Automated Selection of Differentially Methylated Regions in Microarray Data

Pavlos Antoniou, Spiros Michalakopoulos, Elisavet A. Papageorgiou,
Philippos C. Patsalis and Carolina Sismani

***Abstract*— Differentially methylated regions (DMRs) are segments or islands of consecutive sequence positions, showing methylation enrichment or depletion compared to each other in different samples or tissues. The identification of DMRs is a crucial first step in the discovery of biomarkers for non-invasive prenatal diagnosis of aneuploidies such as Trisomy 21. In this paper we describe an algorithm to automatically identify the manifestation of DMRs on arrays. Our approach, methylation status mask AND (MS-AND), influenced by the SHIFT-AND methodology, uses bit operations and masking and can be applied to any microarray dataset in General Feature Format (GFF). We show the effectiveness and utilization of our algorithm using data from Methylated DNA Immunoprecipitation arrays for the identification of DMRs in chromosomes 13, 18 and 21. The algorithm runs on Linux and on Windows systems and an implementation is available at sourceforge (`http://sourceforge.net/projects/ms-and`).**

I. INTRODUCTION

The completion of the whole human genome project coupled with the introduction of high throughput microarray and sequencing technologies has provided scientists with powerful means to study and analyse DNA. The human *epigenome* however, still uncharted, provides a larger amount of information than merely the sequence of a gene. There is thus strong interest for its analysis and study. *Epigenetics* is the study of heritable traits other than those carried by DNA sequences and entails DNA methylation and chromatin modifications.

DNA methylation is a chemical modification of DNA that occurs when methyl groups are added to its bases. Methylation most often includes the cytosine of CG pairs in DNA, usually in both strands [9]:

$$5' -^{m} CpG - 3'$$

$$3' - GpC^{m} - 5'$$

where m denotes the methylated cytosine.

Pavlos Antoniou, Philippos C. Patsalis and Carolina Sismani are with the Department of Cytogenetics and Genomics, Cyprus Institute of Neurology and Genetics, 6 International Airport Avenue, 1683 Nicosia, Cyprus. `pavlosa@cing.ac.cy`. This work is supported from ANGELAB: A new Genetic Laboratory for non-invasive prenatal diagnosis, `EU FP7 Programme contract no. 317635`

Spiros Michalakopoulos is with MediaNet Software, Llano Castellano 13, 28034 Madrid, Spain. `spiros.michalakopoulos@medianet.es`

Elisavet A. Papageorgiou is with NIPD Genetics, Neas Engomis 31, Engomi, Nicosia, 2409 Cyprus. `e.papageorgiou@nipd.com`

DNA methylation occurs at CpG dinucleotides and is one of the most extensively studied epigenetic modifications [18]. It is found throughout the genome except in short unmethylated regions called CpG islands (CGIs) [22], which are usually (but not always) unmethylated. CGIs are DNA sequences approximately 1000 bp long within which the dinucleotide CG is present in a frequency of around 65% of the sequence base composition.

In a eukaryotic species approximately 5% of the cytosine residues are methylated but the extent of methylation can be tissue specific and can vary from 2% to 7% [9]. This epigenetic modification regulates gene expression, silences the activity of genes and transposable elements and stabilizes gene dosage in X-inactivation and genomic imprinting [3]. DNA methylation information is multiplex with hundreds of potential methylated cytosines in a gene. It is also quantitative because tissues can maintain partial methylation at a locus which means that the extent to which a site is methylated can vary [3]. Methylated regions are present in imprinted genes, which are essential for growth and development. Imprinted genes carry parental allele specific methylation profiles and are characterised by differentially methylated regulatory regions (DMRs) [14].

A. DNA methylation analysis and discovery of differentially methylated regions

The detection of DNA methylation is based on the ability to distinguish cytosine from its methylated version of 5-methylcytosine (m5C) [3]. Investigation of methylation differences between samples, leads to the identification of differentially methylated regions (DMRs).

A number of different techniques are used for identifying DMRs. The preprocessing can be performed by:

- Digestion of DNA by a methylation sensitive restriction endonuclease.
- Bisulfite conversion: sodium bisulfite converts unmethylated cytosine to uracil, whereas methylated cytosines are protected from conversion.
- Methylated DNA immunoprecipitation (MeDIP): using an antibody, DNA containing methylated cytosines is immunoprecipitated and selected.

