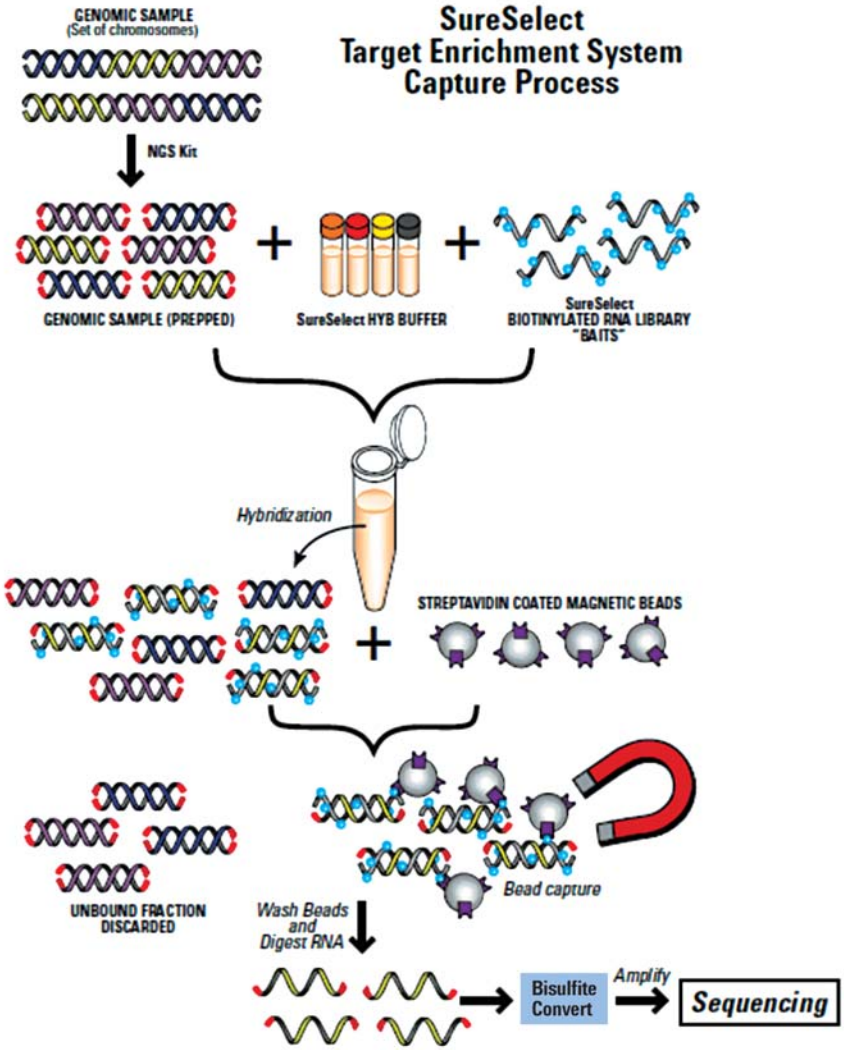

Figure 1. SureSelect target enrichment system workflow.

DESIGN CONTENT – 84 Mb Design, 3.7M CpGs

- CpG islands
- Cancer, tissue-specific DMRs
- Gencode promoters
- DMRs or regulatory features in:
  - CpG Islands, shores and shelves ±4kb
  - DNase I hypersensitive sites
  - Refseq genes
  - Ensembl regulatory features

SureSelect^{XT} Human Methyl-Seq Results are Highly Correlated with Whole Genome Data

Using fetal lung fibroblast (IMR90), we found that our data was highly correlated (R > 0.93) with published data. In Figure 2, each hexagon represents the density of mean methylation levels (filtered for > 10 reads). SureSelect^{XT} Human Methyl-Seq data from IMR90 cells was compared with the average methylation levels of published whole genome sequencing (WGS) data.[2]

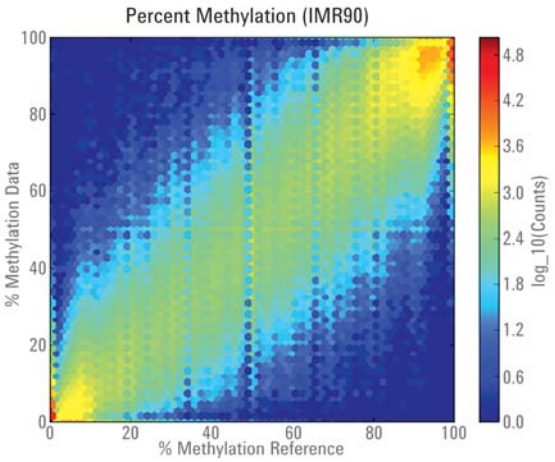

Figure 2. A comparison of SureSelect^{XT} Human Methyl-Seq data from fetal lung fibroblasts and published WGS results shows strong correlation (R > 0.93).

