Genome analysis

PRIMEGENS-v2: genome-wide primer design for analyzing DNA methylation patterns of CpG islands

Gyan P. Srivastava[1], Juyuan Guo[2], Huidong Shi[2] and Dong Xu[1,*]

[1]Computer Science Department and Christopher S. Bond Life Sciences Center and [2]Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, MO, USA

ABSTRACT

Motivation: DNA methylation plays important roles in biological processes and human diseases, especially cancers. High-throughput bisulfite genomic sequencing based on new generation of sequencers, such as the 454-sequencing system provides an efficient method for analyzing DNA methylation patterns. The successful implementation of this approach depends on the use of primer design software capable of performing genome-wide scan for optimal primers from *in silico* bisulfite-treated genome sequences. We have developed a method, which fulfills this requirement and conduct primer design for sequences including regions of given promoter CpG islands.

Results: The developed method has been implemented using the C and JAVA programming languages. The primer design results were tested in the PCR experiments of 96 selected human DNA sequences containing CpG islands in the promoter regions. The results indicate that this method is efficient and reliable for designing sequence-specific primers.

Availability: The sequence-specific primer design for DNA methylated sequences including CpG islands has been integrated into the second version of PRIMEGENS as one of the primer design features. The software is freely available for academic use at http://digbio.missouri.edu/primegens/.

Contact: xudong@missouri.edu

1 INTRODUCTION

The importance of epigenetic effects in biological processes and diseases has been more and more recognized. Methylation of cytosine residues at CpG dinucleotides is the best studied epigenetic modification in mammalian genomes and is known to have profound effects on gene expression. Over the past 3 years, an international consensus has emerged in the epigenetics research community for the need of an organized *Human Epigenome Project* (HEP) aimed at generating a high-resolution DNA methylation map of the human genome in all major tissues (Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004). The recently initiated HEP will provide a 'reference epigenome' by resequencing different normal tissues and adding 5-methylcytosine to the DNA sequencing datasets (http://nihroadmap.nih.gov/epigenomics). The pilot HEP in Europe utilized direct sequencing of bisulfite PCR products to provide single

methyl-cytosine resolution mapping of thousands of amplicons (Eckhardt *et al.*, 2006). In this method, the methylation present at any given CpG site is estimated by taking the average of all fragments (thousands) generated during PCR, which results in a more statistically robust representation of the methylation patterns as compared to sub-cloning. Recently, we applied an innovative massively parallel sequencing-by-synthesis method (454-sequencing) for ultra-deep bisulfite sequencing analysis of multiple tumor methylome (Taylor *et al.*, 2007). This highly parallel sequencing system has many potentially important applications, for example development of a high-throughput, large-scale bisulfite genomic sequencing approach that provides an efficient method for deeply exploring the human epigenome.

The successful implementation of above-mentioned approach depends on the use of automatic primer design program capable of performing genome-wide scans for optimal primers from *in silico* bisulfite-treated human genome sequences. Several methods have been proposed to address this issue partially. MethPrimer (Li and Dahiya, 2002) and PerlPrimer (Marshall, 2004) transform the target sequences according to the bisulfite treatment for primer design. These methods do not provide a mechanism to detect non-specific amplification in bisulfite PCR. Bisearch (Tusnády *et al.*, 2005) provides an important feature of similarity search for potential non-specific PCR product with the selected primer pairs on a bisulfite-treated genome. It uses a simple string matching search method to detect potential cross-hybridization of a designed primer pair. The string matching search can find exact match of the primer but cannot detect highly similar sequences (e.g. with 1 nt mis-match) in the genome to the primer, which could also be potential binding site for the primer. In addition, this method is practically not suitable for analyzing the primer pairs for mispriming sites in case of high-throughput primer design, which is required for highly parallel sequencing system to develop high-throughput, large-scale bisulfite genomic sequencing. To address this issue, we developed an efficient method and integrated it into the second version of our software system PRIMEGENS (Xu *et al.*, 2000, 2002; Srivastava and Xu, 2007), PRIMEGENS-v2. PRIMEGENS builds on third-party, open-source software tools like Primer3 (Rozen and Skaletsky, 2000) and BLAST (Altschul *et al.*, 1990) and has various new features for genome-scale primer design. PRIMEGENS has been widely used and cited by the research community (Bertone *et al.*, 2005; Chen *et al.*, 2006; Ehses *et al.*, 2005; Haas *et al.*, 2003; He *et al.*, 2005). However, our early version did not have the feature of primer design