The use of methylation sensitive restriction enzymes, requires the presence of restriction sites in the target sequences, therefore there is a limitation on the number of regions that can be tested. On the other hand, the accurate analysis of the methylation status after bisulfite conversion, depends on the conversion of all unmethylated cytosines to uracils, which

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

rarely occurs, as DNA is degraded by extensive bisulfite treatment [15].

After preprocessing the samples using one of the above methods, the methylation information is obtained by a corresponding technology:

- DNA oligonucleotide arrays, after bisulfite conversion or MeDIP.
- SNP arrays.
- BeadArray (Illumina), after bisulfite conversion.
- Sequencing (Sanger or Next Generation), after bisulfite conversion or MeDIP.

These high throughput approaches have been used to analyse DNA methylation across the genomes of *A.thaliana*, the mouse genome, and parts of the human genome. The first genome wide, single base resolution map of the human methylome was presented in [12], where bisulfite sequencing was used to sequence two human cell lines, embryonic stem cells and fetal lung fibroblasts. Given the great interest in this cutting-edge technology of next generation sequencing, several algorithms based on statistical analysis have been developed that detect differentially methylated regions after bisulfite sequencing.

BSmooth [8], first aligns the bisulfite treated reads to the genome, compiles quality assessment metrics based on methylation estimates by read position, and applies local average to detect DMRs taking into account any available replicates. The *eDMR* algorithm [11], extends an R pipeline (methylkit) [20], to discover boundaries between the DMRs from bisulfite sequencing reads using bimodal normal distribution, and then calculating the statistical significance of the DMRs with the Stouffer-Liptak test. *COHCAP* [23], takes targeted bisulfite sequencing data and Illumina methylation array data, as input, to identify CGIs that show a consistent pattern of methylation among CpG sites; the authors define a cutoff value for methylation status and categorise the regions to methylated and unmethylated depending on this value; they use a p-value and a false discovery rate to statistically assess their findings. *CpG_MPs* [21], identifies CpG methylation patterns in bisulfite sequencing data, by first normalising the sequencing reads into methylation levels of CpGs, and then identifying differentially methylated regions using the hotspot extension algorithm coupled with the Shannon entropy method.

Algorithms for identifying DMRs using microarrays have also been developed. These are again mostly based on statistical analysis. *BioTile* [6], is a Perl application that identifies DMRs in tiling microarray datasets, by first calculating the slope of a linear model between dependent and independent variables for each array probe, and subsequently returning regions of adjacent probes with slope values above or below zero, for three or more consecutive probes, as potential DMRs; finally, the distribution of DNA methylation across each probe in a DMR, is evaluated by a permutation corrected statistical meta-analysis. The software package *QDMR* [24], similarly to *CpG_MPs*, is based on the Shannon entropy method to identify DMRs and quantify methylation differences between the regions.

In this work, we focus on the MeDIP-chip approach, wherein the samples are preprocessed with methylated DNA immunoprecipitation and the selected regions are co-hybridized on an oligonucleotide array with control DNA, to find DMRs. The presented solution intentionally avoids the complexities of statistical analysis, for the express purpose of allowing its use by a wide range of practitioners, including those who are inexperienced in this domain. It is simple to use and identifies DMRs based on the methylation values of the array regions, without requiring statistical decisions or assumptions to be made prior to algorithm use. The program has been implemented (`http://sourceforge.net/projects/ms-and`) to allow for user-input threshold values, which return more finely grained results. The execution time is very low, which permits multiple manual permutations. However, in the paper, for clarity and simplicity, we present the solution where the threshold values are predefined to zero.

B. Differentially methylated regions for non-invasive prenatal diagnosis of aneuploidies

Prenatal diagnosis of chromosomal aneuploidies is currently performed by cytogenetics or DNA analysis of fetal DNA material. This is obtained by amniocentesis, chorionic villus sampling or chordocentesis. All these methods are invasive and have a significant risk of fetal loss (0.5 to 1% for chorionic villus sampling and amniocentesis) [15]. The discovery of free fetal DNA in maternal circulation during pregnancy [13], has shifted the focus of research to the discovery of non-invasive prenatal diagnosis (NIPD) methods. By sampling peripheral blood from the pregnant woman it is now possible to differentiate fetal from maternal DNA, and analyse the fetal DNA in the laboratory. One of the most promising demonstrations of NIPD is based on the use of DMRs, which show differences in methylation status between fetal and maternal DNA [17].