Proof of Concept in Colon Cancer Cells

Highly sensitive and accurate methylation detection following SureSelect target enrichment demonstrates the DNA methylation differences between HCT116 and its corresponding methyltransferase DKO (DNMT1-/- and DNMT3b-/-). Figure 3 shows that DNA methylation dramatically decreases in the HCT116 DKO cell, as expected since methyltransferases drive methylation of genomic DNA.

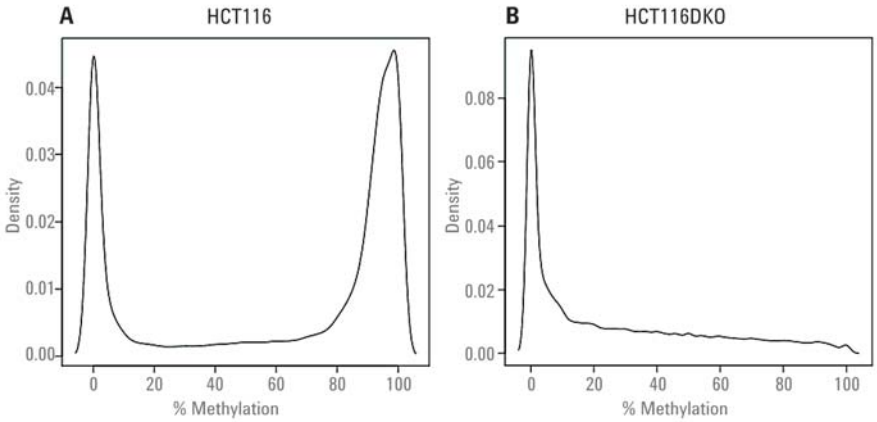

Figure 3. Methylation differences between wild type HCT116 (A) and its methyltransferase double-knockout (B).

Discover New Methylation Regions with Higher Coverage

Since Agilent probes are not methylation-state dependent they can all target known methylation regions within the genome, allowing for an increased coverage of target regions. With increased coverage, it is possible to uncover minor subpopulations of new methylated sites that would otherwise be missed with lower sequencing depth. SureSelect^{XT} Human Methyl-Seq (4 Gbp of uniquely mapped reads) outperforms whole-genome bisulfite sequencing (91 Gbp of uniquely mapped reads) in percent CpG covered in our design regions or specifically for CpG islands (Figure 4).

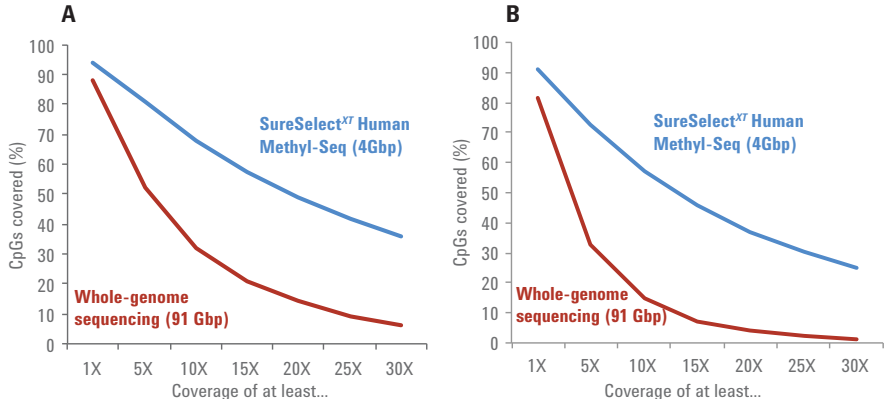

Figure 4. Coverage vs. % CpGs for ~3.7 million CpGs covered in targeted regions (A) and for ~1.9 million CpGs covered in CpG islands (B) using whole-genome sequencing (180Gb of raw sequenced data for 91Gb of uniquely mapped data) and SureSelect^{XT} Human Methyl-Seq (10 GB of raw sequenced data for 4 Gb of uniquely mapped data).

Discover Differential Methylation States for Various Tissues Types

Tissue-specific DMRs were successfully identified in 5 different normal tissues using SureSelect^{XT} Human Methyl-Seq. The average methylation levels were determined in 200 bp windows and clustered. As a result, 208,000 commonly methylated and 111,000 unmethylated regions were identified. Figure 5 shows the uniquely methylated regions identified in each tissue.