---

*To whom correspondence should be addressed.

for bisulfite sequencing. It did not have a graphical user interface (GUI) and it did not run under Windows operating systems. We extended the PRIMEGENS algorithm to include sequence-specific primer design for bisulfite PCR and align these primers using Mega BLAST (Zhang *et al.*, 2000) to check cross-hybridization across *in silico* bisulfite-treated human genome. We also developed PRIMEGENS as a standalone tool with GUI to run under both Linux and Windows.

2 METHODS

The goal of our prime design experiment is to design primer for specific region of genes to cover both the transcription start site (TSS) of the gene and part of a CpG island, which is located in the vicinity of the gene either in the promoter region or in the transcription region. Recent studies show that methylation of CpG sites near the TSS is critical to the expression of the *hTERT* gene in cancer cells (Zinn *et al.*, 2007). Since PRIMEGENS will be likely most useful for the genome-wide bisulfite sequencing experiments such as the Human Epigenome Pilot Project (Eckhardt *et al.*, 2006; Rakyan *et al.*, 2004), we aimed to automate our primer design pipeline so that only a list of gene name is required to perform the primer design. Based on the gene name and the TSS information associated with the gene, we can automatically retrieve the target sequences from the human genome database and design the primers based on a few parameters associated with the TSS such as the distance to the TSS. The main computational challenge in such a primer design is to avoid cross-hybridization of the fragment-specific primers to the other place of thymine-rich methylated genome. To address this problem, we modified our previously developed PRIMEGENS algorithm as shown in Figure 1. The algorithm is composed of two basic components. The first component performs primer design using Primer3, which provides a set of primer pairs and the second component applies Mega BLAST to search the selected primer pairs against bisulfite-treated genomic sequence to find potential non-specific PCR products.

In order to design primer for any sequence, we first convert the target sequence and the complete human genome into bisulfite-treated sequences, where all the cytosine (C) sites in original sequence are converted into thymine (T) except places where cytosine is preceding guanine (G) known as methylation of the CG. In order to run Mega BLAST for the designed primers, we consider the bisulfite-treated human genome as a database. For each chromosome sequence, four variants are generated as a model of the bisulphite-treated sequences (Pattyn *et al.*, 2006): (1) bisulfite methylated forward sequence, (2) bisulfite methylated reverse sequence, (3) bisulfite unmethylated forward sequence and (4) bisulfite unmethylated reverse sequence. As an example, suppose a fragment of human genome has nucleotide sequence 'agctagccagtcga', then this fragment is modified to generate four variants as follows (Pattyn *et al.*, 2005):