One such successful method, for NIPD of aneuploidies, the MeDIP and real time qPCR method by Papageorgiou et al. [16], has been used to detect Trisomy 21 from maternal peripheral blood. This group has previously investigated DMRs on chromosome 21 as well as on chromosomes 13, 18, X, and Y, by using methylated DNA immunoprecipitation, coupled with high resolution oligonucleotide arrays. The method is based on identifying regions that are hypermethylated in placenta DNA, both in first and third trimester, and hypomethylated in female whole blood. From these regions, qPCR primers were designed and the amplicons serve as markers that differentiate fetal from maternal DNA.

In order to discover these regions the authors in [15] used the SW-ARRAY algorithm [19], which is a dynamic programming algorithm to detect copy number variations in array comparative genome hybridization (array-CGH) DNA data. The algorithm detects very large regions of copy number variations. These were taken as starting points for subsequent manual selection, by visualizing them with SignalMap (Nimblegen Software), of the required regions. It is difficult to manually perform this task considering the complexity

of high throughput array data. There is thus a strong need for an automated method to solve the problem of finding regions in all three datasets of hypermethylation status in first and third trimester placenta DNA, and hypomethylated status in maternal whole blood, in order to directly use them as potential markers for NIPD. The algorithm in this paper addresses this need.

The rest of the paper is organized as follows. In the next section we provide the preliminaries: a description of the biological data in Section II-A, and the necessary definitions for the algorithmic solution in Section II-B. Then, in Section III we formally define the problem and present our solution, and in Section IV we present and discuss our experimental results. Finally, we conclude in Section V and discuss further works.

II. Preliminaries

A. Description of input data

We applied our method to the chromosome 13, 18, and 21 datasets of first and third trimester placenta DNA array, and whole blood DNA arrays obtained by Papageorgiou et al. [15] after MeDIP-chip assays. The high resolution tiling oligonucleotide arrays, specific for chromosomes 13, 18 and 21, had a median probe span of 225 bp for chromosome 13, 170 bp for 18 and 70 bp for chromosome 21. The input genomic control DNA and the immunoprecipitated DNA of each sample were differentially labelled with fluorescent dyes, (Cy3, Cy5), and were co-hybridized on the arrays [15]. The ratio of the signal intensities in each datasets was then $\log_2$ normalised and the normalised *$\log_2$ratio* array datasets were used for analysis. We used datasets generated by Papageorgiou et al. [15], to test our method. The total number of probes for the chromosome 13 arrays was 385178 probes, for chromosome 18, 385482 and for chromosome 21, 385183 probes.

In the datasets, $\log_2$ratio values represent methylation values. Hypermethylated probes are the regions that have a $\log_2$ratio value larger than zero and hypomethylated regions are the regions that have a $\log_2$ratio less than zero. In our analysis, we require at least three consecutive probes to have a methylation value larger than a threshold of zero to be selected as a candidate region for placenta DNA arrays, whereas for whole blood array data we require that the corresponding probes have methylation values less than a threshold of zero.

B. Definitions

A *region*[1] is a sequence of base pairs, starting at location r and stopping at location r'. A region's *span* is $[r..r']$. A *probe* p is a 3-tuple (α, ω, μ), where α is the start location, ω is the stop location, and μ is the methylation status in the $[\alpha..\omega]$ region. A *sequence of probes* $\mathcal{P} = p_1, \ldots, p_n$ is a sequence of 3-tuples $(\alpha_1, \omega_1, \mu_1), \ldots, (\alpha_n, \omega_n, \mu_n)$. The *length* $|\mathcal{P}|$ of a sequence of probes $\mathcal{P}$, is the number of probes in the sequence.

[1]A region corresponds to the probe span.

A sequence of probes $\mathcal{P}' = p'_{i_1}, \ldots, p'_{i_k}$ is said to be a *subsequence* of $\mathcal{P}$ if $p'_{i_1}, \ldots, p'_{i_k}$ are consecutive in $\mathcal{P}$. Formally, $\mathcal{P}' = p'_{i_1}, \ldots, p'_{i_k} = (\alpha'_{i_1}, \omega'_{i_1}, \mu'_{i_1}), \ldots, (\alpha'_{i_k}, \omega'_{i_k}, \mu'_{i_k})$ is a subsequence of $\mathcal{P}$ if and only if $(\omega'_{i_1} < \alpha'_{i_2}) \wedge \ldots \wedge (\omega'_{i_{k-1}} < \alpha'_{i_k}) \wedge \{\nexists\ p''_i = (\alpha''_i, \omega''_i, \mu''_i) \in \mathcal{P} \wedge \notin \mathcal{P}'$ such that $(\alpha''_i > \alpha'_1) \wedge (\alpha''_i < \omega'_k)\}$.