Cancer-Specific DMRs at Single Base Resolution

Tumor-specific DMRs were compared between normal and colon cancer tissue. SureSelect^{XT} Human Methyl-Seq detects not only known DMRs, but also potential novel DMRs at single base resolution. As an example, Methyl-Seq can also confirm the presence of DMRs in the HOXA3 gene as identified by comprehensive high-throughput arrays for relative methylation (CHARM) and pyrosequencing (52 % in normal and 72 % in tumor in pyrosequencing). In addition, various DMRs are found in the HOXA3 promoter and the genic and intergenic regions of HOXA3/HOXA4 using SureSelect^{XT} Human Methyl-Seq (Figure 6).

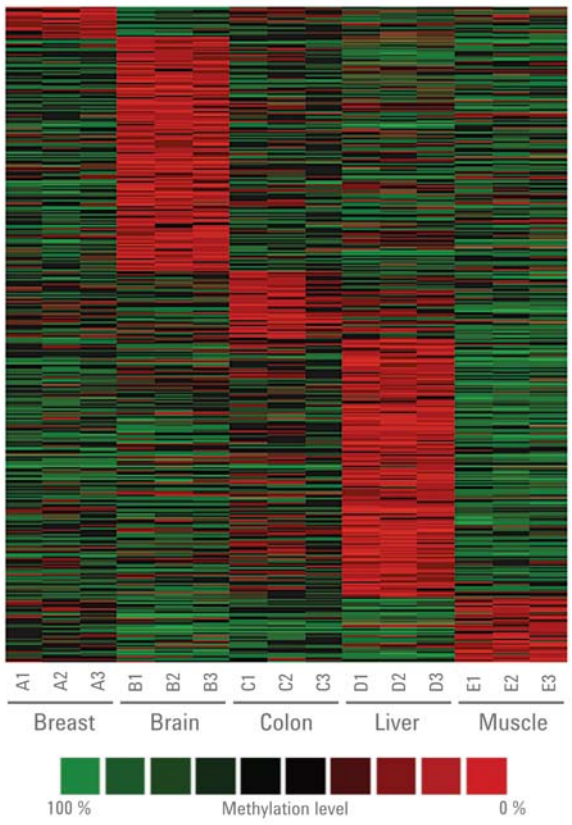

Figure 5. Uniquely methylated regions identified in five different tissues using SureSelect^{XT} Human Methyl-Seq.

Figure 6. A comparison of tumor-specific DMRs between normal (top) and colon cancer (bottom) tissue showing detection of known and novel DMRs at single base resolution.

Data Analysis

The goal of SureSelect^{XT} Human Methyl-Seq is to combine DNA bisulfite treatment with targeted high-throughput sequencing results to obtain a cell's complete epigenomic state by identifying methylation states of individual cytosines at single base resolution. Following capture by solution hybrid selection, there is a preprocessing stage prior to data analysis. This preprocessing stage encompasses DNA bisulfite treatment, high throughput sequencing of bisulfite treated (BST) DNA, and demultiplexing when multiple samples are being analyzed.

The SureSelect^{XT} Methyl-Seq analysis pipeline consists a series of steps: alignment, duplicate removal, normalization, detection, % methylation computation, and methylation identification (Figure 7).

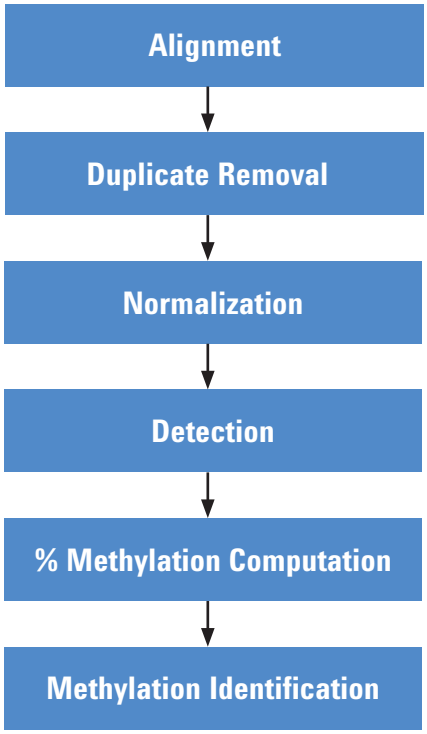

Figure 7. SureSelect^{XT} Methyl-Seq data analysis workflow.