    agctagccagtcga—original
    agttagttagttga—unmethylated forward
    agttagttagtcga—methylated forward
    ttgattggttagtt—unmethylated reverse complementary
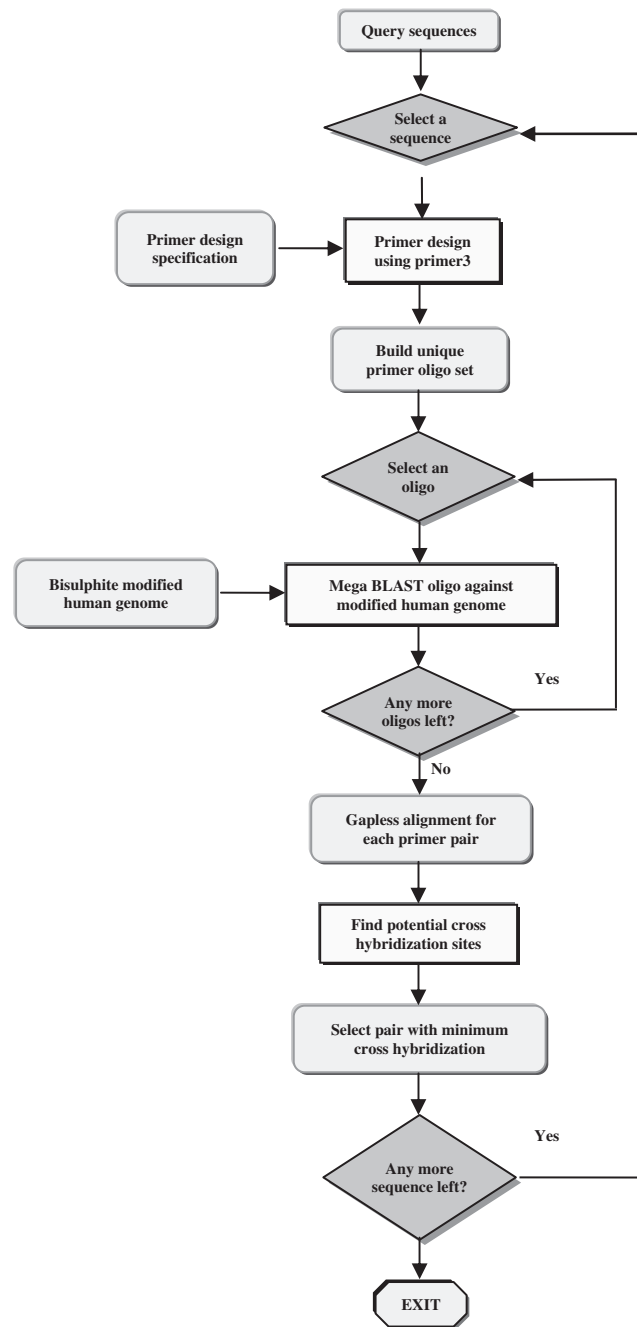   tcgattggttagtt—methylated reverse complementary

Fig. 1. Basic primer design flow chart for PRIMEGENS-v2.

2.1 Prepare input data

Before starting the primer design, appropriate query sequences are generated. As a fact, the location of any CpG island with respect to TSS could be in three ways.

(1) CpG island completely on the left side of the TSS;
(2) CpG island completely on the right side of the TSS;
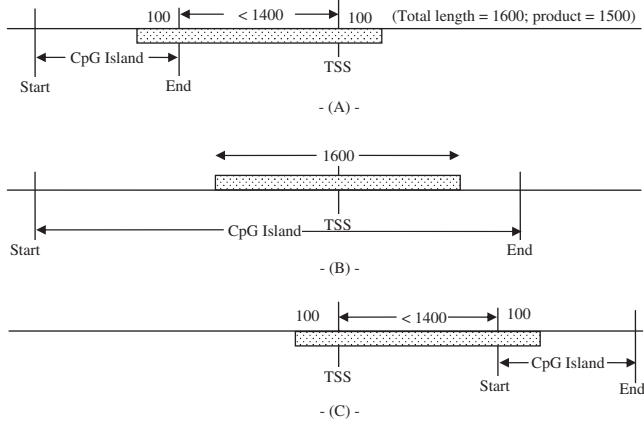(3) CpG island contains the TSS.

Fig. 2. A CpG island can be located near the TSS in three ways. Based on the location of CpG island, the query sequence could be designed to cover partial CpG island region and the TSS together.
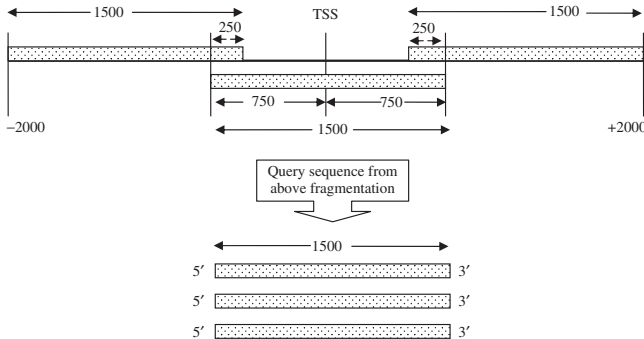

Fig. 3. Partitioning method of generating query sequence fragments to cover the CpG island region located far from the TSS.