$\mathcal{P}_{1st}$ ($\mathcal{P}_{3rd}$) is a sequence of probes with $\log_2$ratio normalised data from the first (third) trimester placenta DNA array, and $\mathcal{P}_{wb}$ is a sequence of probes with $\log_2$ratio normalised data from the whole blood peripheral DNA array.

A subsequence $p_i, ..., p_{i+k}$ of $\mathcal{P}_{1st}$ ($\mathcal{P}_{3rd}$) is a region with span $[\alpha_i..\omega_{i+k}]$, and is said to be a *candidate region* of $\mathcal{P}_{1st}$ ($\mathcal{P}_{3rd}$) if it is of length at least three and the methylation status of each probe in the subsequence is greater than zero. Formally, subsequence $p_i, ..., p_{i+k} = (\alpha_i, \omega_i, \mu_i), ..., (\alpha_{i+k}, \omega_{i+k}, \mu_{i+k})$ of $\mathcal{P}_{1st}$ ($\mathcal{P}_{3rd}$) is a candidate region of $\mathcal{P}_{1st}$ ($\mathcal{P}_{3rd}$) if and only if $(k \geq 3) \wedge (\mu_\ell > 0\ \forall\ \ell \in [i..i+k])$. Similarly, a subsequence $p_i, ..., p_{i+k}$ of $\mathcal{P}_{wb}$ is said to be a candidate region of $\mathcal{P}_{wb}$ if it has length of at least three and the methylation status of each probe in the subsequence is less than zero.

III. The MS-AND algorithm

The inputs to the *methylation status-and* algorithm (MS-AND), and program, are the $\log_2$ratio normalised arrays, from which it is required to identify the DMRs. Using the definitions of Section II, we formally define our problem:

Problem 1 ($\mathcal{DMR}$ Detection): A $\mathcal{DMR}$ is defined to be the region of overlapping candidate regions of $\mathcal{P}_{1st}$, $\mathcal{P}_{3rd}$, and $\mathcal{P}_{wb}$, with a length of at least three. Given the three sequences of probes, $\mathcal{P}_{1st}$, $\mathcal{P}_{3rd}$, and $\mathcal{P}_{wb}$, find the $\mathcal{DMR}$s.

The solution presented below is influenced by the SHIFT-AND algorithm of [2], originally called SHIFT-OR, but that has come to be known as SHIFT-AND [7]. In that algorithm, the authors use bit masking and operations to quickly and efficiently identify exact patterns in strings, when the patterns are relatively small. Many other solutions to generic string processing problems, and more specific ones, for example in the areas of music information retrieval [4], and computational molecular biology [1], have also been influenced by the SHIFT-AND algorithm, and taken advantage of bit operations and implementation constructs, such as the C++ Standard Template Library's bitset container, which uses minimal space and time to perform said operations.

A. Outline

An outline of the algorithm:

Preprocessing: Align the arrays

A prerequisite for the MS-AND method is that the three input arrays are aligned, that is, the start and stop locations of the probe regions are the same: $(\alpha_{1st_i} = \alpha_{3rd_i} = \alpha_{wb_i}) \wedge (\omega_{1st_i} = \omega_{3rd_i} = \omega_{wb_i})\ \forall\ i \in [1..n]$, where $p_{1st_i} = (\alpha_{1st_i}, \omega_{1st_i}, \mu_{1st_i}) \in \mathcal{P}_{1st}$, $p_{3rd_i} = (\alpha_{3rd_i}, \omega_{3rd_i}, \mu_{3rd_i}) \in \mathcal{P}_{3rd}$, and $p_{wb_i} = (\alpha_{wb_i}, \omega_{wb_i}, \mu_{wb_i}) \in \mathcal{P}_{wb}$. Thus, this is a requirement when gathering the data.

Step 1: Mask the arrays

In this stage, we mask the three DNA arrays. Specifically, for the placenta DNA arrays ($\mathcal{P}_{1st}$ and $\mathcal{P}_{3rd}$) we set the bit to a 1 if there are three or more consecutively hypermethylated probe regions, and to a 0 otherwise. Similarly, for the whole blood DNA array $\mathcal{P}_{wb}$ we set the bit to a 1 if there are three or more consecutively hypomethylated probe regions, and to a 0 otherwise.