1. Alignment

Bisulfite treatment of DNA leaves methylated cytosines unaffected, while non-methylated cytosines are converted into uracils. This specific characteristic of BST DNA allows one to identify methylated cytosines, but at the same time it adds complexities that make it more challenging to identify the true position of sequencing reads within the reference genome. In general, the methylation state of a BST DNA read must be inferred by comparison to an unmodified genomic reference sequence. This comparison is possible by aligning the BST read to the genomic sequence.

In order to do the alignment we use Bismark, a bisulfite sequence aligner and methylation caller (Figure 8). In the alignment process alternative versions of the read sequence are generated: C's are converted to T's and G's to A's (which is equivalent to a C-T conversion on the reverse strand). Similarly, converted versions of the reference genome are also generated and reads are aligned to each of these versions. By doing so, Bismark is able to predict the most likely bias-free alignment of the reads based on the least number of mismatches. In instances where the read aligns to more than one version of the genome with an equal number of mismatches, this read is discarded. Subsequently, each read is aligned to preconverted C-to-T and G-to-A versions of the reference genome. Here Bismark is trying to find the true strand and genomic position of the read without having to worry about the bias introduced by the bisulfite treatment. At this stage, if a sequence is aligned to multiple places with the same number of lowest mismatches it is discarded. Otherwise, the correct alignment is chosen as the one with the minimum number of mismatches.

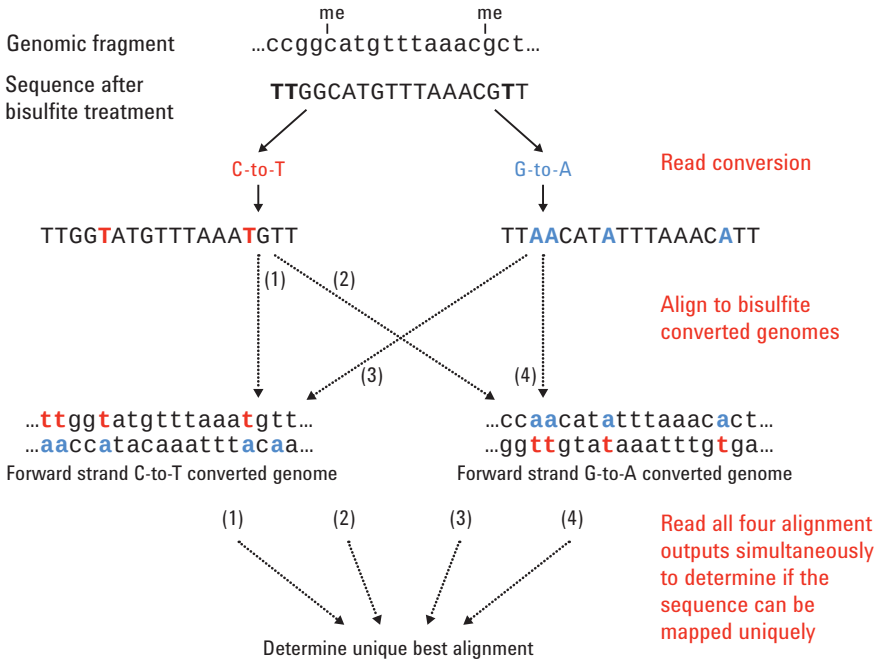

Figure 8. Application of Bismark[1] bisulfite sequence aligner and methylation caller.

2. Duplicate removal

In order to remove PCR duplicates all read pairs with identical strand, chromosome, and genomic position are removed and only one read pair is considered (Figure 9).

3. Normalization

When dealing with multiple samples, we randomly sample reads and assign the same number in order to make them comparable to each other. This is done in order to have a similar average read depth across the regions analyzed.

4. Detection

After the genomic position of a read has been determined through alignment, the methylation state of cytosines is determined in Bismark by simple comparison between the read sequence and the genomic reference Figure 10. Depending on the strand the read has been mapped to, this step marks methylated cytosines as those that did not have a C-to-T conversion (or a G-to-A conversion for the opposite strand) when comparing the reference genome sequence to the BST read sequence.

5. Computation of % methylation

After methylation detection Bismark's raw output summarizes the methylation state for each cytosine in every sequencing read in a list. In order to compute % methylation (% m) for one specific cytosine, Bismark's primary output is sorted by genomic position and % methylation is computed as the ratio of the methylated cytosines (k) to the total number of detected cytosines (n) at this position (Figure 11).