Based on the location of a CpG island, a fragment-specific query sequence can be designed so that most of the CpG island region along with the TSS is covered within the PCR product. As an example, Figure 2 explains a strategy to select a query sequence of 1600 nt with a product size ranging 1500–1600 so that any CpG island within the vicinity of 1400 nt from the TSS is covered. If a CpG island contains the TSS, the region of the CpG island that covers the TSS is selected. In case of a CpG island either in the left or right side vicinity of the TSS, the genome region of 1500 nt from the TSS toward the CpG island and additional 100 nt from the opposite site is selected as the fragment-specific fragment.

This strategy can only be applied when a CpG island is not far from the TSS, in particular, closer than 1400 nt to the either side of TSS. In case the TSS and the CpG island are far away from each other, we partition the selected region to generate reasonably long fragments; e.g. we cut the region into multiple and partially aligned fragments as shown in Figure 3.

We designed primer pairs for 1012 cancer related genes and randomly selected 96 genes for experimental validation. We first downloaded 1012 CpG island sequences at promoter regions that associate with these genes from the UCSC genome website http://genome.ucsc.edu and added extra 100 nt upstream (5′) and

Table 1. Primer design parameters

Minimum product size = 250 nt
Maximum product size = 450 nt
Most suitable primer length = 24 nt
Minimum primer length = 20 nt
Maximum primer length = 28 nt
Minimum melting temperature = 50.0°C
Most suitable melting temperature = 58.0°C
Maximum primer melting temperature = 60.0°C
Primer's maximum GC content = 80.0%
Primer's minimum GC content = 20.0%
Primer salt concentration = 50.0 mm
Primer DNA concentration = 50.0 nm

100 nt downstream (3′) region for each sequence. We then performed *in silico* conversion for the sequences into bisulfite-treated sequences. All the *in silico* converted DNA sequences were stored in a single file in the FASTA format. Since the bisulfite modification will result in two single strand sequences that are no longer complementary to each other, first the *in silico* bisulfite converted sense strand sequence is used. If no suitable primer pair is found, the antisense strand is used for designing primers.

2.2 Primer design for target sequence

PRIMEGENS uses Primer3 to search for 20 unique pair of 18mer to 25mer oligonucleotides with a melting temperature (Tm) of 58°C and other specifications (Table 1) with no potential formation of stable DNA secondary structures. These oligonucleotides are also not expected to contain any CpG dinucleotides to avoid possible methylation within primers. For each query sequence, PRIMEGENS outputs the primer pair oligos having least cross-hybridization (see the following Section 2.3) with the length ranging 18–25, Tm = 58°C, and no internal DNA secondary structures and CpG dinucleotides.

2.3 Non-specific PCR amplification detection

The potential non-specific PCR amplification with the designed primer pair is explored using Mega BLAST. Mega BLAST performs gapless alignments for the oligos designed in the preceding step against the four variant *in silico* bisulfite converted human genome sequences to determine the binding capacity of the oligos to the genome. If an oligo sequence has a significant similarity (either identical or with few mismatches, depending on the threshold used) to any part of genome, then it would be a potential binding site for that oligo. In order to amplify non-specific PCR product, two conditions need to be satisfied: (1) both the left primer and right primer should bind at appropriate places and (2) the amplified PCR product length should not be too long, as long PCR products would not be amplified effectively. Based on these two restrictions and the user-defined threshold for the non-specific PCR product size, PRIMEGENS will select the most query sequence-specific primer pairs for further consideration.