Formally, we define $\mathcal{MS}_{1st}$, $\mathcal{MS}_{3rd}$, and $\mathcal{MS}_{wb}$, to be arrays of bits, of length $|\mathcal{P}_{1st}|$ $(= |\mathcal{P}_{3rd}| = |\mathcal{P}_{wb}|)$ with elements set according to equations 1 and 2:

$$\mathcal{MS}_{1st_i(3rd_i)} = \begin{cases} 1, & \text{if } \mu_j > 0,\ \forall j \in [i-2..i] \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$\mathcal{MS}_{wb_i} = \begin{cases} 1, & \text{if } \mu_i < 0,\ \forall j \in [i-2..i] \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathcal{P}_{1st}$ $(\mathcal{P}_{3rd},\ \mathcal{P}_{wb}) = (\alpha_1, \omega_1, \mu_1), \ldots, (\alpha_n, \omega_n, \mu_n)$, for all $i \in [1..n]$.

STEP 2: IDENTIFY CANDIDATE DMRS

In this stage we identify regions that are concurrently hypermethylated in the two placenta DNA arrays and hypomethylated in the whole blood DNA. The output of this stage is a bitmask array, c$\mathcal{DMR}$ (candidate $\mathcal{DMR}$), which is constructed by performing an AND ($\wedge$) bit operation on the three MS arrays:

$$c\mathcal{DMR} = \mathcal{MS}_{1st} \wedge \mathcal{MS}_{3rd} \wedge \mathcal{MS}_{wb} \quad (3)$$

STEP 3: IDENTIFY DMRS

In the final stage, the DMRs are identified as being regions within the candidate DMRs array with three or more consecutive 1 bits:

$$\mathcal{DMR}_i = \begin{cases} 1, & \text{if } c\mathcal{DMR}_j = 1,\ \forall j \in [i-2..i] \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

The resulting array of bits ($\mathcal{DMR}$) is a map of the differentially methylated regions of the chromosome: a 1 for element i of the array signifies that the region $[a_{i-2}..w_i]$ is hypermethylated in 1st and 3rd trimester placenta DNA and hypomethylated in whole blood DNA.

Pseudo-code is shown in Algorithm 1.

Algorithm 1 Methylation status-and (MS-AND)

```
1: function IDENTIFY DMRS
       ▷ Mask the three pre-aligned arrays:
2:     MS_1st ← MASKARRAY("1st trimester")
3:     MS_3rd ← MASKARRAY("3rd trimester")
4:     MS_wb ← MASKARRAY("whole blood")
       ▷ Identify candidate DMRs:
5:     cDMR ← MS_1st & MS_3rd & MS_wb                    ▷ Equation 3
6:     i ← 0
7:     n ← |cDMR|
8:     DMR_1..n ← 0..0                                   ▷ initialise all bits to 0
       ▷ Identify DMRs:
9:     while i < n do
10:        if cDMR_j == 1, ∀j ∈ [i − 2..i] then          ▷ Equation 4
11:            DMR_i ← 1
12:            output region [a_{i−2}..w_i]
13:        increment i

14: function MASKARRAY(array)
15:    i ← 0
16:    MS_1..|array| ← 0..0
17:    while not end of file do
18:        (α_i, ω_i, μ_i) ← (start location, end location, methylation status)
19:        if array is "1st trimester" or "3rd trimester" then      ▷ Equation 1
20:            if μ_j > 0, ∀j ∈ [i − 2..i] then   ▷ 3 consecutive hyperm/d probes
21:                MS_i ← 1
22:                store α_{i−2} and ω_i
23:        else if array is "whole blood" then                      ▷ Equation 2
24:            if μ_i < 0, ∀j ∈ [i − 2..i] then   ▷ 3 consecutive hypom/d probes
25:                MS_i ← 1
26:                store α_{i−2} and ω_i
27:        increment i
28:    return MS
```

B. Algorithmic complexity analysis

Algorithm 1 runs in linear time, with a low hidden constant, and thus a fast execution time. Lines 2, 3, and 4 take $O(n)$ time each, where n is the number of probes in each of the GFF files. The key to the algorithm's and program's speed is the use of bit masking and operations, and the C++ STL's bitset container. For machine word size w, line 5 takes $O(n/w)$ time. Lines 9 to 13 also run in linear time $O(n)$. The hidden constant is of the order of 3, with the addition of $n/w + k_{1_{st}} + k_{3_{rd}} + k_{wb}$, where k_x is the number of occurrences in each of the three input arrays. We can represent $k_{1_{st}} + k_{3_{rd}} + k_{wb} = K$, which gives us a runtime of $O(3n + n/w + K)$. Furthermore, we know that $n/w < n$, and $k_x < n$, thus $O(K) = O(n)$, so this gives a total runtime of $O(n)$ for MS-AND.