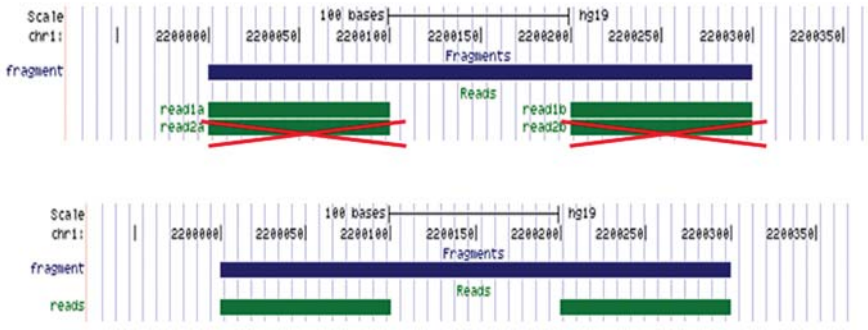

Figure 9. Deduplication: removal of PCR duplicates.

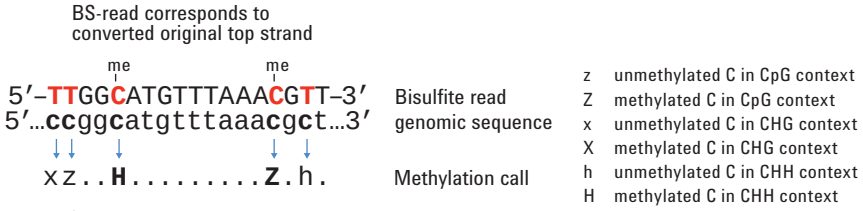

Figure 10. Methylation detection.

6. Methylation identification

The methylation state of a specific cytosine may be further determined based on the hypothetical distribution of methylation levels and statistical significance.[2]

For further information and full documentation of Bismark options and helper scripts, refer to Bismark's online user guide and resource page for the Bismark version you are using (Appendix 1).

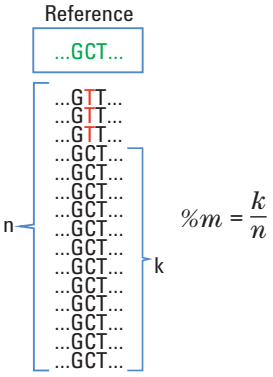

Figure 11. Computation of methylation levels.

Conclusions

Agilent's SureSelect^{XT} Human Methyl-Seq target enrichment platform provides a comprehensive, efficient, robust, and cost-effective method for sequencing subsets of the human methylome. Sequencing bisulfite converted DNA, following SureSelect^{XT} Human Methyl-Seq target enrichment, allows for the characterization of DNA methylation status in targeted regions at single base-pair resolution. The results achieved demonstrate excellent concordance with published whole genome data indicating their reliability. In addition, SureSelect^{XT} Human Methyl-Seq enables high reproducibility of enrichment, depth distribution, and sequence coverage from multiplexed sequencing. SureSelect^{XT} Human Methyl-Seq has also been used to confirm known tissue- and tumor-specific DMRs, as well as to identify identified potential novel DMRs further contributing to its value in comparison with existing methods used for interrogating methylated regions in the genome.

References

1. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq Applications. *Bioinformatics*, **2011**, 27:1571-1572.

2. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic difference. *Nature*, **2009**, 462:315-322.

Appendix: A practical guide to methylation data analysis*

Genome preparation: (This step needs to be done only once for each genome).

USAGE: bismark_genome_preparation [options] <path_to_genome_folder>

Sample command:
bismark_genome_preparation --path_to_bowtie /usr/local/bowtie/ --verbose

/data/genomes/homo_sapiens/GRCh37/

Alignment:
USAGE: bismark [options] <genome_folder> {-1 <mates1> -2 <mates2> | <singles>}

Sample command for a typical single-end analysis of a 40 bp sequencing:

bismark -q --phred64-quals -n 1 -l 40 --directional /data/genomes/homo_sapiens/GRCh37/ s_1_sequence.txt

Duplicate removal:
The deduplicate_bismark_alignment_output.pl helper script for this purpose is available.

Detection:
USAGE: ./methylation_extractor [options] <filenames>

Sample command for a pair-end file:
methylation_extractor -p --merge_non_CpG --comprehensive s_1_sequence.txt_bismark_pe.txt

% methylation computation:
The genome_methylation_bismark2bedGraph_v3.pl helper script for this purpose is available.

http://www.bioinformatics.bbsrc.ac.uk/projects/download.html#bismark

**For full documentation of Bismark options and helper scripts, please refer to Bismark's online user guide and resource page for the Bismark version you are using.*

www.agilent.com/genomics/sureselect

For Research Use Only. Not for use in diagnostic procedures. This information is subject to change without notice.

© Agilent Technologies, Inc., 2012
Published in the USA, May 10, 2012
5991-0166EN

Agilent Technologies