3 EXPERIMENTAL VALIDATION

We successfully designed primer pairs for 1012 query sequences. In order to validate primer design using PRIMEGENS, we randomly picked and synthesized 96 pairs of primers and performed bisulfite
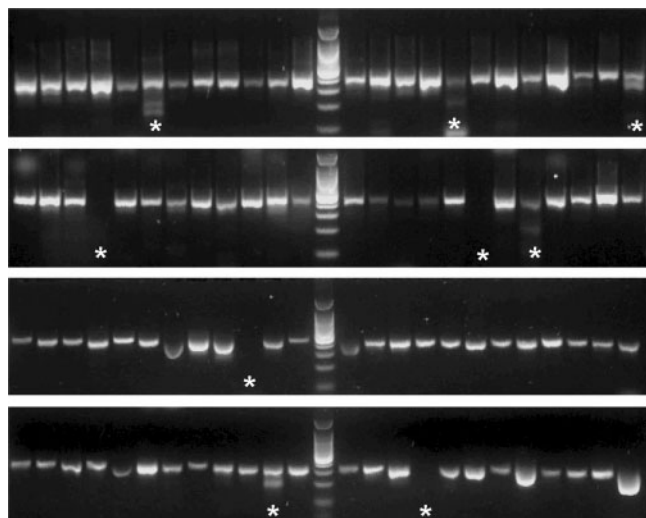
Fig. 4. Experimental validation of the PCR primers designed by PRIMEGEN. Genomic DNA isolated from RL cells was bisulfite-treated and subjected to PCR using primers designed by PRIMEGENS. A total of 96 primer pairs were tested using the same touchdown PCR program in a 96-well PCR plate. The PCR products were examined on 2% agarose gels. Out of 96 primer pairs, 87 successfully yielded expected unique products. *indicates the failed reactions or multiple PCR products.

PCR using bisulfite-treated DNA in a 96-well PCR plate. As shown in Figure 4, 87 out 96 primers (91%) generated unique PCR products and all PCR products are similar in size as designed.

To compare with primer design without considering bisulfite-treated sequences, we also analyzed the percentage of generated primer pairs that appear unique via Mega BLAST on regular human genome without bisulfite-treated human genome sequences. For all the 96 primer pairs used in the PCR experiment, we applied (1) the regular genome sequence to run Mega BLAST and found total 91 (~94%) primer pairs showing unique hybridization; (2) bisulfite-treated genome sequences to run Mega BLAST and found total 81 (~84%) primer pairs showing unique hybridization. It is clear that ~10% of the primer pairs are excluded due to cross-hybridization resulting from bisulfite modification on genome. Interestingly, the experimental success rate (~91%) is higher than what is predicted by PRIMEGENS (~84%). This is because PRIMEGENS uses very stringent criteria (e.g. to allow some nucleotide mismatches of primer as potential cross-hybridization). Some primer pairs are predicted to have possible cross-hybridizations, but the resulting PCR products are experimentally unfavorable or have too low yields to be observed.

4 SOFTWARE IMPLEMENTATION

Figure 5 shows the user operation flowchart. Four basic panels can be controlled automatically in order to minimize wrong steps taken by a beginner. A user can repeat the whole primer design in a single activation of PRIMGENS software. On the other hand, PRIMGENS provides various options for advanced users (see the user manual at http://digbio.missouri.edu/primegens/). The software is general enough to design primer pairs for a PCR product containing any
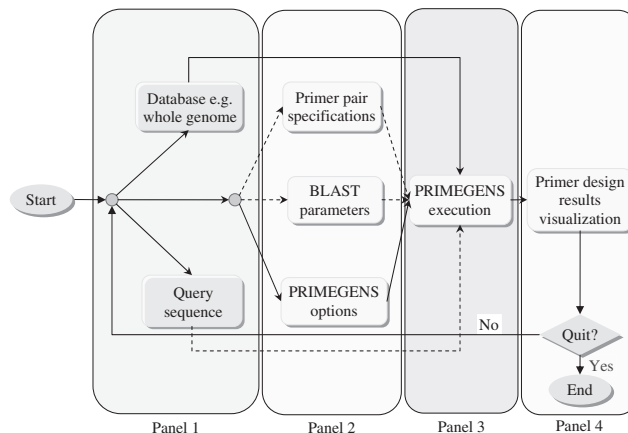

Fig. 5. PRIMEGENS software user operation flow chart.