The space complexity is $3n*$bit $+K*(2*$int $+$ double$))$, which is $O(n/w + K)$ This means that if there are $\sim n/w$ occurrences, the algorithm has a memory consumption of the order of the number of occurrences i.e., the number of hypermethylated (hypomethylated) regions.

IV. EXPERIMENTAL RESULTS

In order to use MeDIP as an enrichment method before non-invasive prenatal diagnosis, the selected markers need to be hypermethylated in the fetus and hypomethylated in the maternal whole blood [15].

The program was implemented in C++ and runs on Linux and Windows environments (`https://sourceforge.net/ms-and`). It takes as input the three array data files in General Feature File (GFF) format and three threshold values, t_{1st}, t_{3rd}, and t_{wb}.

Therefore in our experiments we identified those regions where:

- at least three consecutive oligonucleotide probes in each array have the same methylation status.
- they have a positive methylation status greater than $t_{1st}(t_{3rd})$ in the first and third trimester placenta data (fetal), and a negative methylation status less than t_{wb} in female whole blood data.

Our algorithm however, can be easily modified to identify regions where hypomethylation of fetal DNA and hypermethylation of maternal whole blood is required.

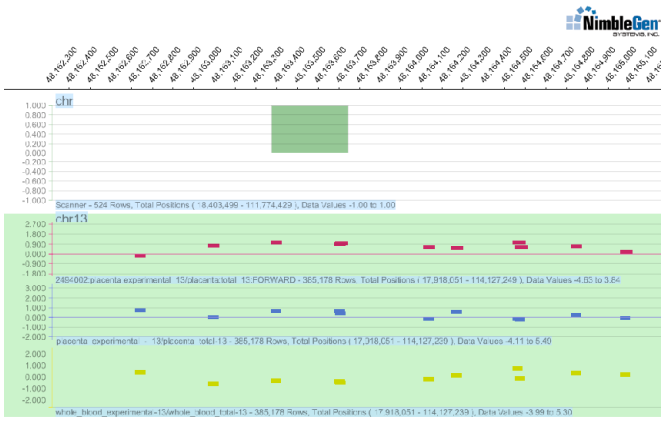

Fig. 1. A differentially methylated region of chromosome 13. The top track shows a DMR region identified by our algorithm. Below it, the track shows the three (at least) positive probes of 1^{st} trimester; the track below this shows the positive 3^{rd} trimester probes, and the bottom track shows the negative valued probes of whole blood for the region identified.

We tested our program on MeDIP array data for chromosomes 13, 18 and 21, with threshold values set to zero. For chromosome 13 we identified 524 regions hypermethylated in fetal DNA and hypomethylated in the female whole blood dataset. In chromosome 18 we identified 378 regions which were hypermethylated in the first two datasets and hypomethylated in the third. The largest number of DMRs were recorded for chromosome 21 where we found 871 regions which were hypermethylated in fetal DNA and hypomethylated in maternal blood.

The results were viewed and validated with the SignalMap software (Nimblegen System). Figure 1 presents a single differentially methylated region of chromosome 13. The top track shows our results, the chromosome 13 DMRs found by our algorithm. The second track shows chromosome 13 first trimester placenta array data. The third track is the third trimester placenta data for chromosome 13 and the bottom track is the female chromosome 13 whole blood array data. As we see from the figure, the region identified by our algorithm has positive hypermethylated values for the first trimester and third trimester placenta tracks and negative hypomethylated values for the female whole blood track. This differential methylation is as required for MeDIP quantitative assays.

Using the results of our automated procedure one can directly cross reference these DMRs with the gene regions of each chromosome as well as the CpG islands of each chromosome. We downloaded gene data tracks for chromosomes 13, 18, and 21, and the CpG islands (CGIs) of chromosomes 13, 18, and 21, from the UCSC genome browser (NCBI build 36), (http://genome.ucsc.edu). Figure 2 shows this investigation for all three chromosomes. Visual representation of DMRs, gene and CpG island tracks for chromosomes 13, 18, and 21. In the outer-most circle we can see the ideogram of each chromosome. In the inner-most circle, the DMRs reported from our algorithm are shown in green. Above these tracks, the second circle as we move away from the centre of the image, we view in blue the tracks corresponding to the genes of each chromosome. The next circle of tracks, shows the CpG islands of each chromosome in purple. The figure shows the high density of DMRs found in chromosome 21 compared to the other two chromosomes which is attributed not only to the fact that we have reported more DMRs on chromosome 21 than the other chromosomes (13 and 18), but also due to the higher resolution array used for chromosome 21 and due to the smaller size of chromosome 21 compared to the other two. We also observe that many of the chromosome 21 DMRs seem to fall outside CpG island regions, a majority of which are unmethylated [5].