Fig. 6. The main GUI for PRIMEGENS.

specific site (the TSS as an example) and/or a specified segment (CpG island as an example).

In the software development, JAVA is used to develop the GUI, whereas the core program is written in C for computational efficiency. The program also links to the executables of Primer3 and Mega BLAST. PRIMEGENS also supports running the computational jobs in the batch mode. It has been tested on various Unix/Linux platforms, including high-performance Linux clusters. PRIMEGENS is a standalone package and can be downloaded from http://digbio.missouri.edu/primegens. Once the software is installed, a user will be able to see the interface as shown in Figure 6.

On average it takes 4–5 min on a local desktop machine to design primer pairs for a given query sequence, where most of the running time is consumed by Mega BLAST search across the bisulfite-treated genomic sequences. Once the primer design is done, a user can visualize and analyze the design primer pairs along with the respective query sequence and search database as shown in Figure 7.
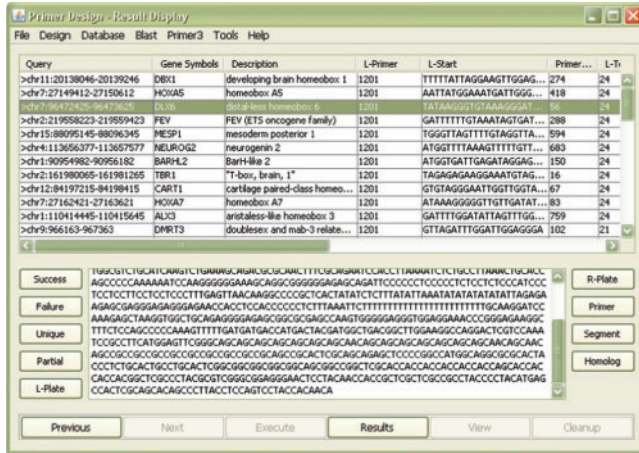
Fig. 7. GUI for visualizing PRIMEGENS primer design result.

Table 2. This table summarizes various features of available primer design tools

Tools	Free	Scale	Specificity	Avail.	B/S	Reference
Primer3	Yes	One	No	Web	No	(Rozen and Skaletsky, 2000)
MethPrimer	Yes	One	No	Web	Yes	(Li and Dahiya, 2002)
Meth-BLAST	Yes	One	Manual	Web	No	(Pattyn *et al.*, 2006)
PerlPrimer	Yes	One	No	GUI	Yes	(Marshall, 2004)
BiSearch	Yes	One	Manual	Web	Yes	(Tusnády *et al.*, 2005)
Methyl Primer Express	No	One	No	GUI	Yes	Applied Biosystems, Foster City, CA, USA
EpiDesigner	Yes	Large	No	Web	Yes	Sequenom Inc., San Diego, CA, USA
PRIMEGENS	Yes	Large	Automatic	GUI	Yes	This article

The acronyms used for column headers are; Free: whether the tool is freely available or not; Scale: whether the tool designs primers for one gene at a time or on a large scale; Specificity: whether the tool checks cross-hybridization for designed primers or not; Avail.: availability of the software (Web server or standalone GUI); B/S: whether the tool designs primers for bisulphate-treated sequences or not.

5 DISCUSSION

In this article, we present a method capable of designing sequence-specific primer pairs for bisulfite-treated sequences at the genomic scale. This method is incorporated into the second version of PRIMEGENS. It not only searches for appropriate primers but also checks for non-specific PCR amplification.