TABLE I

TESTS ON CHROMOSOME 13 WITH VARYING THRESHOLD VALUES FOR t_{1st}, t_{3rd}, AND t_{wb}

threshold values			# identified regions
t_{1st}	t_{3rd}	t_{wb}	
0	0	0	524
0.25	0.25	-0.25	2
0.1	0.1	-0.1	92
0.1	0.15	-0.075	80

Tests were also run where the threshold values were non-zero. Table I shows the number of identified regions, for a few test runs for different threshold values, on chromosome 13. Although these results are promising, they need to be validated by biologists as to their usefulness.

V. CONCLUSION

We have presented an efficient algorithm for identifying differentially methylated regions, and shown how this is important as it is the first step in the discovery of biomarkers for certain NIPD techniques. After providing the biological setting, we defined our problem and the necessary constructs to then present the MS-AND algorithm. We then discussed our solution, its algorithmic complexity, and our experimental results on real data.

In our future research we will work on ways to more accurately identify clinically interesting biomarkers, with the MS-AND algorithm as a basis. The user-defined threshold values method is being validated by biologists, and if this indeed proves useful, machine learning statistical techniques can be used to automatically determine optimal threshold values during a learning period for specific datasets. Furthermore we plan to compare our implementation against other algorithm implementations, for accuracy of biomarker identification and algorithmic efficiency.

Another two approaches being considered by our group, also require human validation with regards to its usefulness, before further development. The first one is to sort and rank the output regions by the difference in methylation status between the hypermethylated and hypomethylated data. With this approach the runtime will depend on the number of outputs. For K identified regions, the running time is $O(n+KlogK)$. The second approach will classify the output

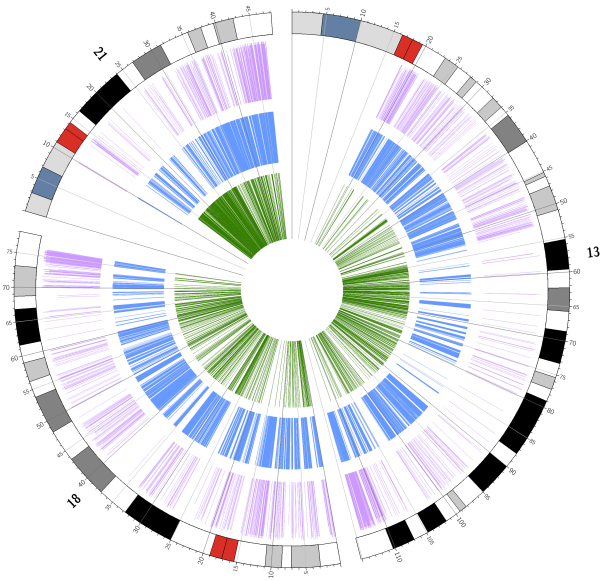

Fig. 2. Visual representation of DMRs, gene and CpG island tracks for chromosomes 13, 18, and 21. In the outer-most circle we can see the ideogram of each chromosome. In the inner-most circle, the DMRs reported from our algorithm are shown in green. Above these tracks, the second circle as we move away from the centre of the image, we view in blue the tracks corresponding to the genes of each chromosome. The next circle of tracks, shows the CpG islands of each chromosome in purple. Image created with the software circos [10].

regions depending on the difference in methylation status between the hypermethylated and hypomethylated data, into a small range of classes. This approach has the advantage of keeping the runtime of the algorithm linear.

VI. ACKNOWLEDGEMENTS

The authors would like to thank Marios Ioannides and Dr. Michael Hadjidaniel for their helpful discussion and comments on the subject and for testing the algorithm implementation.

REFERENCES

[1] Pavlos Antoniou, Costas S. Iliopoulos, Inuka Jayasekera, and M. Sohel Rahman. Implementation of a swap matching algorithm using a graph theoretic model. *Bioinformatics Research and Development*, 13:446–455, August 2008.

[2] Ricardo A. Baeza-Yates, Blanco Encalada, and Gaston H. Gonnet. A new approach to text searching. *Communications of the ACM*, 35:74–82, Oct 1992.

[3] P. A. Callinan and A. P. Feinberg. The emerging science of epigenomics. *Human Molecular Genetics*, 15 Spec No 1, April 2006.

[4] Joseph Wun-Tat Chan, Costas S. Iliopoulos, Spiros Michalakopoulos, and M. Sohel Rahman. Exact and approximate rhythm matching algorithms. *International Journal on Digital Libraries*, 12(2-3):149–158, August 2012.

[5] S. D. Fouse, R. O. Nagarajan, and J. F. Costello. Genome-scale DNA methylation analysis. *Epigenomics*, 2(1):105–117, Feb 2010.

[6] Jerry Guintivano, Michal Arad, Kellie Tamashiro, Todd Gould, and Zachary Kaminsky. Biotile, a perl based tool for the identification of differentially enriched regions in tiling microarray data. *BMC Bioinformatics*, 14(1):76, 2013.

[7] Dan Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[8] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83+, October 2012.

[9] William S. Klug and Michael R. Cummings. *Concepts of Genetics*. Prentice Hall, New York, sixth edition, 2000.

[10] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.

[11] Sheng Li, Francine Garrett-Bakelman, Altuna Akalin, Paul Zumbo, Ross Levine, Bik To, Ian Lewis, Anna Brown, Richard D'Andrea, Ari Melnick, and Christopher Mason. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics*, 14(Suppl 5):S10, 2013.

[12] Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh M. Ngo, Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren, and Joseph R. Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

[13] Y. M. Dennis Lo, Noemi Corbetta, Paul F. Chamberlain, Vik Rai, Ian L. Sargent, Christopher W.G. Redman, and James S. Wainscoat. Presence of fetal DNA in maternal plasma and serum. *The Lancet*, 350(9076):485 – 487, 1997.

[14] Susan K. Murphy, Zhiqing Huang, and Cathrine Hoyo. Differentially methylated regions of imprinted genes in prenatal, perinatal and postnatal human tissues. *PLoS ONE*, 7(7):e40924, 07 2012.

[15] Elisavet A. Papageorgiou, Heike Fiegler, Vardhman Rakyan, Stephan Beck, Maj Hulten, Klea Lamnissou, Nigel P. Carter, and Philippos C. Patsalis. Sites of differential dna methylation between placenta and peripheral blood: Molecular markers for noninvasive prenatal diagnosis of aneuploidies. *The American Journal of Pathology*, 174(5):1609 – 1618, 2009.

[16] Elisavet A. Papageorgiou, Alex Karagrigoriou, Evdokia Tsaliki, Voula Velissariou, Nigel P Carter, and Philippos C Patsalis. Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. *Nat Med*, 17(4):510–3, 2011.

[17] Philippos C. Patsalis. A new method for non-invasive prenatal diagnosis of down syndrome using MeDIP real time qPCR. *Applied & Translational Genomics*, 1(0):3 – 8, 2012.

[18] Nina Plmke, Diana Santacruz, and Jrn Walter. Comprehensive analysis of DNA-methylation in mammalian tissues using MeDIP-chip. *Methods*, 53(2):175 – 184, 2011. Functional Mouse Genomics.

[19] Thomas S. Price, Regina Regan, Richard Mott, Asa Hedman, Ben Honey, Rachael J. Daniels, Lee Smith, Andy Greenfield, Ana Tiganescu, Veronica Buckle, Nicki Ventress, Helena Ayyub, Anita Salhan, Susana Pedraza-Diaz, John Broxholme, Jiannis Ragoussis, Douglas R. Higgs, Jonathan Flint, and Samantha J. Knight. Swarray: a dynamic programming solution for the identification of copy-number changes in genomic DNA using array comparative genome hybridization data. *Nucleic acids research*, 33(11):3455–3464, 2005.

[20] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[21] Jianzhong Su, Haidan Yan, Yanjun Wei, Hongbo Liu, Hui Liu, Fang Wang, Jie Lv, Qiong Wu, and Yan Zhang. Cpg_mps: identification of cpg methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Research*, 41(1):e4, 2013.

[22] Miho M. Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, 2008.

[23] Charles D. Warden, Heehyoung Lee, Joshua D. Tompkins, Xiaojin Li, Charles Wang, Arthur D. Riggs, Hua Yu, Richard Jove, and Yate-Ching Yuan. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Research*, 41(11):e117, June 2013.

[24] Yan Zhang, Hongbo Liu, Jie Lv, Xue Xiao, Jiang Zhu, Xiaojuan Liu, Jianzhong Su, Xia Li, Qiong Wu, Fang Wang, and Ying Cui. Qdmr: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Research*, 39(9):e58, 2011.