PRIMEGENS provides a unique software package for the research community, compared to other available tools for primer design as shown in Table 2. While most other tools design primers for one gene at a time, PRIMEGENS can design primers for thousands of genes (fragments) by one run. Furthermore, PRIMEGENS is the only available tool which checks primer specificity automatically on a large scale. Our experiment shows that the success rate for the PCR experiment is about 91%. This is a significant improvement over our earlier manual design with the same parameters using the Methyl Primer Express software (Applied Biosystems, Foster City, CA, USA, http://products.appliedbiosystems.com), which was the best for such a design to our knowledge. There, only 130 out of 227 primer pairs worked under the same PCR condition, which gave a success rate of 57.3%. Even under other PCR conditions, only 35 additional primer pairs worked, which gave a success rate of 72.6% (165 out of 227). Sixty two primer pairs failed under three PCR conditions and they were not tested further.

The efficiency of similarity search lies in using Mega BLAST, which is well known as one of the fastest DNA sequence alignment algorithms. Since the typical memory requirement in Mega BLAST against the human genome is higher than what most desktop machines have, the human genome is split into chromosomes and all the four variants of each chromosome are saved, making a total of 96 chromosome files (24 chromosomes with 4 variants for each chromosome). Most running time is taken by Mega BLAST on four variants for each of the human chromosome sequence. Due to the scaling problem, users are not advised to use the software on a low-memory desktop machine. Later version of PRIMEGENS will better serve desktop users for bisulfite primer design, by having precomputed data and indexing for the human genome bundled with the software. We also plan to develop a web server for PRIMEGENS so that users can apply the software more easily. As more and more experimental studies are conducted for methylation, we expect our tool will benefit many users.

REFERENCES

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bertone,P. *et al.* (2005) Design optimization methods for genomic DNA tiling arrays. *Genome Res.*, **16**, 271–281.

Chen,Y.A. *et al.* (2006) A multivariate prediction model for microarray cross-hybridization. *BMC Bioinformatics,* **7**, 101.

Eckhardt,F. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.

Ehses,S. *et al.* (2005) Optimization and design of oligonucleotide setup for strand displacement amplification. *Biochem. Biophys. Methods*, **63**, 170–186.

Haas,S.A. *et al.* (2003) Genome-scale design of PCR primers and long oligomers for DNA microarrays. *Nucleic Acids Res.*, **31**, 5576–5581.

He,Z. *et al.* (2005) Empirical establishment of oligonucleotide probe design criteria. *Appl. Environ. Microbiol.*, **71,** 3753–3760.

Li,L.C. and Dahiya,R. (2002) MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, **18**, 1427–1431.

Marshall,O.J. (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, **20**, 2471–2472.

Pattyn,F. *et al.* (2006) methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. *BMC Bioinformatics,* **7**, 496.

Rakyan,V.K. *et al.* (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.*, **2**, e405.

Rozen,S. and Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In Misener,S. and Krawetz,S.A. (eds.) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp. 365–386.

Srivastava,G.P. and Xu,D. (2007) Genome-scale probe and primer design with PRIMEGENS. In Yuryev,A. (ed.) *Methods in Molecular Biology,* Vol 402*: PCR Primer Design*. Humana Press, Totowa, New Jersey, pp. 159–175.

Taylor,K.H. *et al.* (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, **67**, 8511–8518.

Tusnády,G.E. *et al.* (2005) BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. *Nucleic Acids Res.*, **33**, e9.

Xu,D. *et al.* (2000) *Currents in Computational Molecular Biology: A Computer Program for Generating Gene-Specific Fragments for Microarrays*. Universal Academy Press, Tokyo, pp. 3–4.

Xu,D. *et al.* (2002) PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. *Bioinformatics*, **18**, 1432–1437.

Zhang,Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.

Zinn,R.L. *et al.* (2007) hTERT is expressed in cancer cell lines despite promoter DNA methylation by preservation of unmethylated DNA and active chromatin around the transcription start site. *Cancer Res.,* **67**, 194–201.